Language as Design:
Adapting Language to Different Online Audiences

Tal August

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2022

Reading Committee:
Katharina Reinecke, Co-Chair
Noah A. Smith, Co-Chair
Dan Weld
Emma Spiro

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Language as Design:

Adapting Language to Different Online Audiences

Tal August

Chair of the Supervisory Committee:

Katharina Reinecke

Noah A. Smith

Computer Science and Engineering

One of our most powerful language capabilities is our ability to adapt written and spoken language use to different audiences (sometimes referred to as audience design or linguistic accommodation). How we explain a topic to a fifth grader differs from how we explain it to a college student, or how we write about it in a paper. Targeting language messages to different receivers enriches and empowers our communication. However, as online audiences expand in size and demographics, it becomes increasingly difficult to adapt to all potential receivers. Though research papers, news articles, legal documents and social media posts proliferate on the internet, much of their language appeals to an ever-narrowing audience segment. New techniques in natural language processing (NLP) have the potential to make such language adaptation automatic. However, developing systems that effectively rewrite language require an understanding of what language is important to change.

In this thesis, we show that language style changes, similar to other interface design changes, influence user behavior and introduce automated systems that design language for different people. We begin by focusing the study of language style changes to the subreddit *r/science* and show how language in it is associated with changes in people's behavior, potentially restricting access to scientific information. To

understand what language is important to change when adapting to different people, we investigate how experts design scientific language for a general audience. We take inspiration from these expert strategies to build PAPER PLAIN – a reading interface for making medical research papers approachable to a general audience. To adjust language to finer-grained audiences, we investigate how people respond to levels of language complexity based on their background knowledge and develop a novel controllable generation method to adjust the complexity of generated summaries. In two user studies we observed that generated summaries using our method leads to similar reader responses as with expert summaries, establishing the feasibility of generating summaries with varying complexities. Our work provides guidance on designing language for specific audiences and adaptable communication at scale. We conclude with a summary of the contributions and a discussion of future research on designing language to encourage better communication online.

Acknowledgements

First, of course, are my advisors, Katharina and Noah. Katharina, thank you for envisioning this thesis with me. Your unflagging guidance and conviction in this research were a constant support through research and life events. Noah, thank you for welcoming me into a new discipline. Your patience and interest were necessary ingredients to talking about ideas I didn't have the words for when I started. To both of you, thank you for fostering a warm groups of students. They were both sources of joy when I was most struggling with my own work.

Next, to my committee and other mentors through my PhD. Dan Weld, whose questions kept me thinking about what's next. Emma Spiro, for shifting my thinking about the applications of this work. Shamsi Iqbal, who was a wonderful mentor while I was still figuring out what I wanted to do. Tim Althoff, whose excitement about research, and exciting research, I hope continues to rub off on me. Kyle Lo, for spending those hour-turned-to-three-hour meetings figuring out the exact right way to say something. Nigini Oliveira, for being the first person to introduce me to the weeds of research. Keshet Ronen, for your patience and interest learning to align our language. Andrew Head, for your drive to make our work ever better. I look forward to many more conversations with all of you. Thank you also to my collaborators: Lucy Lu Wang, Jonathan Bragg, Marti Hearst, Elizabeth Clark, Sofia Serrano, Suchin Gururangan, Amanda Baughan, Naomi Yamashita, Lauren Kim, Dave Wadden, Qisheng Li, Michael Gamon, Mark Encarnación, Maarten Sap, Dallas Card, Gary Hsieh, Chenhao Tan, Narshion Ngao, Manuel Nordhoff, Spencer Williams, Rock Pang, Lucy Lin, and Sebastin Santy. Also to my friends from CSE who introduced me to tabletop gaming, our weekly games are always something I look forward to. In addition, thank you to all the people on LabintheWild and Mechanical Turk who contributed data to my experiments. Thank you also to Martha Rimniceanu for submitting graduate applications in Tisch with me, all those years ago.

Thank you to my parents and family. My mom, for reminding me that my feelings were developmentally appropriate and for celebrating all the small victories. To my dad, for teaching me to ask questions that might or might not have answers. You showed me that language is an interface, not that you would say it in so few words. To my brother Alon, thanks for reminding me that I sacrificed a lot of other skills to be this smart. To my sister Siona, thanks for horse and truck lessons over long calls. To my wife Kim and my daughter Rose. Kim, thank you for everything. For each word in this dissertation, you have shared with me a conversation, a dinner, a look, that made it possible. To Rose, thanks for showing me how everything can shift without moving. Finally, to our dog Callie, thanks for always reminding me to lay in the sun.

DEDICATION

To my daughter Rose

I so love being outside language with you

Contents

List of Figures

List of Tables

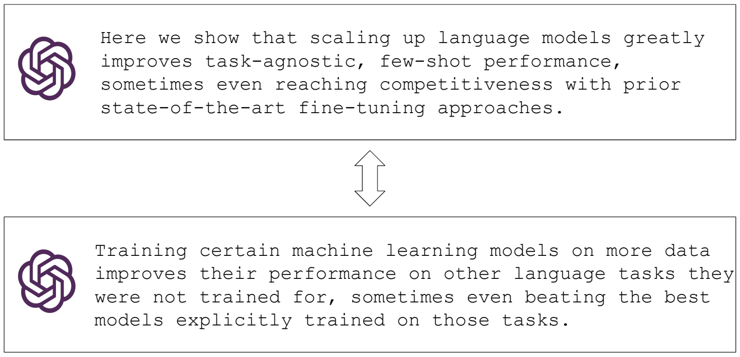
Chapter 1

Introduction

Language is humans' most powerful medium of communication. We use it to exchange information, to build relationships and trust [Scissors et al., 2008; Fusaroli et al., 2012; Milroy and Milroy, 1978; Sharma and De Choudhury, 2018], to construct our identities [Haider-Markel and Joslyn, 2001; Labov, 2006; Danescu-Niculescu-Mizil et al., 2013], and to make sense of the world [Chong and Druckman, 2007; Tversky and Kahneman, 1981; Entman, 1993]. We also change the style of language we use depending on who we are communicating with, sometimes referred to as audience design [Bell, 1984], or linguistic accommodation [Giles et al., 1991]. For example, how we explain an experiment to a fifth grader differs from how we would explain it to a college student, or how we write about it in a paper.

Such language adaptations are necessary to share potentially life-saving information and improve people's lives. Patients can make use of plain language summaries of medical research to find treatments while also feeling more confident in their doctor's recommendations [Zuccala, 2010; National Institutes of Health, 2005; Tennant et al., 2016; Day et al., 2020; Epstein, 1997], and people can build important support networks through conversations in online communities [Danescu-Niculescu-Mizil et al., 2013; Sharma and De Choudhury, 2018]. However, when language isn't well suited to its audience, it can instead restrict access to these resources. A biomedical paper abstract can be useful to a researcher but confusing and demoralizing to a patient seeking new treatments [Day et al., 2020; Nunn and Pinfield, 2014; August et al., 2022b], and successfully engaging with an online community requires aligning with language norms (sometimes unstated) in that community [Danescu-Niculescu-Mizil et al., 2013; August et al., 2020a]. Adapting language to different people (e.g., writing a plain language version of a research paper abstract, as in Figure 1.1) could help them access resources otherwise out of reach.

Because of the way language can either support or impede communication, it can be thought of as an interface. Similar to how an interface defines the communication that can happen between two entities (e.g., a person and a computer), language allows people to interact with information and each other in ways not possible otherwise. While there are many examples of text-centric online interfaces that primarily use language to communicate (e.g., news articles, research papers, forum posts), rarely is language investigated as a mutable element of design in these interfaces. Research on adapting interfaces to different contexts and users has instead investigated changing features such as visual aesthetics and structure [Reinecke and Bernstein, 2013], layout [Gajos et al., 2008b], or language localization (e.g., changing from English to German) [Wang et al., 2019]. While language style is a part of these interface qualities (e.g., lots of text might make a visual layout more cluttered), rarely is the language style itself made flexible. One reason for this is because it is hard to know what parts of language to change and how these changes will affect an audience's behavior.



Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches.

Training certain machine learning models on more data improves their performance on other language tasks they were not trained for, sometimes even beating the best models explicitly trained on those tasks.

Figure 1.1: Example of how changing language can make information more approachable for different audiences. The first sentence is taken from the paper on GPT-3 [Brown et al., 2020], which was designed for NLP researchers. The second sentence reports on roughly the same information but is designed for a general audience.

Previous work, including our own, has shown that language style can influence people's behavior in diverse online settings. For example, in online experiments, changing the framing of a study's slogan (e.g., emphasizing the personal benefit of completing a study vs. how completing a study will further science) can significantly impact the recruitment of a study [August et al., 2018]. Changing the formality of language in online experiments [August and Reinecke, 2019] and security prompts [Stokes et al., 2023] can also impact a user's attention and reported compliance. In clinical settings, creating patient pamphlets based on the interests and health history of individual patients can improve engagement and understanding [Skinner et al., 1994; Marco et al., 2006] This prior work illustrates how language style can influence people's behavior, necessitating a closer look at designing language appropriate for different users and contexts.

Thesis: Adapting language in text-centric interfaces to different audiences using automated techniques can improve engagement and make information more widely approachable on the Internet. In this thesis, we (1) show that language style influences user behavior and (2) introduce methods and systems that design language for different people. We make the following contributions:

1. Empirical results showing how language style (e.g., changing complexity) influences people differently, potentially restricting access to information for some people. These results point to the importance of investigating language as a mutable element of interface design.

2. Technical methods for identifying language style at scale using natural language processing (NLP) in the context of science communication. Our models and results contribute to work on strategies for engaging different audiences in science communication.

3. A novel system leveraging NLP techniques for adapting medical research papers to general audience readers. Our system contributes to work on designing AI augmented reading interfaces.

4. A novel technique for guiding text generation models to adjust the complexity of generated text. This technique contributes to work in controllable text generation, specifically in the context of quantifying and controlling language complexity.

1.1 Outline

In Chapter 2 we provide an overview of research in the topics this thesis draws on and contributes to. We cover the visual features commonly explored in interface personalization, research on language style for different audiences, the benefits and risks of language models, and science communication. We outline how this thesis contributes to each domain of research, pointing to research in this thesis that identifies language style as an additional feature of personalization, develops new generation techniques using language models, and new findings on science communication strategies. Each subsequent chapter additionally covers the relevant work and contributions of that chapter.

In Chapter 3, we show that language style is associated with changes in people's behavior in the online forum *r/science* [August et al., 2020a], potentially restricting access to information. We investigate the large science communication forum, *r/science*, finding that members used specialized language, such as scientific jargon, that was distinct from language common in other subreddits. New users who left *r/science* used less of this specialized language than those that stayed, suggesting that the specialized language can discourage some from contributing to the community. *r/science* has millions of subscribers, suggesting that such language changes could vastly broaden access to scientific information.

In Chapter 4 we identify ways that experts design scientific language for a general audience [August et al., 2020b]. We focus on science communication because it presents many barriers in communication (e.g., lack of public understanding about science) and overcoming these barriers promises substantial public benefits [Nisbet and Scheufele, 2009a] (e.g., reducing vaccine hesitancy [Goldenberg, 2016]). We introduce a set of writing strategies drawn from a wide range of prescriptive sources in science communication. To measure strategy use at scale, we develop a novel text classification task and discriminators using recent NLP advances for applying models to new language domains, including pretrained language models [Liu et al., 2019a] and task and domain-adaptive pretraining [Gururangan et al., 2020]. We find that the use of strategies, such as storytelling and emphasizing the most important findings, varied significantly across publications with different reader audiences in a corpus of 128K science news articles, magazines, press releases, and blog posts. This suggests that while all strategies can be effective at designing scientific language for a general audience, different strategies might be more helpful for different readers.

In Chapter 5 we take inspiration from these strategies to build PAPER PLAIN – a reading interface for making medical research papers approachable to a general audience [August et al., 2022b]. PAPER PLAIN supports general audience readers of medical research papers by guiding readers with four features powered by NLP: definitions of unfamiliar terms, in-situ plain language section summaries, a collection of key questions that guide readers to answering passages, and plain language summaries of the answering passages. In a user evaluation of PAPER PLAIN, participants who used PAPER PLAIN had an easier time reading and understanding research papers compared to those who used a typical PDF reader.

Our goal is to adapt language to different audiences, and people have different knowledge that can impact how they respond to scientific information [Nisbet and Scheufele, 2009b; Forzani, 2016; Bliss, 2019].

Paper Plain used a single version of text for all potential readers, limiting its ability to communicate with diverse stakeholders. In Chapter 6, we investigate how people respond to finer-grained levels of language complexity and if readers' background knowledge influences their response to language complexity. We conduct a within-subjects study that presents readers with summaries of scientific papers at three levels of complexity and measures their paper comprehension and self-reported reading experience. We find that the least complex summaries were easier to understand and perceived as more interesting by participants who had little background knowledge in a topic. In contrast, those with high topic familiarity gained no such benefit from low complexity text. To understand the feasibility of generating such summaries, we develop a novel controllable generation method to adjust the complexity of generated summaries [August et al., 2022a]. We observe highly similar impacts of complexity using generated summaries curated for factuality in this follow-up study, establishing the feasibility of generating summaries at varying complexities [August et al., 2023]. Our work provides guidance on designing language for specific audiences and lays the foundation for tools supporting adaptable science communication at scale. We conclude this thesis with a summary of the contributions and a discussion of future research on designing language to improve access to information and encourage better communication online.

1.2 Terminology

In this thesis we refer to **language choices** or **language style** to denote changes in the way language conveys information (often referred to as style) rather than what information is conveyed (often referred to as content). Language style and content can never be entirely divorced from one another (e.g., changing language framing will change what information is made salient, and changing language complexity will change how detailed the information is), but our focus is changing style without dramatically changing content. We consider language style in English only for this thesis, though we discuss in Chapter 7 future work exploring contexts with multiple languages. We refer to **general audience** readers as readers without expertise in a given scientific or technical topic, sometimes called a lay audience.

Chapter 2

Related Work

This thesis draws on research in adaptive user interfaces, language style, language models, and science communication. Each chapter will cover prior work relevant to the individual project. Here we will broadly summarize these research trends and how they relate to the thesis as a whole.

2.1 Adaptive User Interfaces

There is a rich body of research on adapting interfaces to different users. The most relevant to this thesis are online interfaces that adapt visual characteristics or show the benefit of doing so in controlled settings. Aesthetic judgments vary widely by user, suggesting that personalising website aesthetics could benefit user trust and usability [Lavie and Tractinsky, 2004; Reinecke and Bernstein, 2013]. Lindgaard et al. [2011] and Moshagen and Thielsch [2010] showed that judgments of website visual aesthetics were associated with perceived usability and trust. Reinecke and Gajos [2014] showed that users' visual preferences for website complexity and color were influenced by their age, gender, and cultural background. Reinecke et al. [2011] introduced the adaptive system MOCCA, which could automatically adapt visual features based on a user's culture. Users' visual preferences are also associated with how fast they are able to get information from a website [Baughan et al., 2020]. Other work has identified common interface elements that are associated with aesthetic judgements, such as typography, images, layout, whitespace, and color [de Souza Lima and von Wangenheim, 2022; Mõttus and Lamas, 2015].

Specific to text, Wallace et al. [2022] showed that fonts, the visual representation of language in an interface, can impact readability and that different readers prefer different fonts. Rello and Baeza-Yates [2016] found that typographic features like san serifs and monospacing can have a significant effect on webpage readability for readers with dyslexia, necessitating a personalised approach to webpage fonts. Miniukovich et al. [2017] compiled a set of 12 design guidelines for text, such as using larger fonts, for personalizing website readability for people with dyslexia. In a similar vein, adjusting the language an interface is in, known as language localization, is a common practice for adapting interfaces to different countries [Wang et al., 2019].

In contrast, to visual features or language localization, this thesis explores language features within a single language (e.g., the formality or complexity of the text) as an important element of adaptation in text-centric online interfaces. We explore how language features are associated with changes in user behavior and how adjusting these features can improve communication online.

2.2 Language Style and Variation

While adapting interface language is rare, we adapt language all the time when speaking to someone. In sociolingustics, language style is the linguistic variation a speaker exhibits in different social contexts [Eckert, 2004]. Style has been considered a way of constructing or expressing identities [Eckert, 2004; Coupland, 2002] and adjusting to different audiences [Bell, 1984]. In much of this early research, style was localized to specific linguistic variables (e.g., pronunciation of the /r/ sound [Labov, 1973, 2006]), but research has also shown that people align language styles in terms of words and phrases to one another when speaking, known as linguistic accommodation [Giles et al., 1991], and that this can improve communication. [Scissors et al., 2008] showed that people in text message conversations had improved trust when they had higher accommodation. Linguistic accommodation is also associated with more meaningful mental health conversations Wadden et al. [2021], higher community acceptance [Sharma and De Choudhury, 2018], and team performance [Fusaroli et al., 2012; Gonzales et al., 2010].

In a similar vein of research, Bell [1984] identified how people shape their language style to their audience, a phenomenon known as Audience Design. Work has shown evidence of audience design in both offline [Ferreira, 2019; Bell, 1984; Danescu-Niculescu-Mizil et al., 2012] and online interactions [Wadden et al., 2021; Rudat et al., 2014]. For example, Tan et al. [2014] found that aligning language to one's audience on Twitter can increase retweets. In similar work, Danescu-Niculescu-Mizil et al. [2013] found that adopting the language of a community is predictive of how long someone will stay in that community.

As more communication moves online, people's audiences are becoming more diverse and difficult to distinguish. How can someone writing a blog post know all of their potential readers? Someone could write additional text (such as explanations of concepts unfamiliar to some, but not all, readers) to meet each potential reader, but it is not clear what each reader might need, and even if it was, writing additional text versions would be prohibitively costly and time consuming. This thesis identifies what language it is important to change for whom and introduces methods to make this language design scalable.

2.3 Language Models Augmenting Communication

Recent advances in NLP have the potential to improve online communication. Driving these advances are language models (LMs). LMs are a type of unsupervised machine learning model whose basic task is to predict either the next word in a sequence, or a set of masked words within a sequence (e.g., "The small dog walked in the ___." or "The small ___ ___ in the park."). A growing body of research in NLP has shown that training large language models (now approaching 1 trillion parameters) on large amounts of text (approaching 1 trillion words) can lead to downstream improvements on many other language understanding tasks (e.g., question answering or summarization) with little or no additional training [Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019a; Chowdhery et al., 2022]. LMs like GPT-3 [Brown et al., 2020] have generated text that is in some contexts indistinguishable from human-authored language [Clark et al., 2021]. Recent work has used LMs to generate versions of text suitable for audiences different from the original text, such as a plain language summaries of a medical paper [Guo et al., 2021].

As promising as LMs are, they also have significant downsides. Because the models are trained on so much text, it is hard to know or control what a model is learning. LMs can generate hateful and toxic language [Gehman et al., 2020], and learn undesirable social biases from the training data [Brown et al., 2020]. Furthermore, because LMs generate by sampling a probability space, there is no guarantee that the language is factual or consistent with previous generations [Maynez et al., 2020]. Bommasani et al. [2021] provide a review of LMs and their risks. There has also been work on encouraging factuality in LMs generating summaries by constraining generation with logical constraints [Lu et al., 2020], providing evaluation frameworks for factuality [Gabriel et al., 2020; Fabbri et al., 2020; Scialom et al., 2019a], and encouraging models based on discriminator architectures [Gabriel et al., 2021a].

The work in this thesis introduces a novel method of using LMs to generate text at varying levels of complexity [August et al., 2022a] and develops systems that leverage LMs, among other NLP advances, to augment and adapt language for different people. Because the context of language we explore is science communication, we are particularly mindful of the risks LMs pose to factuality. We discuss these issues further in Chapters 5 and 6.

2.4 Science Communication

This thesis focuses on the domain of science communication for designing language, contributing to research on expanding access to scientific information online. Over the past twenty years, science communication has shifted from improving scientific literacy to fostering participation in science [Hetland, 2014]. A growing body of research has shown that scientific literacy is only one of many factors that influence public decision making and cannot be divorced from cultural values [Nisbet and Scheufele, 2009b; Bubela et al., 2009]. Science writing today sometimes still addresses information deficits, but often also tries to encourage two-way interactions and engagement in science [Burns et al., 2003].

There is a wealth of work exploring writing in scientific journals (i.e., when scientists communicate within their discipline). Because of the natural structure of scientific journal papers, much work has looked at ways of automatically identifying content in these papers [Liakata et al., 2010; Guo et al., 2010; Liakata et al., 2012] and the use of guidelines to structure journal papers [Kröll et al., 2014].

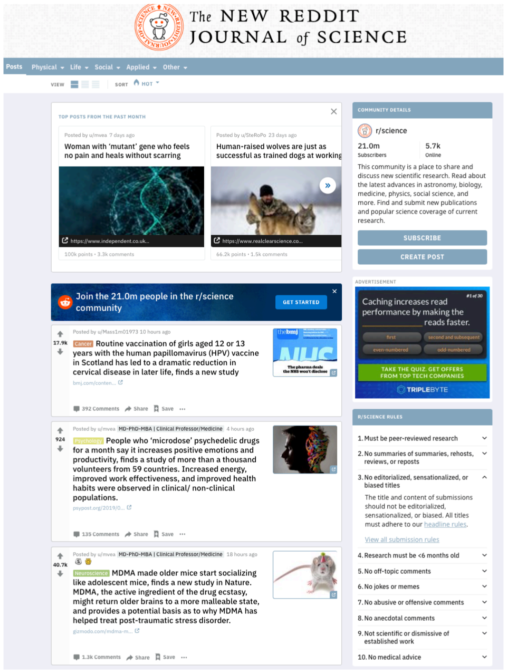
Previous work on improving science communication (i.e., text used to communicate outside of a discipline) has introduced tools that identify or restrict jargon (e.g., xkcd's Simple Writer[1]) or summarize journal articles [Vadapalli et al., 2018]. Luzón [2013] evaluated science blogging strategies, such as length of posts and update frequency, to engage readers with different levels of science knowledge. Fan et al. [2019] introduced models for generating explanations of complex questions from the *r/explainlikeiamfive* corpus. Cachola et al. [2020] introduced the task of generating short summaries (TLDRs) of research papers, while Dangovski et al. [2020] focused on generating lay summaries of scientific papers. Additional

[1] `https://xkcd.com/simplewriter/`

work has also examined effective techniques for engaging readers [Putortì et al., 2020]. This thesis builds on work in science communication by exploring how to augment scientific text to make it approachable to more people.

Chapter 3

The Linguistic Barriers of Entry to *r/science*



Posted by u/mvea 7 days ago
Woman with 'mutant' gene who feels no pain and heals without scarring
https://www.independent.co.uk...
100k points • 3.3k comments

Posted by u/SteRoPo 23 days ago
Human-raised wolves are just as successful as trained dogs at working
https://www.realclearscience.co...
66.2k points • 1.5k comments

Join the 21.0m people in the r/science community

Posted by u/Mass1m01973 10 hours ago
Cancer Routine vaccination of girls aged 12 or 13 years with the human papillomavirus (HPV) vaccine in Scotland has led to a dramatic reduction in cervical disease in later life, finds a new study
bmj.com/conten...

Posted by u/mvea MD-PhD-MBA | Clinical Professor/Medicine 4 hours ago
Psychology People who 'microdose' psychedelic drugs for a month say it increases positive emotions and productivity, finds a study of more than a thousand volunteers from 59 countries. Increased energy, improved work effectiveness, and improved health habits were observed in clinical/ non-clinical populations.
psypost.org/2019/0...

Posted by u/mvea MD-PhD-MBA | Clinical Professor/Medicine 18 hours ago
Neuroscience MDMA made older mice start socializing like adolescent mice, finds a new study in Nature. MDMA, the active ingredient of the drug ecstasy, might return older brains to a more malleable state, and provides a potential basis as to why MDMA has helped treat post-traumatic stress disorder.
gizmodo.com/mdma-m...

This community is a place to share and discuss new scientific research. Read about the latest advances in astronomy, biology, medicine, physics, social science, and more. Find and submit new publications and popular science coverage of current research.

1. Must be peer-reviewed research
2. No summaries of summaries, rehosts, reviews, or reposts
3. No editorialized, sensationalized, or biased titles

   The title and content of submissions should not be editorialized, sensationalized, or biased. All titles must adhere to our headline rules.

   View all submission rules
4. Research must be <6 months old
5. No off-topic comments
6. No jokes or memes
7. No abusive or offensive comments
8. No anecdotal comments
9. Not scientific or dismissive of established work
10. No medical advice

Figure 3.1: The main page of *r/science* with top posts on the left and a description of the community and its rules on the right.

Online science forums are a prime example of the growing focus on two-way conversations in science communication. These sites also offer a ideal context in which to study language choices in science communication, as each post and comment represents language choices that can either facilitate or hinder communication. Because many of these forums are public, new members across the internet are able to join to discuss and learn about new scientific findings. We begin our study of language as design in the online science communication forum, *r/science*. We show how members of *r/science* use scientific jargon and impersonal language more often than other large or topically related subreddits. New members use common *r/science* language less than more frequent members, and members who end up leaving *r/science* use less of this specialized language than those that end up staying. Below we cover our analyses and findings in detail and discuss the implications of such language choices in online science communication. This chapter includes materials originally published in August et al. [2020a].

3.1 Introduction

r/science, a sub forum ("subreddit") on the social news aggregation platform Reddit and shown in Figure 3.1, has emerged as one of the largest platforms for disseminating and discussing scientific findings outside of the social circles of the scientists themselves. As an online community for discussing research findings, *r/science* has the potential to contribute to science outreach and communication with a broad audience. However, past work has shown that active contributors in *r/science* are largely already involved in scientific activity, and broader public engagement is lacking [Jones et al., 2019].

A possible reason why *r/science* does not attract broader contribution is that the community might have developed specialized language that deter some users from actively participating. Commonly used insider words [Danescu-Niculescu-Mizil et al., 2013], politeness or formality [Burke and Kraut, 2008; Nguyen and Rosé, 2011], and even pronoun usage [Tran and Ostendorf, 2016] are all examples of linguistic behaviors developed within a community that characterize its language. Just as with other normative behaviors, the language is important for new users to adopt: users who don't adopt it receive fewer and less supportive responses [Jaech et al., 2015; Sharma and De Choudhury, 2018] and are more likely to leave an online community [Danescu-Niculescu-Mizil et al., 2013]. Given that science communication is often hindered by specialized language [Plavén-Sigray et al., 2017], it is possible that the *r/science* community has developed language norms that prevent a wider public from engaging.

To explore whether *r/science* contains specialized language that may present a barrier to entry, we analyzed differences in the language of 68,560,317 publicly available posts and comments on Reddit. We compared the language used in *r/science* to the language used by contributors of 11 other subreddits using language models, a commonly used technique for measuring language differences [Danescu-Niculescu-Mizil et al., 2013].

We found our hypothesis to be true: the language used in *r/science* posts and comments differs from the language used in other large or topically related subreddits. Many of the distinctions in language in *r/science* reflect *r/science*'s rules for language norms, such as hedging (qualifying statements with words such as 'suggest' and 'possibly'), impersonal language, and scientific terminology. Our results show that transient contributors (i.e., those who only post once) on *r/science* fail to adapt to this specialized language more often than more experienced authors, suggesting that *r/science*'s language norms do not always come naturally to people. Our findings indicate that *r/science*'s guidelines and community norms, while useful to maintain a high standard of rigor and discourse, have the side effect of limiting contributions from a broader audience by enforcing a specialized language that people wanting to post and comment first have to learn.

3.2 Related Work

Here we cover work relevant to the current chapter on online community norms and guidelines and studies of specialized language in online communities.

3.2.1 Online Community Norms and Guidelines

Online communities develop norms and guidelines that users follow to be productive members. These guidelines are often decided as a rough consensus of members around what behavior is or is not acceptable in the community [Kraut and Resnick, 2012]. There are some community standards governing the entire Reddit platform ("reddiquette"), but it is common for individual subreddits to have their own set of specialized rules and norms [Fiesler et al., 2018]. Chandrasekharan et al. [2018] characterized Reddit community rules into three levels: Macro (i.e., rules shared across all of Reddit), Meso (i.e., rules shared across groups of subreddits), and Micro (i.e., subreddit specific rules). More popular subreddits usually have more structured norms to handle and socialize large influxes of new members without losing community values [Fiesler et al., 2018]. Moderators also play a strong role in developing and enforcing community norms, which can range from being highly active in a community, such as posting content often, to only being involved when a member seriously violates a rule [Seering et al., 2019].

Although community norms are often publicly displayed, newcomers can be overwhelmed by these rules, risking community rejection by unknowingly violating norms [Halfaker et al., 2009; Kraut and Resnick, 2012], or turned off from the community by guidelines clashing with their own values [Oliveira et al., 2018]. Research suggests that Wikipedia's sharp decline in retention of desirable new editors (i.e., not vandals) from around 40% in 2003 to less than 10% in 2010 can partly be attributed to inflexible rules [Halfaker et al., 2013]. Community guidelines can also deter some users from joining a community due to an unintended clash in values, such as StackOverflow's rule of "No thank you's", that conflicts with many users' beliefs around healthy community support [Oliveira et al., 2018].

Publicly displaying community norms supports new users in interacting with the norms of the community, which can help increase normative behavior in newcomers [Matias and Mou, 2018]. Cialdini et al. [1990] characterized two ways norms influence behavior: injunctively, where norms prescribe acceptable behaviours in the group, and descriptively, where member behaviour provides examples of norms. Morgan and Filippova [2018] characterized these injunctive and descriptive norms in Wikipedia sub-communities, identifying injunctive norms as posted guidelines and descriptive norms as active community threads. Both injunctive and descriptive norms can increase normative behaviour in online communities, especially when injunctive norms are reinforced with descriptive norms, or vice versa.

While many of the guidelines and moderation in online communities focus on acceptable behaviors, like how people should treat other members or what they can post about, communities also develop unique *language* norms, such as specific words members use that are important for new members to follow in order to receive support from the community [Danescu-Niculescu-Mizil et al., 2013; Jaech et al., 2015; Sharma and De Choudhury, 2018]. In this chapter, we identify how the language norms that *r/science* has developed can act as an additional gatekeeper to participation.

3.2.2 Community Language

Studying community language norms has a long history in sociolinguistic research. Labov [2006] studied fine-grained phonetic differences in New York City, showing that different socio-economic classes, and even small peer groups within these classes, use significantly different vocalizations, such as a different pronunciation of the /r/ sound [Labov, 1973, 2006]. Milroy and Milroy [1978] showed in their seminal work exploring vernacular English in Belfast, Northern Ireland, that community networks, such as familial ties, were a strong influence on an individual's linguistic variation.

Research has drawn comparable findings for linguistic variation in online communities [Danescu-Niculescu-Mizil et al., 2013; Tran and Ostendorf, 2016; Nguyen and Rosé, 2011]. Cassell and Tversky [2005] explored the formation of a new online community comprised of children from around the world, finding that these children from diverse cultural, economic, and geographic backgrounds converged on a shared language style, such as speaking in the collective voice, and topics of conversation. Danescu-Niculescu-Mizil et al. [2013] had similar findings in online beer enthusiast communities, showing that members adopt certain words (e.g., “aroma” or fruit-related words) that become widespread in the community. Tran and Ostendorf [2016] characterized the language style of eight subreddit communities, showing that stylistic features, such as community-specific jargon and sentence structure, led to close to 90% accuracy in identifying a community. Zhang et al. [2017] built on this research, exploring a larger subset of 300 subreddits and showing that frequently used words within a subreddit were also useful in characterizing how the community was distinctive (different from other communities) and dynamic (how quickly the community shifted to new topics).

Adopting the language style of a community is important for being accepted by that community [Burke and Kraut, 2008; Danescu-Niculescu-Mizil et al., 2013; Jaech et al., 2015]. For example, users who did not adopt the specific words common in the beer enthusiast communities mentioned above were more likely to leave compared to members who did adopt these new words [Danescu-Niculescu-Mizil et al., 2013]. Members of breast cancer online support groups use much more community-specific jargon and informal language the longer they remain part of the community, indicating that language style is a reflection of community socialization [Nguyen and Rosé, 2011]. In addition, members of mental health support groups on Reddit are more likely to receive supportive responses if their language matches the style of the community [Sharma and De Choudhury, 2018], and Twitter users who match the language style of their followers receive more retweets [Tan et al., 2014]. We build on this work by exploring the language norms of *r/science*, and how new members in *r/science* adopt, or don’t adopt, these norms.

3.3 Measuring Language Differences

We conducted linguistic and quantitative analyses of 12 large subreddits to answer our overarching question of whether *r/science*’s potentially specialized language represents a barrier of entry. More precisely, our research questions are:

RQ1: Do posts and comments on *r/science* contain specialized language compared to other large or topically related subreddits? If so, what are the characteristics of this specialized language?

RQ2: How does the language of users who stay and leave differ in their posts and comments? If there is a defined difference, this would suggest that new contributors experience a language barrier, and those

Subreddit	# Posts	# Comments	# Subscribers
r/science	22, 157	604,267	21,015,665
r/news	254,201	5,878,711	17,972,696
r/politics	211,866	14,754,150	4,925,536
r/pics	101,624	3,806,068	21,313,781
r/funny	114,272	4,246,111	23,759,930
r/askreddit	1,096,947	35,665,797	22,153,598
r/askhistorians	41,203	72,056	998,325
r/everythingscience	6,772	36,901	165,852
r/futurology	20,127	799,176	14,037,974
r/truereddit	4,193	198,108	439,334
r/dataisbeautiful	8,437	420,762	13,608,622
r/askscience	12,162	184,249	17,901,914

Table 3.1: Number of posts, comments, and subscribers of each subreddit.

who join *r/science* with a strongly differing language are less likely to stay than those whose language more closely aligns with *r/science*'s.

3.3.1 Data

In addition to *r/science*, we selected 11 subreddits with the goal of comparing the language used in *r/science* to the language used in other large or topically related communities: *r/news, r/politics, r/pics, r/funny, r/askhistorians, r/futurology, r/truereddit, r/dataisbeautiful, r/askscience, r/everythingscience* and *r/askreddit*. While there is no such thing as "the average Reddit user" that we could compare against, our list includes some of the largest subreddits that share a similar subscriber count to *r/science* (*r/news, r/politics, r/pics, r/funny*, and *r/askreddit*) and those with the largest overlap in user participation with *r/science* (*r/askhistorians, r/futurology, r/truereddit, r/dataisbeautiful, r/askscience*, and *r/everythingscience*) [Martin, 2017]. All of these are subreddits with community members that *r/science* should have an interest in engaging, which is why our aim was to characterize how different the language experience is for a user in these subreddits compared to *r/science*.

For each subreddit, we collected all comments and posts from Jan. 2018 to Dec. 2018 using Google's Bigquery.[1] We ignored comments and posts that had been deleted by their original author, by moderators, or that were shorter than 10 words, for a total of 1,893,961 posts and 66,666,356 comments. Table 3.1 provides

[1] https://cloud.google.com/bigquery/

details of our dataset.

3.3.2 Analysis

RQ1: *r/science*'s specialized language We analyzed whether *r/science* uses specialized language by constructing language models trained on its posts and comments. A language model is a conditional probability distribution over each word in a vocabulary given preceding words (left context); the distributions are estimated using a training corpus of texts, which we denote $\mathcal{D}$.

A language model can be used to assign probability mass to a sequence of words by taking the product of conditional probabilities for individual words given their left contexts (i.e., applying the chain rule of probability). The probability that a language model trained on text dataset $\mathcal{D}$, which we denote $LM(\mathcal{D})$, assigns to a new piece of text depends very heavily on the training data $\mathcal{D}$. For example, training a language model on judicial decisions will likely lead to very low probability assignment for a Reddit post about astronomy, but higher probability for a similar-length legal brief.

Given a language model $LM(\mathcal{D})$, we can calculate how different a text (word sequence $\vec{w} = \langle w_1, w_2, \ldots, w_N \rangle$) is from the training data of LM by calculating $\vec{w}$'s **cross entropy** under LM:

$$CE(\vec{w}, LM(\mathcal{D})) = -\frac{1}{N} \sum_{i=1}^{N} \log p_{LM(\mathcal{D})}(w_i \mid w_{i-1}), \tag{3.1}$$

where $p_{LM(\mathcal{D})}(w_i \mid w_{i-1})$ is the "bigram" probability of word w_i given the preceding word w_{i-1} according to the language model $LM(\mathcal{D})$, and w_0 and w_N are special "start" and "stop" tokens included by convention.[2] The higher the cross entropy, the more divergent $\vec{w}$ is from $LM(\mathcal{D})$'s training data, the less likely it is under $p_{LM(\mathcal{D})}$, and hence the more "surprising" it is under the distribution of language model $LM(\mathcal{D})$.

Our language models are Katz-backoff bigram models with Good-Turing smoothing [Chen and Goodman, 1999]—a commonly used technique to improve probability estimates that past work in this space has employed [Zhang et al., 2017]—trained on posts and comments of active authors in *r/science*. All language models used as a vocabulary the words seen during training. We built our language models using the SRILM toolkit [Stolcke, 2002].

[2] While bigram models are a relatively simple choice, they suffice for our purposes and are relatively robust against overfitting for datasets of the size we consider here.

For comments, our language models were trained on 1000 comments, 5 comments sampled from 200 experienced commenters for each month. Following Zhang et al. [2017], we defined experienced commenters as those who had commented at least 5 times in a given month. For posts, the language models were trained on 1000 posts (5 posts sampled from 200 experienced posters for the entire year). We treated experienced posters as those who had posted at least 5 times in the year. This provided roughly the same percentage of users per year that our definition of experienced commenters did for a month. We constructed 100 language models for each month of comments and 100 language models total for the year's posts, resampling authors and their comments and posts each time. We used all language models in each cross entropy calculation for a post or comment, averaging all post or comment cross entropy within a language model.

Past work has identified that longer posts and comments tend to exhibit higher cross entropy, possibly due to the higher probability of esoteric language in longer posts or comments [Danescu-Niculescu-Mizil et al., 2013; Zhang et al., 2017]. To avoid these length effects, we followed past work [Zhang et al., 2017] and only used the first ten words of each comment or post for training and cross entropy calculations, inserting stop tokens at the end of these first ten words.[3]

With language models trained on *r/science* posts and comments, we analyzed whether text from other subreddits diverged from the text of *r/science* by calculating the cross entropy of text sampled from other subreddits and comparing it to text sampled from *r/science*. In particular, we calculated the cross entropy of 50 comments, 5 unique comments sampled from 10 experienced authors (different from those used to train the models), for each month from each subreddit using language models trained on *r/science* comments for that month. We then conducted an ANOVA and Tukey post-hoc tests to compare the cross entropy of *r/science* comments compared to the comments of other subreddits, averaged over all months. We conducted the same analysis for *r/science* posts, sampling over the entire year 12 times to match the number of month samples.

One limitation of using language models trained on *r/science* is that they will overfit to *r/science* posts and comments. This leads us to expect that they will find posts and comments of other subreddits surprising (i.e., have higher cross entropy). All a higher cross entropy means is that something in the text of *r/science* is different from other subreddits, but it doesn't signal what that something is. To characterize the actual

[3] We obtained qualitatively similar results using the entire span of each post or comment for training.

differences between the language used in *r/science*, we therefore additionally analyzed individual word frequencies using uni- and bigram language models for both posts and comments (separately). To do this we first estimated an empirical background frequency for each token (counted as a unigram or bigram) using all other subreddits:

$$\hat{p}_w = (c_w + \alpha) \Big/ \textstyle\sum_{i=1}^{V} (c_i + \alpha) \ , \tag{3.2}$$

where c_w is the observed count of vocabulary token with index w, V is the size of vocabulary, and α is a smoothing constant, which we set to 0.1.

To determine which tokens are unusually common or rare in *r/science*, we used a χ^2 test of independence with a Bonferroni correction. Among those that were significantly different, we identified the tokens that had the most improbably high or low observed counts by modeling them with independent binomial distributions for each token:

$$p\left(c_w^{(s)}\right) = \text{Binomial}\left(N^{(s)}, \hat{p}_w\right) , \tag{3.3}$$

where $c_w^{(s)}$ is the observed count of token w in *r/science* and $N^{(s)}$ is the total number of tokens (uni- or bigrams) in *r/science* (posts or comments).

To summarize these findings, we report the words whose observed counts had the lowest probability according to these models, separating them into those that were used more frequently and less frequently than would be expected. We also used a similar analysis to compare the frequency of words used by transient contributors, those who only contributed once, and experienced contributors within *r/science* comments.

To further evaluate how the language of *r/science* differs from the language of other subreddits, we built a text classifier using only uni- and bigram features to classify posts and comments as either in *r/science* or not. In contrast to the language models, which show how surprising the language of other subreddits is to *r/science*, a classifier will show how easily distinguishable *r/science* language is compared to those of other subreddits. Our classifier is a support vector machine (SVM) using uni- and bigram features. We use a term frequency inverse document frequency (tf-idf) transform to account for common words throughout all posts and comments. We trained two SVMs: one for posts and one for comments. We used one versus many classification, meaning the classifiers classify posts and comments as either in *r/science* or not. We report the F_1 score for both classifiers. Because the number of non-*r/science* posts and comments vastly outweigh the number of *r/science* posts and comments, they are sampled equally to have a balanced training and test set.

RQ2: Language used by transient vs. returning authors Past work has shown that new authors commenting or posting for the first time do not necessarily adopt the language norms of the community immediately [Danescu-Niculescu-Mizil et al., 2013] and are more likely to leave the community. This is know as an *acculturation gap* [Zhang et al., 2017], i.e., the difference in language between users who only ever posted or commented once (transient users) and returning authors, and is predictive of long term engagement [Danescu-Niculescu-Mizil et al., 2013]. To analyze whether there is a defined acculturation gap for *r/science* (which would indicate new users' difficulty in adapting to the specialized language of the community), we calculated the difference in the cross entropy of posts and comments by transient contributors (i.e., those who only contribute once) vs. experienced contributors. Following Zhang et al. [2017], we use experienced contributors as a proxy for community language. It is also important to note that transient users might have contributed to other subreddits or even read *r/science* posts; however, they had not posted or commented in *r/science*. We calculated the cross entropy of experienced contributors by sampling 5 comments for 50 experienced contributors, for a total of 250 comments. We then sampled 250 comments from transient contributors. The acculturation gap was the difference between these two cross entropies. We resampled experienced and transient contributors' comments for each month and 12 times on the entire year for posts. We followed this analysis with comparing the common and rare words used by transient users and experienced users in *r/science*. This allowed us a more nuanced perspective on not only whether the language differed between transient and returning users, but also how it differed.

While the acculturation gap is useful in identifying the distinctiveness of *r/science*'s language to first time users, we sought to further investigate whether language might act as a barrier to new users by examining if those who ultimately stayed in *r/science* matched the language of the community more closely in their first contribution than those who only contributed once and then left. If so, this would suggest that language is a factor for deterring users from engaging with *r/science* beyond one post or comment. Because we only looked at data for a single year, 2018, we were unable to determine whether users contributed before this

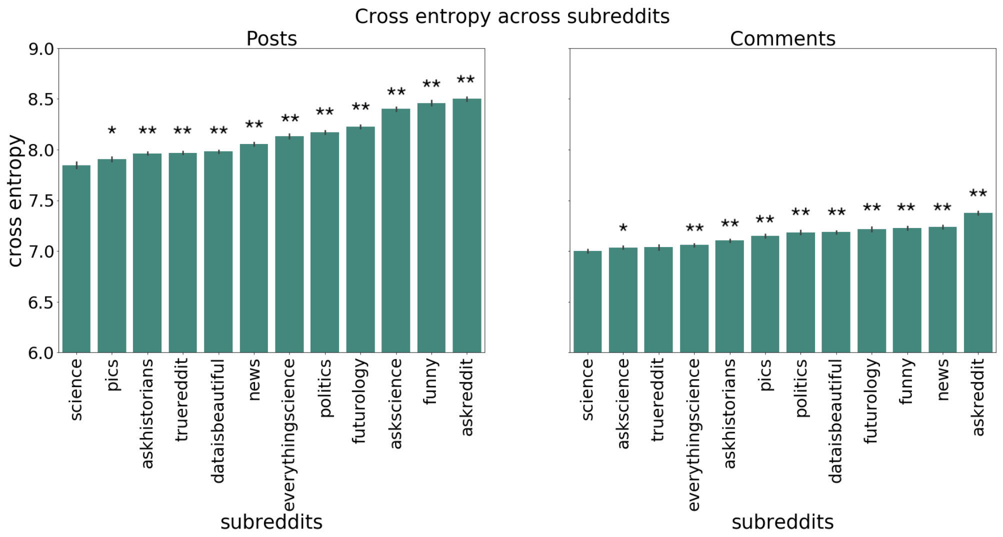


Figure 3.2: Cross entropies of posts and comments from experienced contributors from each subreddit, calculated using *r/science* language models. Bars show bootstrapped 95% confidence intervals. Asterisks denote t-test significance compared to *r/science*. Similar results were found using Mann-Whitney U tests for non-normal distributions. All results are corrected for multiple hypothesis testing using Bonferroni-Holm; $^{*}p < .05$, $^{**}p < .001$.

cutoff. We therefore ignored comments and posts from January and February for this analysis. Considering that over 70% of users contribute for only one month in *r/science*, it is unlikely that users who did not post in the first two months of 2018 were active before that.

We calculated the cross entropy of the 250 first posts or comments from 250 experienced authors, comparing this to 250 posts or comments from authors who only ever commented or posted once in the subreddit. We resampled for each month of comments and 12 times on the entire year for posts.

3.4 Results

RQ1: The *r/science* community uses specialized language compared to other subreddits. The cross entropy of comments and posts significantly differed across subreddits (posts: $F_{11,14388)} = 3995.661$, $p < .0001$, comments: $F_{(11,14388)} = 1615.158$), see Figure 3.2. *r/science* has the lowest cross entropy for both posts and comments, suggesting that there are unique language characteristics (e.g., words and phrases) in *r/science*'s posts and comments that do not occur in other subreddits. The difference holds even for those subreddits that are topically related (e.g., *r/askscience* and *r/everythingscience*).

	Especially Common	Especially Rare
Posts	Terminology (*cells, brain, cancer, disease*) Reporting (*study, researchers, scientists, new*) Hedge words (*may, according, suggests, likely*)	Pronouns (*you, your, it, youve, my, i, he, me*) Questions (*what, whats, if, why, how, who*) Opinions (*serious, best, worst, like, favorite*)
Comments	Terminology (*cells, cancer, species, energy*) Reporting (*study, studies, science, research*) Analysis (*factors, correlation, likely, link*)	Pronouns (*he, i, his, she, my, her, him, me*) Politics (*trump, mueller, clinton, hillary*) Profanity

Table 3.2: Summary of the most unusually common and rare words in *r/science* posts and comments, calculated with χ^2 test of independence based on the frequency of the words in *r/science* relative to other subreddit posts and comments, respectively.

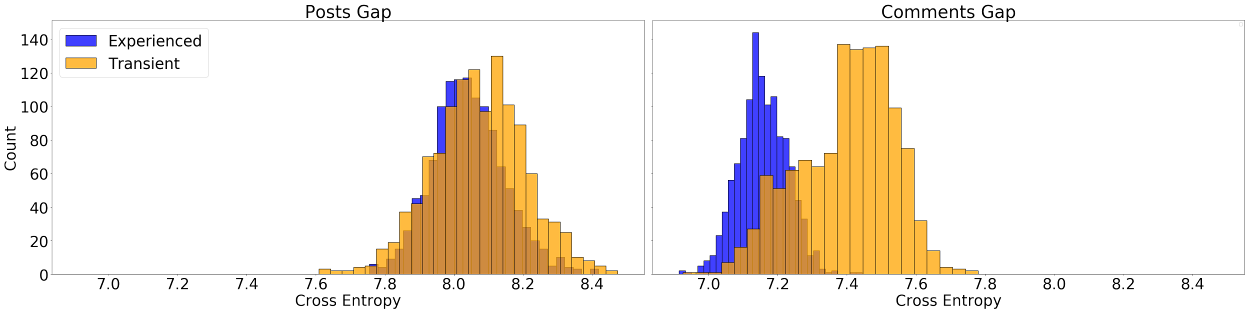


Figure 3.3: Sampled histogram of cross entropies between transient and experienced contributors (all posts and comments). Difference in comment cross-entropy is significant ($p < .001$) though not for posts (independent samples t-test, corrected for mutliple hypothesis testing using Bonferroni-Holm).

While it is not surprising that *r/science* has the lowest cross entropy since the language models were trained on it, it is interesting to see how other subreddits relate to *r/science* in post and comments. For example, post cross entropies are more distinct across subreddits than comments. This is due to posts containing more topic words (e.g., *Trump*, *cells*) than comment text. Interestingly, *r/pics* posts have a similar cross entropy to *r/science* posts. This may be because *r/pics* also has stringent guidelines for post titles that discourage personal words in the title (e.g., no memorial posts, no posts asking for assistance, and no personal information). While *r/politics* also has stringent post title guidelines, the strong topical differences between it and *r/science* most likely contributed to its higher cross entropy.

Table 3.2 summarizes the words and bigrams that comprise the most improbably common and rare terms in *r/science* compared to the other subreddits for both posts and comments (ignoring conjunctions and prepositions). Looking at these words, we can see that in *r/science* posts scientific terminology, references to scientific studies, and *hedge* words (e.g., *may*, *suggests*, *likely*) are all extremely common relative to other subreddits. By contrast, personal pronouns, question words, and expressions of opinions are extremely un-

	Transient Contributors	Experienced Contributors
Posts	*The definition of the kilogram might be about to change for the better!*	*Sulfur isotope has helped reveal surprising information about both the origins of life on Earth.*
	How Reddit (and the rest of the internet) is good (and bad) for you	*Negative experiences on social media carry more weight than positive interactions [...]*
Comments	*Same with adderall in my case. Whenever I'm on it im no longer constantly hungry*	*Yeah I would have wanted a control group just to confirm how fmri changes when you were just exposed to it.*
	I'm convinced that any mouse with a strong background in science could make itself immortal.	*Well, no, this would be enough to be revolutionary if you could build, say, MRI machines with it. It's much cheaper to run a fridge than to keep something chilled with liquid helium.*

Table 3.3: Posts and comments from transient and experienced users in *r/science*. Sampled from all contributions.

common. Similar patterns hold for *r/science* comments, but we did not observe such an extreme use of hedge words, and the most notably underused words (besides personal pronouns) are profanity and terms related to politics (which have a high background frequency in comments in other subreddits). Looking at the bigrams reveals similar findings: science posts and comments contain more references to scientific studies (e.g., *(researchers have)*, *(study finds)*) and hedge phrases (e.g., *(according to)*, *(likely to)*)) and fewer questions and personal references (e.g., *(what is)*, *(do you)*, *(when i)*. While there are topical differences in some word usage comparison (e.g., *Trump* as a common word outside of *r/science*), there are also many examples of stylistic differences (e.g., hedging and impersonal language) in words and phrases. This indicates that *r/science* differs in style and topic from other subreddits, especially with hedging and impersonal speech.

The classifiers achieved mixed results for identifying posts and comments as from *r/science* or not. The comment classifier obtained a test F_1-score of 0.73, while the post classifier reached 0.84. Similar to the cross entropy findings, the classifier scores suggest that posts are easier to differentiate across subreddits than comments. Considering the minimal features used to train both classifiers (simple uni and bigram features with a tf-idf transform), the scores reflect the distinctiveness of *r/science* posts, scoring well above random chance.

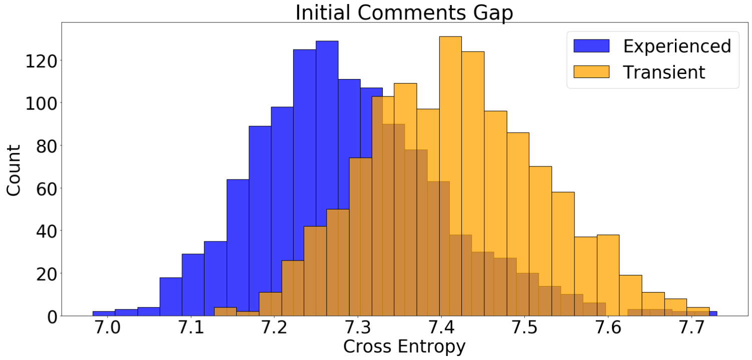


Figure 3.4: Sampled histogram of cross entropies between first time comments of users who leave and who stay. Difference is significant ($p < .001$).

RQ2: *r/science* users that don't match the community's language are more likely to leave. The majority (57%) of users in *r/science* only post or comment once and never return. We found a pronounced and significant difference in cross entropy of these transient authors versus experienced authors in *r/science* for comments (mean = 7.37, s.d. = 0.13, vs. mean=7.16, s.d. = 0.14) ($t_{1199} = 40.85$, $p < .0001$, $d = 1.67$) and posts (mean = 8.10, s.d. = 0.15, vs. mean = 8.05, s.d. = 0.16) ($t_{1199} = 8.07$, $p < .0001$, $d = 0.33$). Figure 3.3 plots the distributions of the cross entropies for posts and comments.

Looking at common words and phrases in these transient user comments, we found that personal words (e.g., *i*, *my*, *feel*) are significantly more common in transient user comments than experienced user comments of *r/science*, and words discussing scientific findings (e.g., *abstract, journal, evidence*) are significantly rarer. The common and rare bigrams for transient users reflect similar differences, with personal and anecdotal phrases (e.g., *(i was)*, *(when i)*) common while references to scientific findings (e.g., *(linked academic)*, *(press release)*) rare. These results mirror those found between *r/science* and other subreddits (see Table 3.2) suggesting that users from other popular subreddits, while possibly matching the language in these subreddits, are faced with a more pronounced language barrier in *r/science*. Table 3.3 provides examples of posts and comments from transient and experienced contributors on *r/science*.

r/science also stands out as having lower new user retention than the majority of other subreddits (see

Figure 3.5), with an average of 10% (s.d. = 3.28%) of new users returning after posting or commenting for the first time in the previous month ($F_{(11,120)} = 13.818$, $p < .0001$). Interestingly, many of the subreddits related to *r/science*, such as *r/askscience* and *r/everythingscience*, have similarly low retention.

Our methods for measuring language differences did not allow quantitative comparisons of differences between transient and returning authors in *r/science* compared to other subreddits since this would have involved comparing significance values and cross entropies calculated from different language models, both of which are improper comparisons. We instead ran our word frequency tests on comments of transient users from subreddits topically related to *r/science* and with similarly low user retention: *r/askscience* and *r/everythingscience*. If the common and rare words fell into similar categories as in *r/science* (e.g., personal words and scientific findings) for transient contributors, this would suggest that the low retention in these communities is related to new users failing to adapt to the same specialized language.

Common words for transient users in *r/askscience* and *r/everythingscience* included more personal words words (e.g., *i*, *my*, *your*), while scientific words (e.g., *species*, *particles*, *genetic*) were rare for transient contributors. However, there were also noticeable differences in these words compared to *r/science*: *r/askscience* contained many more question words (e.g., *what*, *please*, *thank*), which makes sense considering the purpose of the subreddit is to ask questions. These differences fall along the differences in the purposes of the subreddits, while the similarities between these subreddits (a focus on scientific terminology and away from personal words) suggests that this type of language is more difficult for the general Reddit user to adapt to, possibly contributing to the lower user retention they share.

These differences suggest that transient users in *r/science*, those who only post or comment once, use significantly different language than those who are returning contributors in the community. To delve deeper into this difference, we explored how the language of the first post or comment of contributors in *r/science* who would return differed from the posts or comments made by transient users.

Users who end up becoming experienced contributors of *r/science* matched the language in *r/science* in their first comment more closely than those who only contributed once and then left (mean = 7.30, s.d. = 0.13 for experienced users vs. mean = 7.40, s.d. =0.11 for transient users) ($t_{1199} = 19.03$, $p < .0001$, $d = 0.78$). Figure 3.4 plots this difference, showing similar distributions as found in Figure 3.3 for comments. We did not find this to be the case with posts; the cross entropy of experienced users' first posts was not

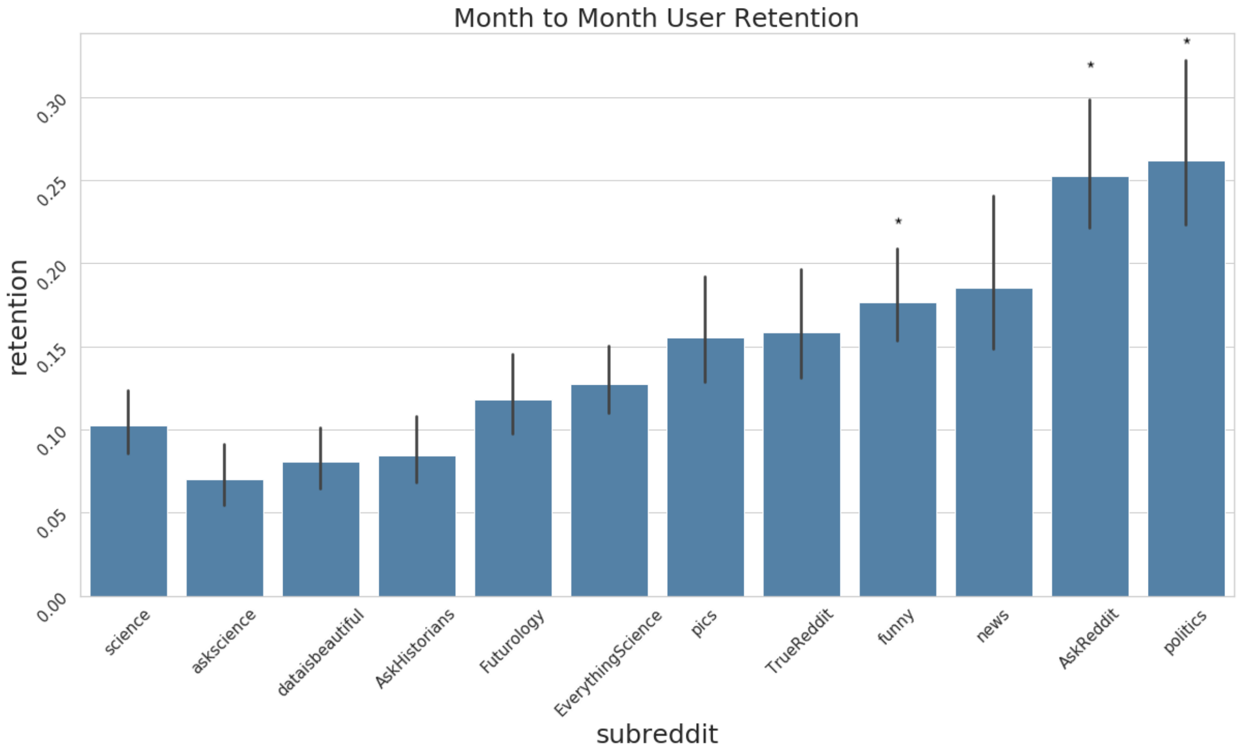


Figure 3.5: New user retention across all subreddits. $*p < .05$ independent samples t-test compared to *r/science*. Corrected for multiple hypothesis testing using Bonferroni-Holm.

significantly lower than the cross entropy for transient users' posts. Considering that posting is speaking to all of *r/science*, this is probably due to people focusing more on what they are writing – and following *r/science*'s posting rules more closely – when they first post compared to when they first comment. These results show that users who leave after commenting once diverge from the language of *r/science* significantly more than users who ultimately stay, suggesting that language is a factor for deterring some users from commenting in *r/science*.

3.5 Summary

As an online community for discussing research findings, *r/science* has the potential to contribute to science outreach and communication with a broad audience. Yet in this chapter we show how the language used by people on *r/science* can be a barrier to entry for potential members. Our analysis of the language used in more than 68 million posts and comments from 12 subreddits shows that *r/science* uses a specialized language that is distinct from other subreddits. Transient authors of posts and comments on *r/science* use less specialized language than more frequent authors, and those that leave the community use less specialized language than those that stay, even when comparing their first comments. These findings suggest that the specialized language used in *r/science* has a gatekeeping effect, preventing participation by people whose language does not align with that used in *r/science*. More broadly, we found that language can prevent engagement for some, but not for others, indicating the need to for methods that can adapt language to certain audiences.

One interpretation of our results is that the community rules are effective and functioning as intended. First, *r/science*'s specialized language is clearly geared toward maintaining rigorous scientific discussion in the community Jones et al. [2019]. Second, it could be that the language of *r/science* promotes stability in readership and trust among passive users; by having a high bar for who gets to talk, people feel most of what they hear is worth listening to. This can encourage a readership that lurks but does not contribute, trusting in those who speak the language to contribute to the discussion. We see some evidence of this in the massive number of subscribers in *r/science* (over 21 million, fifth overall on Reddit at the time of writing) versus an order of magnitude fewer posts and comments compared to other subreddits with similar subscriber counts.

However, another perspective is that there are opportunities to help broaden participation and socialize newcomers. While it may not be necessary for everyone to have the same knowledge about or interest in science, science communication should be accessible to everyone Burns et al. [2003]. Our findings indicate that simply posting community rules on the front page has not been sufficient in guiding first time and transient contributors to use the specialized language expected on *r/science*. The *r/science* community can use these findings and our methodology to explore ways of socializing newcomers without sacrificing community norms, such as by identifying and defining specialized terms in the guidelines or moderator posts.

Chapter 4

Writing Strategies for Science Communication

In this chapter, we identify the ways that expert science writers design scientific language for a different audiences as a first step in understanding how to automatically adjust language for different audiences. Much of the specialized language in *r/science* is indicative of hurdles faced by readers in science communication generally: scientific terminology and complex language can make scientific writing hard to understand for many people [Britt et al., 2014a]. Expert science writers have years of experience communicating this information with different audiences. We compile a set of strategies from a wide range of prescriptive science writing sources in English and develop an annotation scheme allowing humans to recognize these strategies in texts about science. We introduce a new corpus of 128K university press releases, science blogs, and science magazines and annotate a subset of 337 texts. We use the annotations to train transformer-based classifiers to explore the communicative goals of science writing by analyzing variations in the strategies' use across several scientific communication forums. This chapter includes materials originally published in August et al. [2020b].

4.1 Introduction

Communicating scientific discoveries to a general audience of readers is difficult. A researcher or writer interested in doing so is faced with the challenging task of translating complex scientific ideas in an engaging manner without misleading or overwhelming their audience. There are many guides to science communication [e.g., Blum et al., 2006], but they rarely offer empirical evidence for how their advice is used, or proven effective, in practice. The potential science communicator is then confronted with the additional hurdle of understanding how to implement these guidelines in their writing.

Effective science communication requires understanding the unique needs and expectations of different audiences and stakeholders in science [Nisbet and Scheufele, 2009b]. We envision natural language processing technologies that help science writers communicate more effectively. These technologies might automatically classify common strategies in a writer's own text, support writers to adapt language to specific readers, or guide readers through personalized article recommendations.

As a first step, we compile a set of strategies from a wide range of prescriptive science writing sources in English and develop an annotation scheme allowing humans to recognize these strategies in texts about science. We introduce a new corpus of 128K university press releases, science blogs, and science magazines and annotate a subset of 337 texts. We use the annotations to train transformer-based classifiers to explore the communicative goals of science writing by analyzing variations in the strategies' use across several scientific communication forums.

Our paper is the first computational analysis of writing strategies driven by science communication theory. We find that most strategies are prevalent throughout our corpus and that publication venues with varying audiences use strategies differently. For example, press releases emphasize the impacts of science more than magazine articles, which instead tell more stories about the science. We also find that higher quality newspaper articles, as rated by expert journalists, use more storytelling and analogies than lower quality articles. This suggests that (1) such strategies are a meaningful way of communicating scientific language for a general audience and (2) that there is not one single general audience: different venues attract different types of readers and thus communicate with these strategies differently.

4.2 Defining Science Communication Writing Strategies

The goal of general science communication is to increase public awareness, enjoyment, interest, and understanding about science [Burns et al., 2003]. Based on the idea of compositionality in discourse theory [Bender and Lascarides, 2019], we can think of the communicative intent of science writing as being made up of smaller communication goals represented in particular passages of an article [Grosz and Sidner, 1986; Louis and Nenkova, 2013a]. Our computational approach builds on this theoretical assumption by annotating sentences and letting an article inherit the attributes we find in its sentences.

Past work on science communication has taken a similar view [Louis and Nenkova, 2013a] by using syntactic relations to characterize an article's communicative goals, allowing them to emerge inductively rather than from a theory of science communication. Our complementary approach starts with science communication guides to construct theory-driven communicative goals (referred to as "writing strategies" and consisting of lexical to multi-sentence features), and explores their use in a diverse range of science communication text.

To define our writing strategies, we categorized and grouped advice from style guides for science communicators. These guides were a mix of online resources, books, and academic articles (see Table A.1 in the appendix). We selected the guides based on discussions with three expert science communicators at a large research university's press department and through online searches. We stopped adding resources when we reached saturation [Holton, 2007], meaning that each new resource had fewer new strategies and suggesting that our resources provided good coverage.

Two authors open-coded [Holton, 2007] the suggestions from each guide by assigning each piece of advice in a resource a code that represented its high-level strategy (such as "avoid jargon"). The authors then looked at other resources to see whether the same advice appeared there. Each new piece of advice was added with a new code. After coding all resources, the authors grouped the codes into a set of 10 suggested writing strategies. Appendix A.1 provides additional details on the coding and categorization. The strategies are as follows (examples of each are given in Table A.5 in the appendix):

LEDE A few sentences at the beginning of an article, called a lede (spelled "lede" for easier differentiation with its homograph "lead"), that draws a reader in and makes them want to read more.

MAIN Sentences describing the main findings being reported by the original paper in order to not overwhelm the reader with details.

IMPACT Writing about the real world impact of the science or findings being reported in order to excite readers. This can include future technologies, breakthroughs the findings might enable, or their societal implications.

EXPLANATION Explanations about scientific subjects to improve reader understanding. This could be explaining a certain topic or word, or what researchers did in a study and what the findings mean.

ANALOGY Analogies or metaphors used as a way to explain concepts or make ideas in the article more relatable.

STORY Stories to engage readers and make the reported science more interesting. This can include short story snippets, or coming back to an underlying story throughout an article.

PERSONAL Including personal details about researchers in order to make them more approachable and add depth to the story.

JARGON Avoiding specialized terminology or jargon as much as possible as it can overwhelm readers.

ACTIVE Using the active voice to make the writing more lively and engaging.

PRESENT Similar to ACTIVE, using present tense verbs also to make the writing more lively and engaging.

Some of these strategies are specific to science writing, such as emphasizing the real world impact of the findings (IMPACT), while others are often thought of as general rules for good writing, such as using the active voice (ACTIVE). Both types of strategies were commonly referenced in the resources we analyzed, which suggests that engaging science writing shares traits with engaging writing in other disciplines while also containing its own set of unique strategies.

4.3 Dataset

In order to study the use of these strategies by science writers and build classifiers for automatic identification, we collected a corpus of 128K documents from a variety of science communication sources. We focused on four major types of U.S.-based venues, representing a broad spectrum of science communication for different audiences: blog sites, popular science magazines, university press releases, and scientific journal magazines.

Past work has shown that blog sites usually write to scientifically literate and engaged readers [Ranger and Bultitude, 2016], while university press releases often write to other science journalists [Bratton et al., 2019; Sumner et al., 2014b]. We selected popular science magazines since they target a more general audience, and scientific journal magazines as they often write to those involved in research, though not necessarily in the same domain [Nielsen and Schmidt Kjærgaard, 2011]. The choices of website or publication we collected from each venue category were based either on previous work covering those categories [e.g., blog posts; Vadapalli et al., 2018] or as a convenience sample based on what was widely available. One note is that while past work has used the blogs sites we selected as sources for high quality science blogs (*sciencedaily.com* and *phys.org*), these sites also source a large portion of their content from press releases, often only changing headlines and lede sentences.

We scraped articles from each of these sources for all of 2016–2019 using the Wayback Machine,[1] resulting in 137,828 articles. Appendix A.2 provides more details on site selection.

4.4 Filtering

To focus on science communication, specifically, we removed articles matching U.S.-centric political keywords such as *Trump*, *democrats*, and *Senate*. We also removed all articles over 15,000 or under 1,500 characters, since these represented either multiple articles on the same page, article previews, or scraper errors. After filtering we had a total of 128,253 articles. In total 7% of documents were filtered (3.5% removed for political keywords and 3.5% for length). Table 4.1a details the sites for each venue and the number of articles after filtering.[2]

4.5 Annotation

Recall that our goal was to measure the use of strategies from Section 4.2 in our corpus. We sample 337 articles stratified across sites to gather a spread of articles and balance the articles across venues. Each

[1] https://archive.org/web/

[2] URLs for all scraped articles are available at https://github.com/talaugust/scientific-writing-strategies.

article was given to two annotators who were trained on the writing strategies and instructed to annotate the use of strategies at the sentence level. Concretely, each "annotation" corresponded to a contiguous chunk of one or more sentences labeled with one of the seven strategies. A sentence could be labeled with multiple strategies. Three of the strategies, JARGON, ACTIVE, and PRESENT, were not annotated because we believed they could be reliably detected using existing methods based on lexical and syntactic features; see Section 4.7.1.[3] Figure A.2 (in the appendix) presents an example of an annotated excerpt and the task interface.

We conducted annotation in sets of 50 articles. After each set, one author measured agreement and manually evaluated a subset of annotations by both annotators. This author then acted as a coordinator for the annotators, providing suggestions or revisions to annotation guidelines. Additionally, annotators were able to look at the other's annotations after completing an article. Figure A.1 in the Appendix plots Krippendorff's α after each batch of 50 articles. Two categories that achieved low agreement (as measured by α) were EXPLANATION and PERSONAL, which we dropped from further analysis.

Venue	**Site**	**#Articles**
Blogs	*Sciencedaily.com*	38,191
	Phys.org	74,732
Magazines	*The Atlantic*	2,771
	Scientific American	5,174
Press releases	*Harvard*	752
	Stanford	599
	Rochester	219
	Northwestern	197
Journals	*Science*	4,173
	Nature	1,445
Total		128,253

(a) Sites and number of articles in corpus after filtering.

Strategy	**# Sentences**	**Avg. words**
LEDE	595	34.1±20.6
MAIN	1,596	35.9±19.1
IMPACT	1,102	40.0±19.1
EXPLANATION	4,677	55.0±28.1
ANALOGY	410	33.6±17.3
STORY	1,736	74.4±60.9
PERSONAL	727	50.5±33.1
Total	10,843	48.7±32.7

(b) Number of sentences and average number of words per annotation span for each strategy across articles in our annotated dataset (337 articles).

[3] Code for the interface is available at `https://github.com/talaugust/scientific_article_annotation`.

4.6 Hypotheses

Our annotated corpus allowed us to begin to explore how strategies relate to the communicative goals of different science communication venues. To do this, we introduce hypotheses informed by existing literature.

Hypotheses **H1–H3** are based on our expectations for how strategies can differentiate science writing venues in our corpus based on their underlying goals. For these hypotheses, we evaluated strategy use across our corpus.

Hypothesis **H4** builds on past research in science communication exploring article quality [Louis and Nenkova, 2013b] which is introduced with its own annotated dataset. Since we had no strategy annotations for this dataset, we report only on the aggregated predictions of our classifiers.

H1: LEDE is used once or twice within an article, but consistently across our entire corpus. Because the LEDE strategy is well adopted in general journalism [Pöttker, 2003], and LEDE sentences are only used at the beginning of an article, we expect low but consistent use of LEDE across the corpus.

H2: Press releases use higher IMPACT than other venues. One goal of press releases is to encourage other science writers to pick up a story. Because a key component to selecting stories is impactful findings [Hayden et al., 2013], we expect that press releases will emphasize this more.

H3: Magazines use lower JARGON, higher ACTIVE and PRESENT and higher STORY than other venues. Magazines target a broader readership compared to other venues, making it likely they use these strategies that are common in prescriptive guides for general good writing [Strunk, 2007] to relate to a wider audience.

H4: Higher quality science news articles use more STORY and ANALOGY than lower quality articles. Past work on science news quality has suggested that features related to storytelling and figurative language (e.g., coherence and descriptive language) are associated with higher quality articles [Louis and Nenkova, 2013b].

4.7 Strategy Classification

We used our collected corpus and annotations to automate recognition of writing strategies and to evaluate our hypotheses. We describe our methods for classifying strategies with rules and with human annotations (Sections 4.7.1 and 4.7.2). We then discuss methods for using these classifiers to estimate the use of strategies in our corpus and overall classifier performance (Section 4.7.3).

4.7.1 Rule-Based Strategies

As discussed earlier, three of the strategies could be reasonably identified using rules, and were not annotated.

JARGON We used common science jargon word lists drawn from Rakedzon et al. [2017] and Gardner and Davies [2013] to detect jargon use. The word list from Rakedzon et al. [2017] consists of 2,949 words common in scientific journal abstracts and articles while rare in common usage. We augmented this list with the core Academic Vocabulary List [AVL, Gardner and Davies, 2013], which is the top 3,000 word lemmas based on 120 million words of academic texts from the Corpus of Contemporary American English [Davies, 2009]. High JARGON means higher use of these specialized terms, which is negatively associated with the strategy (i.e., since the recommended strategy is to avoid specialized terms).

ACTIVE We identified active and passive clauses by counting the 'nsubj' and 'nsubj:pass' words from a parse of each article using Stanford NLP's dependency labels in the Stanford NLP Pipeline.[4] We normalized all active clauses by the number of verbs in an article.

PRESENT For measuring present tense, we normalized the number of present tense verbs using Stanford NLP's Universal Features, which are similar to POS tags and part of the same Stanford NLP Pipeline [Manning et al., 2014], over all verbs in an article.

4.7.2 Sentence Classifiers

For the remainder of our strategies, we trained classifiers based on the annotations collected to estimate the prevalence of each strategy in our corpus. Each classifier takes a single sentence as input and provides a binary label (present or absent) for a given strategy. Apart from pretraining, the classifiers were trained

[4] `https://stanfordnlp.github.io/stanfordnlp/pipeline.html`

Strategy	Prec.	Recall	F_1	Calibr. Err.	Acc.	MFC Acc.
LEDE	$.31_{.02}$	$.56_{.03}$	$.40_{.02}$	$.05_{.004}$	$.95_{.002}$	.95
MAIN	$.38_{.04}$	$.51_{.03}$	$.43_{.03}$	$.11_{.01}$	$.86_{.005}$	.86
IMPACT	$.40_{.03}$	$.55_{.03}$	$.46_{.02}$	$.09_{.01}$	$.91_{.004}$	.90
ANALOGY	$.52_{.04}$	$.60_{.02}$	$.55_{.03}$	$.04_{.01}$	$.96_{.001}$	.95
STORY	$.38_{.03}$	$.49_{.04}$	$.43_{.02}$	$.12_{.01}$	$.84_{.009}$	.84

Table 4.2: Mean and standard deviation of strategy classifier performance, including calibration error, on held-out test set based on 5 random seeds. Baseline accuracy for most frequent class (MFC), which always predicts the negative class, is included.

separately for each strategy. We based our classifiers off RoBERTa [Liu et al., 2019b] as it is a high-performing contextual word representation learner that has achieved state-of-the-art results on multiple NLP benchmarks, and which comes pretrained. We used Huggingface's RoBERTa implementation.[5]

We started by continuing to pretrain RoBERTa on additional in-domain text to tailor the model more closely to our task. This additional pretraining followed two phases as in Gururangan et al. [2020]: pretraining on 11.90M general news articles from REALNEWS [Zellers et al., 2019] for 12.5K steps (domain-adaptive pretraining), and then pretraining on a held-out subset of 100k documents from the unannotated portion of our own corpus for 10 epochs (task-adaptive pretraining). Appendix A.5 includes details for both pretraining steps.

Finally, we finetuned our pretrained RoBERTa model on each sentence-level classification task separately, making 5 binary classifiers (LEDE, MAIN, IMPACT, ANALOGY, STORY). Our pretrained RoBERTa models were finetuned using a 80%, 10%, 10% train, validation, test setup using all annotated articles. Articles were randomly split across the sets, meaning no two sentences from the same article could occur across sets. Appendix A.6 includes additional finetuning details.

Using classifiers optimized for individual classification can lead to biases when estimating category proportions [Hopkins and King, 2010]. Past work has suggested that using a **well-calibrated** classifier leads to better proportion estimation in large unlabeled corpora [Card and Smith, 2018]. Calibration refers to the long-run accuracy of predicted probabilities, where a well-calibrated probabilistic classifier at the level μ is one that predicts class k with probability μ when the proportion of instances correctly assigned to k is also μ.

[5] `https://huggingface.co/transformers/model_doc/roberta.html`

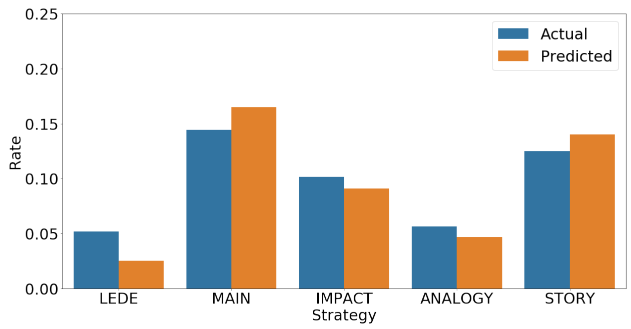


Figure 4.1: The rate of occurrence of predicted versus actual strategies on our human-annotated test set based on PCC and the classifiers.

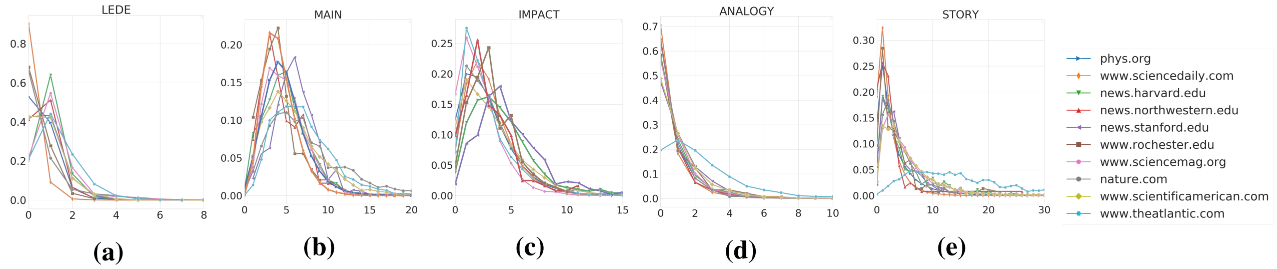
Following Card and Smith [2018], we performed model selection based on calibration error on held-out data during hyperparameter tuning. We estimated calibration error using the adaptive binning procedure from Nguyen and O'Connor [2015]. After picking our most well-calibrated classifiers, we measured the rate of each strategy across a collection of documents by averaging the classifiers' predicted posterior probabilities of a positive label. This is referred to as **Probabilistic Classify and Count** [PCC; Bella et al., 2010] and is a standard method for predicting label distributions in a corpus using a probabilistic classifier [Card and Smith, 2018].

4.7.3 Evaluation

We evaluated our classifiers in two ways: on individual examples (i.e., for reporting F_1), and in aggregate on a held-out annotated test set.

Our goal for the classifiers was to estimate aggregated proportions in our corpus, not to achieve perfect performance. For this reason, we report on classifier F_1 performance only to establish that the classifiers were reasonably able to detect strategies. Table 4.2 details the precision, recall, F_1 scores, average calibration error and accuracy of the trained classifiers, and baseline accuracies for the most frequent class predictions for each strategy. Our classifiers achieved F_1 scores between 0.40 and 0.55, which was comparable to other classifications of communicative goals, such as propaganda technique detection [e.g., F_1 scores between 0.39 and 0.61 in Da San Martino et al., 2019].

Because we were most interested in estimated proportions, we also compared the classifiers' predicted strategy rates in our held-out test set with the actual rates of the annotated strategies. Actual strategy rate was calculated as the number of sentences containing a strategy divided by the total number of sentences in the test set. Figure 4.1 illustrates this comparison. While we did see some discrepancies between actual rates and our predicted rates, these differences were small ($< .05$ rate difference, or less than 5% of sentences)



(a) **(b)** **(c)** **(d)** **(e)**

Figure 4.2: Histogram of number of sentences per article (x-axis) estimated to use each strategy (proportions on y-axis). Two categories that achieved low agreement (as measured by α) were EXPLANATION and PERSONAL, which we drop from this analysis. For an enlarged figure, see Figure A.4 in the appendix. Figure A.3 plots the estimated proportion of sentences using each strategy for all sites, including EXPLANATION and PERSONAL.

and the trend of each strategy remained the same (e.g., STORY and MAIN are the most common, LEDE and ANALOGY are the least), suggesting that the classifiers estimate strategy rates with sufficiently high accuracy to begin comparing rates across strategies.

We additionally evaluated how accurate our automatic measures for JARGON, ACTIVE, and PRESENT were by randomly sampling 5 sentences from the top and bottom 10% of articles containing JARGON, ACTIVE, and PRESENT (as measured by our rule-based approaches) and manually inspecting them for correct word classifications. The rules for each measure were in line with our intuitions about JARGON, ACTIVE, and PRESENT with a large majority of words (Over 80% in the 30 sentences evaluated) being identified correctly as jargon, active voice, or present tense.

4.8 Evaluating Strategy Applications

Evaluating our classifier output against gold-standard human annotations, as reported in Section 4.7.3, establishes the validity of our classifiers. We next turn to our hypotheses introduced in Section 4.6 to illustrate how we can use the strategies, classifiers, and corpus to explore the communicative goals of science writing. We introduce each hypothesis and report on its results separately.

H1: LEDE is used once or twice within an article, but consistently across our entire corpus. Figure 4.2a plots the estimated number of LEDE sentences per article across each site. Supporting **H1**, the majority of sites peaked at either 0 or 1 LEDE sentences, with all sites tapering off quickly after that. *theat-*

lantic.com does have a higher number of predicted LEDE sentences (with 20% of articles containing more than 2 sentences). This might be due to *theatlantic.com* articles being longer (since they are full magazine articles) and therefore using more text to entice readers.

H2: Press releases use higher IMPACT than other venues. We find support for **H2**: press release sites like *news.harvard.edu*, *rochester.edu*, and *news.stanford.edu* had larger modes than other sites for IMPACT sentences in Figure 4.2c. For example, close to 15% of articles in *rochester.edu* had over 5 estimated IMPACT sentences, compared to 7 or 8% of articles in *scientificamerican.com* or *theatlantic.com* having that same number. This is especially striking because *scientificamerican.com* and *theatlantic.com* generally had much longer texts, since they are full magazines, compared to press releases.

H3: Magazines use lower JARGON, higher ACTIVE and PRESENT and higher STORY than other venues. Texts from *theatlantic.com* and *scientificamerican.com*, the two magazine sites, had the lowest and third lowest use of JARGON in the corpus with average rates below 0.2 (i.e., less than 20% of words), macro-averaged across articles. Magazines also had the highest use of ACTIVE and some of the highest PRESENT. Additionally, *theatlantic.com* was the only site to have close to 5% of articles estimated to have more than an 15 STORY sentences (Figure 4.2e).

H4: Higher quality science news articles use more STORY and ANALOGY than lower quality articles. To evaluate this hypothesis, we used the corpus of *New York Times* science articles introduced by Louis and Nenkova [2013a].[6] The corpus consists of three labels of article quality: TYPICAL, VERY GOOD, and GREAT. These labels were drawn from whether the article appeared in that year's "Best American Science Writing" anthology (GREAT), was written by an author whose work had appeared in the year's anthology (VERY GOOD), or was neither (TYPICAL). The articles were drawn from the *New York Times* annotated corpus [Sandhaus, 2008] and filtered for science-related keywords (e.g., biology, biologist).

For a clear differentiation of article quality, we applied our strategy classifiers to only the GREAT and TYPICAL articles in the dataset. We selected science articles from the years 2001 to 2007 for a total of 55 GREAT articles (6,211 sentences) and 15,532 TYPICAL articles (1,079,768 sentences).[7] To test for signifi-

[6] Available at `http://www.cis.upenn.edu/~nlp/corpora/scinewscorpus.html`

[7] We obtained similar results when uniformly sampling 55 TYPICAL articles.

cance we performed χ^2 tests of independence and augment these with the ϕ coefficient, which is similar to Cohen's d as an effect size calculation for categorical variables [Fleiss, 1994].

Results: Sentences in GREAT articles used STORY and ANALOGY slightly but significantly more often than TYPICAL articles (STORY: 0.38 vs. 0.33, $p < 0.001$, $\phi = 0.01$, ANALOGY: 0.05 vs. 0.03, $p < 0.001$, $\phi = 0.01$), supporting **H4**.

4.9 Summary

Writing strategies that can be automatically recognized could greatly support science communication efforts by enabling tools to detect and suggest strategies for writers or automatically adapting text to different readers. In this chapter we compiled writing strategies from theory and practical advice, collected a large corpus and annotated a subset of it to measure strategies' use. We observed how strategies covary with intended audience. For example, blog sites, which target researchers, use more jargon and focus on the main findings of a paper, while magazine articles, which target a much broader audience of readers, tell more stories and use more active voice. Our findings also suggest that science newspaper articles judged by experts to have higher quality use more metaphorical language and tell more stories. We expect that our strategy formulations, classifiers, annotations and dataset will enable NLP-powered tools to support effective science communication for different audiences.

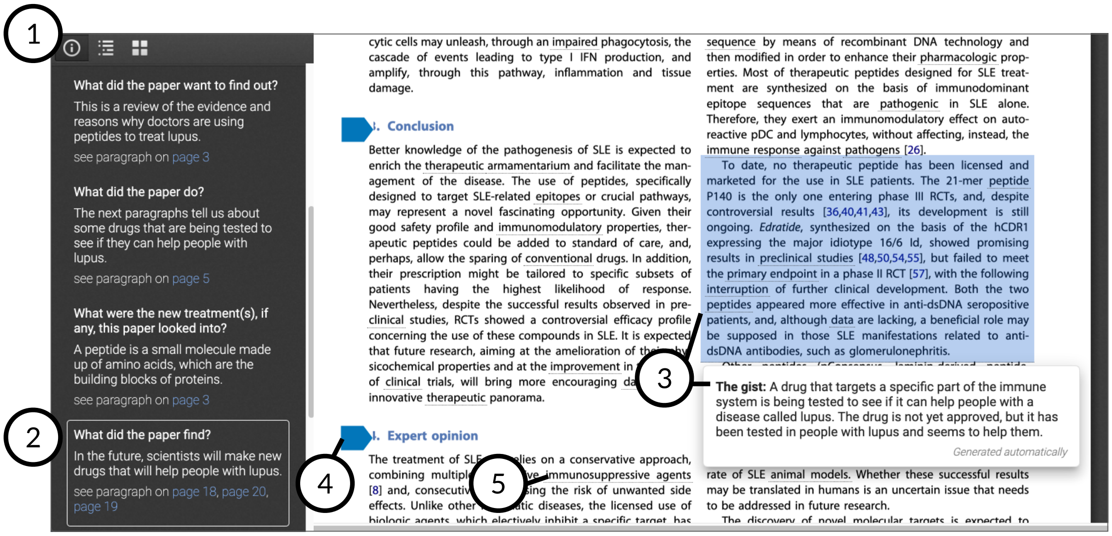
Chapter 5

PAPER PLAIN: Making Medical Research Papers Approachable to Healthcare Consumers with NLP

The writing strategies we identified in Chapter 4 highlight ways to adapt scientific language to a general audience. We take inspiration from these strategies by develop an augmented reading interface—PAPER PLAIN—for making medical research papers approachable to a general audience. PAPER PLAIN leverages strategies we identified to provide key questions that guide readers to areas of relevant information in a paper and assist them in making sense of complex medical language. We specifically focus here on healthcare consumers, such as patients, their families, friends, and other caregivers, staying apprised of the latest research in order to better manage their care and treatment options. This chapter includes materials originally published in August et al. [2022b].

5.1 Introduction

The latest medical knowledge often appears solely in the medical research literature [Zuccala, 2010; National Institutes of Health, 2005; Tennant et al., 2016; Day et al., 2020; Epstein, 1997]. For healthcare consumers, like patients, their families, friends, and other caregivers, staying apprised of the latest research



What did the paper want to find out?

This is a review of the evidence and reasons why doctors are using peptides to treat lupus.

see paragraph on page 3

What did the paper do?

The next paragraphs tell us about some drugs that are being tested to see if they can help people with lupus.

see paragraph on page 5

What were the new treatment(s), if any, this paper looked into?

A peptide is a small molecule made up of amino acids, which are the building blocks of proteins.

see paragraph on page 3

What did the paper find?

In the future, scientists will make new drugs that will help people with lupus.

see paragraph on page 18, page 20, page 19

cytic cells may unleash, through an impaired phagocytosis, the cascade of events leading to type I IFN production, and amplify, through this pathway, inflammation and tissue damage.

3. Conclusion

Better knowledge of the pathogenesis of SLE is expected to enrich the therapeutic armamentarium and facilitate the management of the disease. The use of peptides, specifically designed to target SLE-related epitopes or crucial pathways, may represent a novel fascinating opportunity. Given their good safety profile and immunomodulatory properties, therapeutic peptides could be added to standard of care, and, perhaps, allow the sparing of conventional drugs. In addition, their prescription might be tailored to specific subsets of patients having the highest likelihood of response. Nevertheless, despite the successful results observed in preclinical studies, RCTs showed a controversial efficacy profile concerning the use of these compounds in SLE. It is expected that future research, aiming at the amelioration of their physicochemical properties and at the improvement in of clinical trials, will bring more encouraging da innovative therapeutic panorama.

4. Expert opinion

The treatment of SLE relies on a conservative approach, combining multiple ve immunosuppressive agents [8] and, consecutiv sing the risk of unwanted side effects. Unlike other atic diseases, the licensed use of biologic agents which electively inhibit a specific target has

sequence by means of recombinant DNA technology and then modified in order to enhance their pharmacologic properties. Most of therapeutic peptides designed for SLE treatment are synthesized on the basis of immunodominant epitope sequences that are pathogenic in SLE alone. Therefore, they exert an immunomodulatory effect on autoreactive pDC and lymphocytes, without affecting, instead, the immune response against pathogens [26].

To date, no therapeutic peptide has been licensed and marketed for the use in SLE patients. The 21-mer peptide P140 is the only one entering phase III RCTs, and, despite controversial results [36,40,41,43], its development is still ongoing. *Edratide*, synthesized on the basis of the hCDR1 expressing the major idiotype 16/6 Id, showed promising results in preclinical studies [48,50,54,55], but failed to meet the primary endpoint in a phase II RCT [57], with the following interruption of further clinical development. Both the two peptides appeared more effective in anti-dsDNA seropositive patients, and, although data are lacking, a beneficial role may be supposed in those SLE manifestations related to anti-dsDNA antibodies, such as glomerulonephritis.

The gist: A drug that targets a specific part of the immune system is being tested to see if it can help people with a disease called lupus. The drug is not yet approved, but it has been tested in people with lupus and seems to help them.

Generated automatically

rate of SLE animal models. Whether these successful results may be translated in humans is an uncertain issue that needs to be addressed in future research.

The discovery of novel molecular targets is expected to

Figure 5.1: When a reader opens a paper in PAPER PLAIN, they see a side pane containing a reading guide (1), consisting of key questions the reader might ask of the paper, brief generated plain language answers, and pointers to passages in the paper the reader can read more. When a reader clicks a question (2), the paper jumps to the passage that provides that answer and shows a paragraph-length plain language answer (3). Plain language summaries can be accessed for any section of the paper by clicking a label next to the section header (4). The reader can view definitions of medical jargon by clicking underlined terms (5).

may mean becoming familiar with the literature. In the words of one patient [@access Working Group, 2021]:

> *I had been studying CLL [Chronic Lymphocytic Leukemia] through free access articles on PubMed and Google Scholar… Reading these NIH papers enabled me to have an intelligent dialogue with a CLL specialist, ultimately leading me to the selection of a clinical trial.*

However, a healthcare consumer's success in understanding the medical literature is by no means assured. Healthcare consumers report that, unsurprisingly, medical papers are difficult to read [Day et al., 2020; Nunn and Pinfield, 2014]. This is in part due to being overwhelmed by the amount of unfamiliar jargon. It is also because healthcare consumers are unaccustomed to the norms of how research is conducted and how reports of it are structured [Britt et al., 2014b; Day et al., 2020]. The result is that reading medical papers can be an experience that is challenging and at times demoralizing.

In this chapter, we ask how interactive information interfaces can make medical research articles approachable to non-expert healthcare consumers that need it, whom we refer to as "readers" in this paper. In particular, we study how articles can be imbued with new affordances to help readers navigate and evaluate their contents. The human-computer interaction literature demonstrates myriad ways that reading interfaces can assist readers, including by helping them understand unfamiliar terminology [Abekawa and Aizawa, 2016; Head et al., 2021], hiding sections that are predicting to be irrelevant [Bohn and Ling, 2021], and answering user-written questions [Zhao and Lee, 2020]. Drawing on this work as inspiration, we ask what combination of affordances would be necessary to help bridge the often enormous gap between a reader's current knowledge of biomedical research and the contents of a paper.

The key insight of PAPER PLAIN is that medical papers can be made more approachable by judiciously incorporating plain language summaries to supplement the paper's original content. A reader can engage with the original text through plain language summaries—which we refer to as "gists"—that contain simplified sentences and reduced jargon and are presented alongside passages in the paper. The reader can approach any content in the paper by first inspecting its gist, only committing attention to a dense passage after learning if it is likely to be relevant. In this way, the reader has the support to engage meaningfully with the original paper text: skipping passages of little relevance and spending time reading those of consequence.

This chapter begins with a formative observational study of 12 non-expert readers to identify barriers in reading medical research papers. We observed that, in addition to the expected pervasive difficulties of understanding passages dense with unknown terminology, readers struggled to know what parts of a paper to read and often spent considerable effort making sense of sections with limited usefulness to them. These findings suggest that reading medical papers is uniquely challenging for our envisioned readers due to their lack of domain knowledge and understanding of how medical research is communicated. Many of the questions participants had while reading papers were associated with the strategies we identified in Chapter 4, suggesting that the scientific writing strategies we identified could be useful in adapting medical papers to these audiences.

PAPER PLAIN is designed to make medical papers approachable with four features (illustrated in Figure 5.1) that together help readers understand content at multiple levels of granularity (i.e., entire sections, paragraphs, and individual terms) and throughout the reading process. PAPER PLAIN helps a reader find information relevant to them in the paper by providing a "key question index," a list of important questions a healthcare consumer may wish to ask about a medical study and plain language answers to those questions generated from the text. The key questions overlap with many of our identified strategies, including IMPACT, MAIN and LEDE. When a reader clicks one of these questions, they are taken to a paragraph that answers the question in the paper along with an "answer gist," a plain language summary of that paragraph. PAPER PLAIN conveys the essence of jargon-dense passages with "section gists", in-situ plain language summaries available for each section of the paper. Both gists mirror the EXPLANATION writing strategy. Finally, as an implementation of the JARGON strategy, PAPER PLAIN assists readers in understanding unfamiliar terms by allowing a reader to look up definitions by clicking the term. To assess how PAPER PLAIN supports the reading experience, we conducted a 24 partial within-participant usability study where participants read papers with variants of PAPER PLAIN or a typical PDF reader during a timed reading task. The study showed that PAPER PLAIN lowered participants' self-reported difficulty in reading the paper and increased confidence that they found all of the information of interest to themselves. When asked to answer questions that tested their understanding of the paper, participants answered questions neither significantly more or less accurately when they had access to PAPER PLAIN; though a follow-up analysis suggested that participants with PAPER PLAIN were more accurate on questions whose answers were easily identified given the key question index.

The clear favorite feature was the key question index and answer gists. Participants also used, and appreciated, in-situ section gists and term definitions, though participants tended not to use them when the key question index was available. Altogether, the study suggests that reading interfaces that provide guidance and plain language summaries can indeed lead readers to find papers more approachable than they would with conventional reading tools.

5.2 Background and related work

5.2.1 Healthcare consumers reading medical research

Research on consumer health information seeking suggests that trustworthy online health information can empower healthcare consumers, improve clinician-patient interactions, and increase adherence to medical recommendations [Tan and Goonawardene, 2017; Cocco et al., 2018; Broom, 2005; Johansson et al., 2021]. Tan and Goonawardene [2017] reviewed consumer health seeking behavior and perceptions on using internet information in consultation with clinicians; they found that people did not feel like internet information adversely affected consultations, and that it helped them feel more confident in the consultations and in following clinicians' suggestions. Cartright et al. [2011] distinguished two types of health information searching behaviors: evidence-based, which focused on details of symptoms, and hypothesis-based, which focused on understanding a particular diagnosis. In a related setting, Cocco et al. [2018] studied how people search for health information while in an emergency room, showing that many searched for information online on trusted sites like university or hospital websites. Kivits [2006] explored why healthcare consumers search the internet for medical information, finding that the motivations for searching included helping oneself and filling in missing information from their clinician. Choudhury et al. [2014] studied health searching and sharing behavior on search engines and social media, finding that search engines are often used for serious medical conditions, but social media can be used to share information about more benign symptoms or conditions. Work has also studied how medically concerned users search for health information online [Philipp and White, 2014] and how online searching can lead to real-world healthcare utilization [White and Horvitz, 2014].

While the internet is a good source of consumer health information, there are also many barriers to interacting with this information [Sommerhalder et al., 2009; Storino et al., 2016]. White and Awadallah [2014] analyzed top search results for common health information queries and found that top search results returned for health interventions skewed positively, meaning that more search results said that an intervention will help a condition than suggested by medical evidence. Sommerhalder et al. [2009] found that healthcare consumers searching for information online also struggled with information overload. Information overload can be caused by searches returning unrelated results (e.g., searching a particular symptom and getting results about different diagnoses or home remedies), complex text, or different trusted sites providing contradictory guidance [Sommerhalder et al., 2009; Storino et al., 2016; Kalavar et al., 2021; Baskin et al., 2020]. Most people could not resolve these issues themselves, instead needing to discuss the information during consultations with their clinicians [Sommerhalder et al., 2009].

While many people start out on consumer-facing sites, medical literature is an important source of highly specific, up-to-date information for them [Zuccala, 2010]. In 2005, the NIH established an open access policy in part to encourage "individuals [to] become educated consumers about their healthcare and related research, and to consult with healthcare professionals for specific guidance." [National Institutes of Health, 2005] Subsequent research has shown the public benefit of this open access policy, such as improved access to new research findings for healthcare workers and consumers [Tennant et al., 2016]. While the traditional debate for open-access journals have focused on wider dissemination within research communities, there is an increasing recognition that public stakeholders, including advocacy groups and healthcare consumers, can effectively make use of primary medical research findings [Day et al., 2020; Epstein, 1997]. Indeed, there is a movement in the medical community to involve patients more in the research process, including understanding lab reports [National Academies of Sciences and Medicine, 2018], reviewing research papers [Richards and Godlee, 2014] and leading research efforts [McCorkell et al., 2021; Nair et al., 2012].

At the same time, medical research, and scientific research more broadly, present unique barriers to readers without research expertise [Münchow et al., 2020]. Britt et al. [2014b] argued that science literacy is the ability to evaluate scientific texts effectively, but that this is challenging due to complex arguments and unfamiliar text structures. Bromme and Goldman [2014] highlighted hurdles that the general public face when reading scientific information, including the ability to determine what is relevant and lack of domain expertise. Day et al. [2020] outlined additional barriers specific to searching through medical research, such as lack of adequate scientific literacy, the potential to draw inaccurate conclusions from the findings, and fraudulent journals without sufficient peer review. Nunn and Pinfield [2014] interviewed healthcare consumers on reasons for accessing medical literature and their response to lay summaries written for medical papers. They found that readers appreciated the lay summaries, but often wanted to read the article themselves anyway. At the same time, other work has found that lay summaries help improve reader comprehension compared to journal abstracts [Kerwer et al., 2021]. This chapter illustrates how interactive reading interfaces can make medical research papers accessible to healthcare consumers through a novel interactive system, PAPER PLAIN.

5.2.2 Interactive reading interfaces

PAPER PLAIN draws inspiration from prior affordances in interactive reading systems that have used term definitions [Head et al., 2021], question answering [Zhao and Lee, 2020; Chaudhri et al., 2013], and guided reading [Dzara and Frey-Vogel, 2019] to support reading medical text [Marshall et al., 2016; Brusilovsky and Pesin, 1998], dialogue [Li et al., 2021], news [Bohn and Ling, 2021], and search results [Collins-Thompson et al., 2011]. Inquire Biology [Chaudhri et al., 2013] is a biology textbook augmented with artificial intelligence (AI) features to support student learning. The textbook allows students to view concept definitions and ask open-ended questions about information in the textbook. If students are unsure of what questions to ask, the textbook also recommends possible questions based on highlighted passages. In another resource for students, Dzara and Frey-Vogel [2019] introduced a new method for conducting reading groups that required no prior reading preparation through developing questions about a paper's methodology and findings. They found that these interactive discussions can help pediatric residents analyze medical papers effectively. Also in the clinical context, UpToDate [UpToDate, 2021] provides expert-written summaries of current research for healthcare providers.

In the context of reading research papers, Head et al. [2021] introduced ScholarPhi, a PDF reader that surfaces position-aware definitions for terms defined in a paper (Nonce words) and features for revealing these terms across a paper. In a usability study, researchers were able to read papers more easily using the interface. Zhao and Lee [2020] introduced "Talk to Papers," a natural language question answering system for exploring research papers. "Talk to Papers" allows users to query papers with natural language questions and provides passages where answers are taken from. Other work has explored tools for adaptive summarization in news articles [Bohn and Ling, 2021], evaluating research literature [Letchford et al., 2017; Marshall et al., 2016], navigating concepts within a paper [Abekawa and Aizawa, 2016; Jain et al., 2018] and providing reading guidance in textbooks [Weber and Brusilovsky, 2015; Brusilovsky and Pesin, 1998]. There are also interactive systems for collaborative reading of research papers, such as Fermat's Library [Fermat's Library, 2021], which provides community annotations on popular research papers, and Hypothes.is [Hypothes.is, 2021], which allows users to annotate and share annotations on any webpage.

In contrast to previous reading interfaces for research papers that focus on clinicians, researchers, or students, this chapter focuses on interactions to make papers understandable to healthcare consumers. There are key ways in which previous designs would not support these envisioned readers. Medical research text is so jargoned that a reader has to invest considerable effort learning the background knowledge to understand it. Previous interfaces that assume readers know what important questions to ask [Zhao and Lee, 2020], where to look for their answers [Chaudhri et al., 2013] or know how to make sense of definitions of terms within a paper [Head et al., 2021; Jain et al., 2018] can make reading exceedingly difficult for these readers. PAPER PLAIN goes beyond the typical capabilities of interactive readers to instead help readers understand where to find information of interest in a paper according to the language they already know. To do this, the system incorporates plain language alongside original paper content.

5.3 Observations of Non-Expert Readers

Prior work on reader barriers have focused on consumer health information [Sommerhalder et al., 2009], scientific research in other domains [Münchow et al., 2020], for students [Shanahan et al., 2011], or searching through medical literature [Day et al., 2020], but it is unclear how these barriers manifest for non-experts reading medical research papers. To gather more direct and comprehensive evidence of barriers for this population, we conducted a think-aloud reading study.

5.3.1 Participants & recruiting

We wanted to observe the barriers faced by healthcare consumers when reading medical research. However, the timing of these reading episodes was hard to predict, making it difficult to observe authentic reading experiences. As a compromise, we developed scenarios based on interviews with four healthcare consumers who had prior experience reading medical research and two healthcare providers who had discussed findings from medical papers with their patients. Healthcare consumers and providers were recruited through our personal and professional networks and by referral from other interviewees. More details on these interviews are in Appendix B.1. We then recruited participants without medical or research expertise to walk through these scenarios. We provided these participants with a primer about a medical condition and allowed them considerable agency in how they approached the reading task.

We recruited participants who had no experience in the medical profession and in undertaking research via Upwork, a crowd-work site for hiring freelancers. We listed our job under both “Editing & Proofreading” and “Customer Research” (i.e., workers partaking in user surveys) to attract a broad sample of workers with varied degrees of reading and writing experience. All participants were paid US$15 for the hour-long study.[1]

A total of 12 participants completed the study (T1–12). Of these participants, 11 had completed college

[1] This is above the federal minimum wage of US$7.25 and above the region’s minimum wage.

and 5 had completed professional or graduate school. 11 participants had taken 3 or fewer STEM courses since high school.

5.3.2 Procedure

In the study, participants were given a scenario about a fictional diagnosis representative of common but serious medical conditions (e.g., a herniated disc) with a goal for reading medical papers (e.g., finding new treatments). To ensure participants were equipped with some prior knowledge before approaching papers, they first read a consumer health webpage (MedlinePlus) about the medical condition. This MedlinePlus step was meant to more closely approximate realistic circumstances, in which a participant would have received some information from their doctor about their diagnosis.

We designed the scenarios such that participants would benefit from the additional information found in research papers. To uncover a comprehensive set of barriers, we created four scenarios varied across the following dimensions: diagnosis, demographics (i.e., common or uncommon for a diagnosis), relationship to patient (i.e., patient vs. caretaker), and motivation. There were two possible diagnoses for each scenario: a herniated disc or systemic lupus erythematosus (SLE, also called Lupus). These diagnoses were selected because they are relatively common and represent serious, long-term issues for a patient. Motivations were: learning background-specific information, becoming aware of emerging treatment options, and comparing treatment options. These scenarios were validated as realistic by a healthcare researcher familiar with consumer health. For more information on these motivations, see Appendix B.1.

Participants were randomly assigned into one of the four scenarios. Each scenario was assigned to the same number of participants. After reading a description of the scenario, participants read the MedlinePlus page on the diagnosis then browsed a list of 11 research articles related to the diagnosis. To make these papers representative of the papers healthcare consumers would find in their own searches, we selected only from PubMed articles linked from the MedlinePlus page. MedlinePlus is a patient-facing resource for medical information, so we reasoned that papers linked from it would be representative of those readers would look to first. We selected papers that were 1) review articles or randomized control trials and 2) relevant to the scenarios (e.g., covering possible new treatments). Papers varied in how relevant they were for a scenario (e.g., some papers covered treatments not clinically available), though all papers had some relevance.

While in real-world health information seeking, readers would undoubtedly come across irrelevant information [Sommerhalder et al., 2009], the study's focus was on barriers in reading papers rather than searching through papers and determining their relevance. Participants chose which papers to consult, which permitted us to see how the contents of a paper affected a participant's choice to read it deeply. Most participants had enough time to read one or two papers (all were asked to read at least one).

Participants were asked to read for a total of 40 minutes, split between the MedlinePlus summary page and the papers they chose to read. Participants thought aloud while reading. They were also asked to take notes or speak aloud on any barriers they had encountered every 5 minutes if they had not already volunteered this information. The researcher present would sometimes ask participants to elaborate on these barriers. Following the reading, the researcher interviewed participants to ask what was difficult about reading the research articles and how they thought intelligent reading tools could help them read more effectively. After the interview, participants completed a questionnaire to report their medical literacy and prior research experience.

To analyze the barriers readers faced, a reflexive thematic analysis [Braun and Clarke, 2006; Blandford et al., 2016] was performed on the think-aloud and questionnaire data. We followed Braun and Clarke [2006]'s six phases of thematic analysis. One author familiarised themselves with the interview data by rereading transcripts and rewatching interviews, making notes on barriers readers faced. This author generated initial codes for barriers based on these observations and iteratively revised the barriers with four other authors through discussion (both in meetings, and asynchronously over Google Docs). The authors reviewed each barrier and the strength of the supporting evidence. Through these discussions, barriers were refined and assigned candidate names. After refining the barriers, the first author revisited the data and checked for consistency between barriers and observations from the study. Through discussions with the first author and four other authors the barriers were further refined and assigned descriptive names. The authors drafted a report of the barriers, along with associated observations during the study.

5.3.3 Findings

Our study illustrated barriers readers face when reading medical research papers. The barriers were: unfamiliar terminology; overwhelmingly dense text; not knowing what to read; difficulty finding answers; and

Barrier	Description	Representative Quote	Readers
Unfamiliar terminology	Readers did not understand individual terms and symbols unique to the biomedical research literature.	*“What does this word mean?”*	T1–3, 5–8, 10–12
Overwhelmingly dense text	Readers had difficulty understanding the essence of passages that contained an overabundance of jargon.	*“I am not going to act like I understand what any of this means.”*	T1–8, 11–12
Not knowing what to read	Readers did not know which sections were worth their attention, and expended considerable effort reading uninformative sections.	*“Why did I waste all that time trying to understand what that was?”*	T1–3, 5–12
Difficulty finding answers	Readers had specific questions they wanted to have answered but lacked a sense of where in the text to find answers.	*“Where does it talk about how to treat this condition?”*	T4, 6, 9–10, 12
Difficulty relating findings to personal circumstances	Readers could not find enough information about whether prognoses and outcomes described in the text would apply to them.	“I would love to know how someone with the same demographics as me responded to this treatment”	T2, 5, 8–9, 11

Table 5.1: Five barriers readers encountered when they sought an understanding of medical research papers without having prior medical research experience. All barriers were caused, or exacerbated by, their lack of expertise in medical research.

difficulty relating findings to personal circumstances. Table 5.1 lists these barriers. Below we illustrate how these barriers manifested for non-experts reading medical papers, confirming the presence of these difficulties and highlighting concrete instances of difficulties that inform opportunities for design.

Unfamiliar terminology Nearly all (T1–3, 5–8, 10–12) participants mentioned struggling to make sense of the information in the papers because of medical terminology or acronyms that they did not know. These terms ranged from only appearing in some areas of biomedical research (e.g., "therapeutic peptides") to commonly used medical terms (e.g., "comorbidities," "meta-analysis"). The two participants that did not mention struggling with specific medical jargon (T4 & 9) often skimmed over these terms or were able to infer them from context. Interestingly, while others reported medical terminology as a barrier, they still made some sense of an article without knowing terms by making assumptions about the terms' meanings. At the same time, some terms had meanings that were integral to understanding an article. Incorrect assumptions about these terms could mean misunderstanding the article (T6 & 10). For example, T10 did not know that "in vitro" referred to pre-clinical, non-human studies. They only realized this after reading the majority of the article, which dramatically changed their perception of its usefulness (i.e., that none of the studied drugs were in clinical trials).

While terminology is a common barrier in scholarly communication [Martínez and Mammola, 2021], past interactions to address it present additional issues for our reading context. Past work has addressed researchers not knowing terms in a paper by providing definitions of terms based on earlier references in a paper [Head et al., 2021]. There are two issues with such an approach for our reading context: (1) the sheer number of terms could make it difficult for a reader to know which are important and (2) there is no guarantee a reader in our context would understand references drawn from the paper, considering that almost all text in medical papers has technical jargon. These issues suggest that a different approach to defining terminology for our envisioned readers is needed.

Overwhelmingly dense text While participants could ignore individual terms, such as T4 & 9, sentences were so filled with these terms, and paragraphs were so filled with these sentences, that participants were overwhelmed by passages of dense text (T1–8, 11–12). This dense text included unfamiliar terminology, but also statistics and complex wording or arguments. Because of the amount of text in the articles and the high cost of reading any of it, participants were quickly overwhelmed. As T8 put it, “Honestly reading that stuff it was. . . overwhelming just how much terminology I didn’t know to start off with. . . It’s not like I didn’t understand it at all, it was just hard to follow because I had to keep going back, like ‘Oh what does that acronym mean?’ ” T8 was reading a section containing multiple acronyms defined earlier in the paper, including ‘QoL’, ‘DORIS remission,’ and ‘SLEDAI.’ The beginning of one paragraph reads as such: “In some cases, modifiable causes like anaemia or hypothyroidism may be found, but in most patients, fatigue is unexplained. . . In contrast, SLEDAI or BILAG do not correlate with fatigue.” [Kernder et al., 2020] T5 expressed a similar sentiment when describing a results passage they were reading: “I am not going to act like I understand what any of this means. . . I would have to take the time to understand what these terms mean.” Continuously having to reference earlier sections of a paper, or searching for term definitions on the internet, can be a major distraction, especially when multiple terms appear in a single sentence. Multiplying this by every sentence in a medical paper creates a categorically different barrier than one term might present.

Dense text is a barrier that every reader has encountered when learning to read in a new language or domain and is a core motivation for text simplification research. The nuance to this barrier in the context of medical research papers is that readers might have little interest or capacity mastering the language and norms of a particular paper, given that other papers they might read could use different language, and that they may be pressed for time and emotionally and mentally drained from handling their diagnosis. This barrier is also a broad issue in science communication, not just medical papers, evidenced by our JARGON writing strategy (Section 4.2).

Not knowing what to read Of the 12 participants, 11 (T1–3, 5–12) had a difficult time knowing if a paper held relevant information and invested intense reading effort to determine this. They read papers exhaustively top-to-bottom, reading most of the text, spending time making sense of dense results sections and descriptions of statistical analyses that later they had no use in understanding (T2–3, 5–8). Much of the dense text participants reported struggling with (discussed in the previous barrier) ended up being in sections that they later discovered were less important to read (e.g., a detailed statistical results section).

One clear example of this was T5, who reported struggling to read the entire first paper they selected because they wanted to do their due diligence by understanding the results. After getting to the discussion they realized that it provided an accessible overview of the results, so for future papers they ignored the technical results sections. As they explained, “the results, which in my mind would be the first place I would want to go to. . . are very technical and I am not going to know what that means. . . so a general discussion of the results will be more helpful. . . knowing what I know now I would probably skip the results section.” This quote highlights that non-expert readers lack the knowledge of what they should–and shouldn’t–read in a paper, leading them to take much longer learning what a paper has to offer. Other participants had similar experiences as T5, though did not quickly determine what the best passages were for them to read (T2–3, 6–8).

While some used a paper’s introduction to determine how useful a paper would be, many participants did not trust their ability to know what a paper would contain without exhaustively reading it (T3, 6–8). T6 and 8, for example, both suspected that certain papers would not be useful after reading the abstract or introduction, but continued reading the papers because they hoped they would still find something that was helpful. As we will discuss more in the next barrier—searching for answers—sometimes there was indeed information not surfaced in the introduction or abstract that participants wanted to know, such as low-level details on participant demographics. Participants could invest immense effort to determine if a paper contained this information. In the case of T6, they spent 40 minutes reading a single paper. In another case, T7 reported that they suspected there was useful information in a paper, but it would take them too much time to find it. T3 provided a similar sentiment of wanting a way to know exactly what to read first in a paper: “I would love some sort of. . . thousand foot-view, which is kind of what I needed in the beginning. Make [the paper] less designed for doctors, and make it more patient friendly, where you are less overwhelmed by all the information all at once, where you can search it out in smaller bites.” When asked to elaborate, T3 explained that the smaller bites of information could provide high-level findings that they could follow-up on for more details if they were interested. It is worth noting that some biomedical papers do structure abstracts with high level summaries of all sections first or include article highlights at the beginning of the paper, which could help non-expert readers as well as scientists reading these papers.

Difficulty finding answers Participants in our study had specific information they tried to find in the paper, but struggled to do so (T2, 4, 6, 9–10, 12). In contrast to the previous barrier where participants struggled to know what to read in a paper, sometimes participants knew what they wanted to read, but couldn’t find this in the paper. The two most salient examples of this barrier were searching for patient demographics and previous treatment options. T2 tried to find information on specific demographic groups in the study to see if they matched their scenario. They had to read through the entire article to find a table with patient demographics and a single sentence within the discussion section making reference to the patient group most relevant to them. Abstracts also did not talk about study demographics or current best practices for treating an illness. Introductions would often include useful information, but it was hidden in background paragraphs or quickly mentioned before moving on to the novel results. Participants therefore had to sift through headers and paper sections, making sense of unfamiliar terms and dense text (two previously discussed barriers) while trying to determine if each sentence was relevant to them.

Difficulty relating findings to personal circumstances Some participants also wanted additional information from the papers that were personally relevant to them (T2, 5, 8–9, 11). T2 and 8 imagined a tool that could explain how a treatment would affect them, such as by providing patient testimonials for treatments in the paper or results for slices of patients based on demographics. For example, T2 read a paper that reported a 60% reduction in pain after a surgery, but they wanted to know whether patients regretted the surgery or would recommend it. They also wanted results for a slice of patients most similar to their hypothetical scenario, a 20 year-old male smoker, but the paper only presented average reductions across all patients. T5 found it helpful when an article made reference to the monetary cost of different treatments as a way of referencing patient experiences, though this only happened in one paper. While this personally relevant information was not the goal of the research papers, participants sought this information as a way of relating the information in the paper to their own lives.

These barriers are unique, or uniquely challenging, to our envisioned readers, necessitating a novel approach to ameliorating them. Past interactive reading systems for research papers have assumed readers have extensive domain knowledge, are able to make sense of paper text as a way of resolving unfamiliar terms [Head et al., 2021], know what are the right questions to ask of a paper [Zhao and Lee, 2020], and understand the basic structure of a paper [Abekawa and Aizawa, 2016]. In contrast, the barriers we identify illustrate that these assumptions do not hold for non-expert readers.

Looking back at Chapter 4, many of the strategies we identified could be used to support readers in knowing where to read. The MAIN and strategies both focus on highlighting the most important information for a reader, which could be helpful in overcoming the barriers of not knowing what to read. The IMPACT strategy could help readers know what to take away from the papers, supporting them in overcoming their difficulty finding answers. The JARGON and EXPLANATION strategies could also help alleviate unfamiliar terminology and overwhelmingly dense text. Below we discuss how our system, PAPER PLAIN, seeks to address these barriers using features inspired by the writing strategies: plain language (gists) and a collection of key questions as a reading guide, both novel techniques in the context of interactive reading systems for research papers.

5.4 PAPER PLAIN: Reading Support for Medical Research Papers

PAPER PLAIN makes medical papers approachable to non-expert healthcare consumers. Unlike other systems in the augmented reading space for research papers, PAPER PLAIN focuses on the barriers of non-experts, such as knowing where to invest reading effort. To address this reading context, PAPER PLAIN integrates existing features like term definitions with novel navigational guidance and reading support through a Key Question Index and Answer Gists.

We focus on four of the five barriers discussed in Section 5.3: unfamiliar terminology, overwhelmingly dense text, not knowing what to read, and difficulty finding answers, because these were the most common barriers we observed that hampered readers' ability to get useful information from the papers. In contrast, difficulty relating findings to personal circumstances reflected a desire for additional information and was less focused on understanding information in the paper itself.

We followed an iterative design process for developing PAPER PLAIN. A total of 8 participants used 2 early prototypes of PAPER PLAIN in qualitative usability evaluations. Participants were recruited from our institution, professional networks, and Upwork. Evaluations lasted one hour each. The iterative design is described in more detail in Appendix B.2.

Based on feedback from the iterative design process, PAPER PLAIN was designed with the following features:

1. **Term Definitions** – Tooltips provide definitions of unfamiliar terminology from the open web.

2. **Section Gists** – In-situ plain language section summaries support readers' understanding of dense paper text.

3. **Key Question Index** – Key questions in the sidebar guide readers to relevant answering passages.

4. **Answer Gists** – Plain language summaries of the answering passages help readers understand the important information contained in these passages.

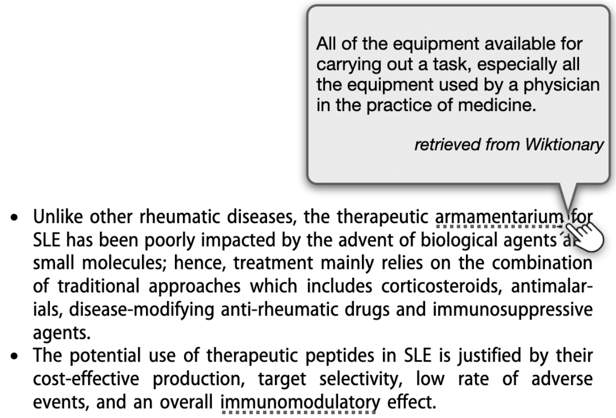
By designing PAPER PLAIN's features to address reader barriers, we found that the features aligned well with the strategies we previously identified. Term Definitions implement the JARGON strategy by resolving unfamiliar terms for a reader. Answer and section gists also resolve terminology, but also provide longer explanations for complex language (EXPLANATION). Finally, the key question index provides an index to the main findings of the paper (MAIN), and highlights the real world impact of the work (IMPACT).

To illustrate how PAPER PLAIN is designed towards the goal of making medical papers approachable to general audience readers, we describe how a fictional reader, Sarah, leverages PAPER PLAIN to achieve her goal of finding more information about new treatment options. Sarah is a first-time reader of medical literature and therefore might differ from some readers who have become familiar with medical terminology because of prior efforts to understand a chronic condition. That being said, we believe that PAPER PLAIN's features for highlighting useful information in a paper can support first-time as well as regular, non-expert readers.

Sarah is a 25 year old woman (pronouns: she/her) who was recently diagnosed with Systemic Lupus Erythematosus (SLE, also called Lupus). Currently her symptoms are mild: some joint pain and tiredness, but symptoms can worsen and become debilitating over time. When Sarah discusses treatment options with her doctor, she wonders if there are treatments the doctor does not mention that might benefit Sarah. Sarah works during the day, but in the evening she starts looking for research papers to be informed on available treatments. Sarah finds a research paper about possible new treatment options, titled:

> "Therapeutic peptides for the treatment of systemic lupus erythematosus: a place in therapy." [Talotta et al., 2020]

After reading the title, Sarah has many questions – *What is the paper about? What are therapeutic peptides? Are they a possible new treatments for SLE?* – and begins reading.



All of the equipment available for carrying out a task, especially all the equipment used by a physician in the practice of medicine.

retrieved from Wiktionary

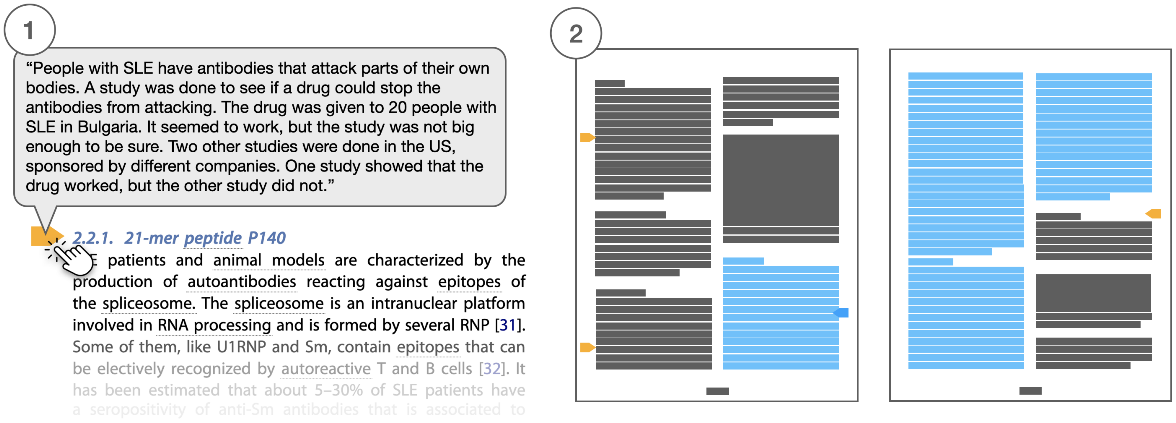
- Unlike other rheumatic diseases, the therapeutic armamentarium for SLE has been poorly impacted by the advent of biological agents and small molecules; hence, treatment mainly relies on the combination of traditional approaches which includes corticosteroids, antimalarials, disease-modifying anti-rheumatic drugs and immunosuppressive agents.
- The potential use of therapeutic peptides in SLE is justified by their cost-effective production, target selectivity, low rate of adverse events, and an overall immunomodulatory effect.

Figure 5.2: Term Definitions on an example passage with technical jargon. Terms with definitions are underlined ("armamentarium", "immunomodulatory"). Clicking a term will open a tooltip with a definition and a reference to the definitional resource.

Term Definitions help Sarah resolve technical jargon without distracting from reading PAPER PLAIN provides definitions for unfamiliar terms in the context of the paper so Sarah can seamlessly integrate new concepts into her reading. While reading the introduction, Sarah reaches a passage full of technical jargon (Figure 5.2). In the first bullet, she is unsure what "therapeutic armamentarium for SLE" means, preventing her from understanding *what* has been "poorly impacted." Rather than open a new tab to search, Sarah clicks on the underlined term and a tooltip appears with a short definition retrieved from Wiktionary[2] explaining that "armamentarium" refers to medical equipment. In the next bullet, Sarah sees a list of promising properties of "therapeutic peptides in SLE," but is unsure what "overall immunomodulatory effect" means. The definition tooltip again helps her understand that therapeutic peptides can help control immune functions. Sarah continues reading, using the tooltips to resolve unfamiliar terms, eschewing the need to constantly switch tabs for finding definitions.

The Section Gists help Sarah decide whether to invest in reading dense passages Equipped with Term Definitions, Sarah manages to learn from the introduction that peptides are indeed possible treatments for SLE and wants to learn more. This particular paper reviews 15 different peptides, each with a dedicated section averaging one page in length; each section includes a description of how the peptide works and its clinical trial results. Sarah is motivated to get a high-level sense of each available peptide, but it will require

[2] `https://en.wiktionary.org/wiki/`



“People with SLE have antibodies that attack parts of their own bodies. A study was done to see if a drug could stop the antibodies from attacking. The drug was given to 20 people with SLE in Bulgaria. It seemed to work, but the study was not big enough to be sure. Two other studies were done in the US, sponsored by different companies. One study showed that the drug worked, but the other study did not.”

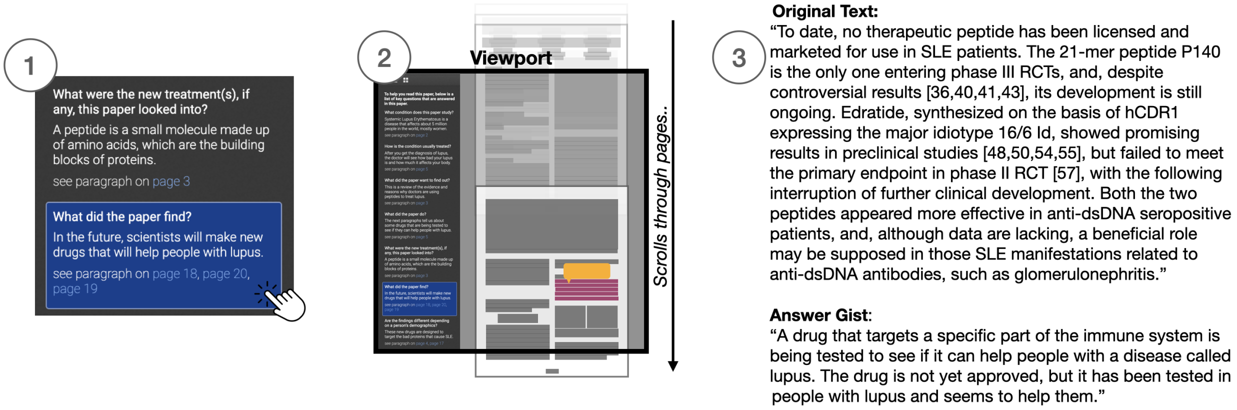
2.2.1. 21-mer peptide P140

SLE patients and animal models are characterized by the production of autoantibodies reacting against epitopes of the spliceosome. The spliceosome is an intranuclear platform involved in RNA processing and is formed by several RNP [31]. Some of them, like U1RNP and Sm, contain epitopes that can be electively recognized by autoreactive T and B cells [32]. It has been estimated that about 5–30% of SLE patients have a seropositivity of anti-Sm antibodies that is associated to

Figure 5.3: Section Gists on an example passage with dense text. (1) Clicking on a tab indicator next to a section title displays a plain language summary of the section. (2) Tabs are positioned throughout the paper, providing summaries that can cover a lot of paper content, even across pages.

reading 20 pages of dense text. From the introduction, Sarah had gathered that not every peptide has equally promising results and each might be used for different treatments of SLE (e.g., more moderate or more severe cases), so she would prefer to only read in depth about the most promising peptides relevant to her mild case of SLE.

Paper Plain makes it easy for Sarah to quickly determine what sections are interesting to her and understand the sections with in-situ plain language summaries (Section Gists). Sarah clicks on an angled flag next to the section title, and a tooltip appears adjacent to the section text (Figure 5.3). The tooltip contains a summary of the section stripped of jargon. Rather than sentences like “SLE patients and animal models are characterized by the production of autoantibodies reacting against epitopes of the spliceosome.”, the summary explains that “People with SLE have antibodies that attack parts of their own bodies.” Sarah learns from the section gist that this particular peptide has had some good preliminary results, but that further studies have had less successful results. She confirms these details by skimming the section and decides this section isn’t so relevant to her. Sarah refers to the Section Gists to develop a surface-level understanding of the rest of the peptide sections, writing down a few peptides that she is interested in keeping track of, without having to parse all the dense, mostly irrelevant text. Sarah completes her reading of these sections in 15 minutes rather than spending hours going through each section.



What were the new treatment(s), if any, this paper looked into?

A peptide is a small molecule made up of amino acids, which are the building blocks of proteins.

see paragraph on page 3

What did the paper find?

In the future, scientists will make new drugs that will help people with lupus.

see paragraph on page 18, page 20, page 19

Original Text:
"To date, no therapeutic peptide has been licensed and marketed for use in SLE patients. The 21-mer peptide P140 is the only one entering phase III RCTs, and, despite controversial results [36,40,41,43], its development is still ongoing. Edratide, synthesized on the basis of hCDR1 expressing the major idiotype 16/6 Id, showed promising results in preclinical studies [48,50,54,55], but failed to meet the primary endpoint in phase II RCT [57], with the following interruption of further clinical development. Both the two peptides appeared more effective in anti-dsDNA seropositive patients, and, although data are lacking, a beneficial role may be supposed in those SLE manifestations related to anti-dsDNA antibodies, such as glomerulonephritis."

Answer Gist:
"A drug that targets a specific part of the immune system is being tested to see if it can help people with a disease called lupus. The drug is not yet approved, but it has been tested in people with lupus and seems to help them."

Figure 5.4: Key Question Index guides readers to answering passages and their Answer Gists. When one of the questions is clicked (1), the interface will scroll (2) to the first answering passage (purple) and display a tooltip (orange) containing the Answer Gist. In (3), we show the simplified Answer Gist alongside the original paper text.

The Key Question Index and Answer Gists help Sarah focus on the most important questions and relevant passages. Sarah gets to the end of the paper using the Section Gists to read only some sections in depth, but is worried she might have missed important information in the paper because she didn't know to look for it. Sarah got a general sense of each section using the Section Gists but is curious if there is information that the general summaries might not have surfaced, especially in larger sections containing lots of relevant information, such as the Discussion or Introduction.

As an alternative to assessing relevance with Section Gists, PAPER PLAIN provides Sarah with key questions linked to answering passages in the paper along with plain language answers to point Sarah to important information. Sarah looks to PAPER PLAIN's sidebar and sees questions about the paper that cover key information, such as "What did the paper do?" and "What did the paper find?" Sarah sees that the question "What did the paper find?" hyperlinks to multiple passages within the Discussion (see (1) Figure 5.4). She clicks on the first link. PAPER PLAIN scrolls through the pages and settles on a highlighted paragraph in the Discussion summarizing the most promising therapeutics peptides (see (2) Figure 5.4). Unfortunately, the answering passage looks dense. Sarah notices a tooltip below the answering passage containing a plain language summary (an "Answer Gist"). This answer gist is a quarter the length of the original paragraph and contains none of the unfamiliar terms (see (3) Figure 5.4). While the answer gist by itself might not contain all the information Sarah wants, she can read the original paragraph along with the answer gist, comparing the complex wording with plain language and get a general understanding of the paragraph without being overwhelmed by technical jargon. Similar to the Section Gists, Sarah can then dive into the original passage with this understanding to get more details. Sarah clicks through the rest of the links for the same question, which scrolls her to individual paragraphs in the discussion that cover the most important findings and interpretations of the paper.

Sarah can also use PAPER PLAIN's key questions to find questions she might not know to ask about a paper. Before finishing reading the paper, Sarah looks through the rest of the questions in PAPER PLAIN's sidebar. Each question is accompanied by a one-to-two sentence plain language answer preview and hyperlinks to one or more paragraphs in the paper that answer the question. With only a handful of key questions and short answers, a majority of the questions can be displayed in the sidebar without scrolling so Sarah can quickly read all the questions and answers with minimal effort (see (1) in Figure 5.4). Sarah sees and clicks on one question she hadn't thought to look for in the paper: "What are the limitations of the findings?" PAPER PLAIN scrolls her to a paragraph in the Conclusion saying that not only are therapeutic peptides currently not licensed for clinical use for SLE (which Sarah had already read), but also that many of the current clinical trials have mixed efficacy results and that future clinical trials might show more promise with different study designs (which Sarah had not already read). Sarah is glad she confirmed and deepened her understanding of the paper's limitations with this final question.

Sarah has spent only a few minutes to learn the most important information about the paper for her: these are not treatments she could ask her doctor to prescribe her, but there might be some promising clinical trials Sarah could look into. She also feels confident that for future papers she could use this key question sidebar to quickly get a high-level summary of the most important information in a paper.

5.5 Implementation

PAPER PLAIN leverages active research in NLP for biomedical question answering [Yoon et al., 2020] and plain language summarization [Guo et al., 2021] to address reader barriers. Below we discuss the implementations powering each feature of PAPER PLAIN. While additional algorithmic advances or human oversight, specifically for ensuring factuality [Maynez et al., 2020], are necessary to make deploying such a system safe, our current implementation indicates the potential for PAPER PLAIN to be deployed at scale

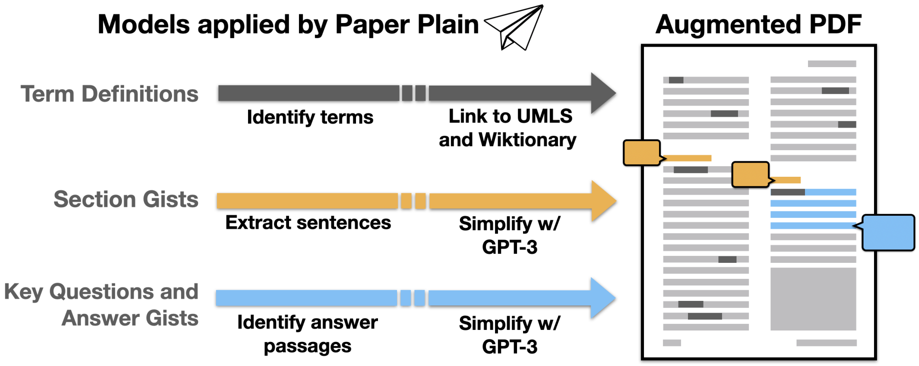


Figure 5.5: PAPER PLAIN uses machine learning models to add term definitions, section gists, and answer gists to the PDF.

over the medical literature.

5.5.1 Term Definitions

PAPER PLAIN identifies medical terms in the paper using `scispaCy` Named Entity Recognition (NER) [Neumann et al., 2019] and links these terms to definitions from the Unified Medical Language System (UMLS)[3] or Wiktionary.[4] The extraction and linking process led to many false positives (e.g., identifying terms like 'expert' or 'negative'), so we additionally filter terms based on word frequency and length. For both Wiktionary and UMLS, we preserve the bottom 80% of terms based on word frequency and remove all terms at or above 30 characters (terms over 30 characters were usually ill-formed, for example, containing a citation string or the beginning of the next sentence). We additionally filter all Wiktionary definitions to those containing at least one of the following tags: 'medicine', 'organism', 'pathology', 'biochemistry', 'autoantigen', 'genetics', 'cytology', 'physics', 'chemistry', 'organic chemistry', 'immunology', 'pharmacology', 'anatomy', or 'neuroanatomy.'

5.5.2 Section Gists

We define section gists for the lowest level subsections in the paper (e.g. "2.2.1"). To generate section gists, we concatenate the first sentence of every paragraph in a section and generate a plain language summary of it using GPT-3 [Brown et al., 2020]. GPT-3 is a pretrained langage model model released by OpenAI that

[3] `https://www.nlm.nih.gov/research/umls/index.html`

[4] `https://en.wiktionary.org/wiki/`

has obtained state-of-the-art results on many language tasks using different prompts for generation [Brown et al., 2020] and is commonly used for many generative tasks (e.g., generating plain language). We engage in prompt engineering, a common practice for achieving fluent text for large generative models [Liu et al., 2021], to encourage fluent and specific plain language summaries. We use a GPT-3 prompt adapted from a preset example that OpenAI provides for simplifying text.[5] We modified the prompt to suggest a fifth grade reading level. We also tested later grades, up to college, but found that the generated text using the fifth grade reading level prompt was the most coherent while still providing some details about the section. Sentences were extracted manually for our prototype system, but could be automatically extracted using PDF parsing methods [Lopez, 2009; Shen et al., 2021]. Using the leading sentence of each paragraph is a common competitive baseline for summarization [Erkan and Radev, 2004]; we chose this strategy rather than inputting the full section text because GPT-3 is prone to copying the text verbatim when given the full section. More details on the GPT-3 prompt are in Appendix B.3.

Because of the risk of hallucinations (i.e., factual inaccuracies) in generated text, we curated gists. If the gist contained clear hallucinations (e.g., if it incorrectly referred to a peptide as a surgical procedure), or was completely incoherent (e.g., repeated the same word over and over), we would regenerate up to five times without modifying the prompt or parameters. If a generated gist was coherent and factually accurate before five tries, we would use that gist. Usually, it only took 1–2 attempts to generate a valid, coherent gist; more details can be found in Appendix B.3.

5.5.3 Key Question Index and Answer Gists

Key questions were drawn from two sources designed to translate medical findings applicable to patients: the PICO framework [Richardson et al., 1995] for clinical questions and Cochrane's guide on writing plain language summaries [Cochrane, 2021]. Both sources focus on information in medical papers that are relevant to patients and caregivers. We curated 8 questions from the two sources for inclusion in PAPER PLAIN; these are listed in Table B.2 in the Appendix.

For each question, PAPER PLAIN extracts relevant passages from the paper using an extractive question answering (QA) system trained on BioASQ, a biomedical question answering task [Yoon et al., 2020]. Be-

[5] `https://beta.openai.com/examples/default-summarize`

cause this QA model extracts single words or phrases rather than full passages, we used the entire paragraph that contains an answer extracted by the model. For our prototype system, we manually labelled sentence boundaries of the extracted answers on the PDF to ensure high quality bounding boxes for display. Recent work has improved the accuracy of automatic sentence bounding box extraction from PDFs [Shen et al., 2021], which could be used to automate this step in the future. We follow prior work on making QA models more robust by including semantically-equivalent variations of questions [Gan and Ng, 2019].

In the system, we highlight the paragraph containing the answer and display an answer gist summarizing the answer. We created answer gists by simplifying the extracted passages using GPT-3 [Brown et al., 2020] with the same prompt and curation procedure we used for simplifying section gists. We also include the first 1-2 sentences of the answer gist in the sidebar along with the question.

5.6 Usability Study

PAPER PLAIN is meant to help readers engage with medical research papers important to them. We ran a partial within-subjects usability study to assess how well PAPER PLAIN's features meet these goals.

The study answers the following questions:

RQ1-How did participants use PAPER PLAIN's features?

- *Did participants prefer some features over others?*
- *Did participants use features throughout the reading session?*
- *Did presence of one feature affect usage of another feature?*
- *Did participants traverse linearly through a paper or employ a jumping reading strategy?*

RQ2-How does PAPER PLAIN affect participants' self-reported reading difficulty, understanding, and ability to identify relevant information?

- *...in comparison with a standard PDF reader?*
- *How does providing reading guidance (i.e., the Key Question Index and Answer Gists features) affect these self-reported metrics?*

- *...in comparison with an interface with only non-guidance features (i.e. Section Gists and Term Definitions)?*

RQ3 - Do we observe any difference in paper comprehension when participants use PAPER PLAIN?

- *...in comparison with a standard PDF reader?*
- *What is participant behavior in the presence of incorrect system predictions (e.g., vague information or factual errors in generated gists)*

5.6.1 Method

Participants

We recruited participants from Upwork using the same recruiting materials as Section 5.3.1. We again recruited from both the "Editing & Proofreading" job category and "Customer Research" to attract a broad sample of workers with varied degrees of reading and writing experience and to remain consistent with Section 5.3.1. All participants were paid US$15 for the hour-long study.

A total of 24 Upworkers (9 male, 1 non-binary, and 14 female) participated in the study. Participants' age ranged from 19 to 67 ($\mu = 35.04$). All participants had completed college, and a third had completed professional or graduate school. 79% of participants (19) had taken 3 or fewer STEM course since high school and 92% (22) had never been involved in publishing a research paper. Similar to Section 5.3, no participants had professional medical experience.

Procedure

The usability study consisted of two parts, each corresponding to a scenario involving a patient with a particular diagnosis—systemic lupus erythematosis (SLE) or a herniated disc—who was interested in exploring new treatments. The scenarios for each paper were drawn from Section 5.3.2. For each scenario, we selected a single paper ([Talotta et al., 2020] for SLE and [Bai et al., 2021] for a herniated disc) for participants to read based on the most common papers readers selected in Section 5.3.

Each participant underwent the following study procedure once for each scenario. First, participants read a description of the scenario, the MedlinePlus page about their diagnosis and the associated research paper. Then, they answered questions about the paper. Participants read the scenario description and had 2 minutes to read the MedlinePlus page on the diagnosis. They went through a short tutorial on PAPER PLAIN then read the paper for 10 minutes. They were told at 5 minutes and 9 minutes how much time remained. After each paper, participants filled out subjective ratings and multiple choice questions about the paper (described in Section 5.6.1). The duration of the reading task was set to 10 minutes following our observations from the formative study (Section 5.3) and pilot studies that this was the typical amount of time participants spent completing an initial overview read of a paper.

After the two scenarios, participants completed a questionnaire on their demographics, education, and research experience. Following the questionnaire, participants completed a short form on their experience using PAPER PLAIN and what features they found most helpful. A researcher was present for the entire experiment and followed up on these answers with additional probing questions in a final interview.

Measures

We collected measures for assessing feature usage (**RQ1**), subjective reading experience (**RQ2**) and comprehension (**RQ3**):

Feature usage To measure how participants used PAPER PLAIN's features (**RQ1**) we collected telemetry data on interactions with PAPER PLAIN's features (e.g., opening a definition tooltip or clicking on a key question). We report feature usage over the 10 minutes of reading each paper. We determine significant patterns of usage if the majority of participants exhibited this pattern, as observed by researchers present in the experiment and corroborated by the rest of the authors when examining usage data.

Subjective reading experience We collected subjective ratings to understand how PAPER PLAIN affected participants' reading experience. Participants completed the ratings after reading each paper. These included:

1. Reading difficulty: Participants rated their reading difficulty on a 1–5 Likert-style scale based on the question: "How hard did you have to work to read the paper?"

2. Understanding: Participants rated their understanding of the paper on a 1–5 Likert-style scale based on the question: "How much do you feel like you understood the paper?"

3. Relevance: Participants rated their confidence they got any relevant information from the paper on a 1-5 Likert-style scale based on the question: "How confident are you that you got all the relevant information from the paper?"

Comprehension We developed multiple choice questions to assess how different interfaces affected participants' understanding of specific details of the paper (**RQ3**). The questions were designed to assess understanding of the paper content without biasing in favor of the experimental condition; therefore, questions were selected that could not be answered directly from the Answer Gists or Key Question sidebar. Table 5.2 shows example comprehension questions and passages of the paper that contained answers to those questions. In summary, our multiple choice questions:

- were specific to the individual papers,
- captured paper information relevant in a clinical context, and
- could not be answered directly from the Answer Gists or Key Question sidebar in PAPER PLAIN.

We achieved these goals by writing 15–20 questions for each paper and having two practicing physicians not involved in the study provide feedback on the questions. The clinicians read the papers without PAPER PLAIN, gave feedback on all questions, and selected 5–7 they thought were most meaningful for the paper and were important in a clinical context. We revised the wording on any questions or answers that were unclear or easy to misunderstand according to the clinicians and two additional pilot studies. At the end, we selected 14 multiple choice questions, 7 for each paper. Paper comprehension was measured as the proportion of questions answered correctly. Participants answered these questions after completing the subjective ratings for a paper.

Interface variants

To understand the impact of PAPER PLAIN's novel guidance-offering features on readers' experience engaging with medical research papers, we evaluated variants of PAPER PLAIN with and without these features.

Question	Correct Multiple Choice Answer	Relevant Passage in Paper
What is hydroxychloroquine?	It is a treatment commonly used for people with mild to severe SLE	SLE patients with a mild involvement can be easily managed with a low dose of oral steroids (to be discontinued as soon as possible), hydroxychloroquine, and symptomatic drugs.
What would one of the eventual uses of therapeutic peptides be for SLE?	They could be used to reduce symptoms of SLE by targeting a specific organ, such as the kidneys	*[from multiple passages]* The potential use of therapeutic peptides in SLE is justified by their cost-effective production, target selectivity, low rate of adverse events, and an overall immunomodulatory effect… Moreover, they could temporarily be utilized to manage SLE flares.
What is the biggest limitation for developing therapeutic peptides?	There isn't enough evidence yet that peptides are effective at treating SLE	Although no therapeutic peptide has been licensed for SLE treatment...they show a good safety profile but have mostly failed to achieve the primary endpoints despite positive results observed in some subsets of SLE patients.

Table 5.2: Examples of multiple choice questions and answers from the usability study.

There were three versions of PAPER PLAIN and one baseline:

1. PAPER PLAIN – The full interface with the Key Question Index and Answer Gists, Section Gists, and Term Definitions.

2. Questions and Answers – The guidance-focused variant with only the Key Question Index and Answer Gists.

3. Sections and Terms – The variant without guidance, providing readers with the Section Gists and Term Definitions.

4. PDF baseline – A typical PDF reader.

Conditions With four interface variants and two papers, our study tested eight conditions, each consisting of one interface-paper pair. Each participant was assigned two conditions, i.e., two of the possible eight interface–paper combinations. No participant experienced the same interface or saw the same paper twice. Each interface–paper configuration occurred the same number of times as the first or second task in the study. All eight configurations were assigned the same number of participants across all study sessions.

Analysis

We compared readers' subjective ratings (for reading difficulty, understanding, and relevance) and number of correct answers to the multiple choice questions across the interface variants (PAPER PLAIN, Questions and Answers, Sections and Terms, PDF baseline) using mixed-effects linear models [Lindstrom and Bates, 1990] with paper type and system variant as fixed effects and participant as a random effect. Using a mixed-effects model for each measurement, we first conducted F-tests for any significant difference across the system variants, and then we conducted t-tests for differences in the estimated fixed-effects between all pairs of system variants. More details are in Appendix B.4.

Historically, usability studies with reading interfaces have often failed to reveal significant differences in how readers answer comprehension questions with and without experimental interfaces (see for instance, studies conducted by Head et al. [2021] and Badam et al. [2019]). A lack of significant difference can be attributed to several reasons: there could be similar comprehension between conditions, or the instrument might be incapable of measuring comprehension, or there may have been too little data to observe an effect amidst high variance. Understanding the nature of an insignificant difference is important, particularly if the interface could have degraded comprehension. Plain language can overly-simplify scientific findings, and might risk leading readers to misunderstandings the material [Scharrer et al., 2017; Sumner et al., 2014a].

Therefore, we also conducted a non-inferiority test [Walker and Nowacki, 2010] to confirm that PAPER PLAIN did not detract from paper comprehension. Non-inferiority tests evaluate the hypothesis that a treatment is no worse than the control. They have been used in psychotherapy research to assess, for example, the effect of remote versus in-person interventions [Lovell et al., 2006; Wagner et al., 2014; Leichsenring et al., 2018]. Non-inferiority tests are conducted similarly to traditional hypothesis testing, but the test evaluates if the difference between a treatment and control is significantly larger than an equivalence margin δ. In this case, we set $\delta = 1$, meaning that our non-inferiority test measured if the difference in the number of correct answers to multiple choice questions between Paper Plain and a typical PDF reader was within 1 correctly answered question. We use the lower bound t-test of the `statsmodels` TTOST package in Python [Seabold and Perktold, 2010] for the non-inferiority test.

For qualitative findings, one author conducted a thematic analysis on the observations of the study sessions similar to the one in Section 5.3. The author discussed findings with four other authors to refine the themes. Themes were identified via open coding and discussed in three weekly meetings with all authors. One author coded all interviews, while another author verified the themes in one of the interviews.

5.7 Results

Below we report our findings from the usability study broken down by research question.

5.7.1 How did participants use PAPER PLAIN's features?

Most participants interacted with all the features of PAPER PLAIN available to them. All participants with access only to the Key Question Index and Answer Gists (Questions and Answers) clicked on at least one Key Question and opened an Answer Gist. Usually they clicked on many more: on average participants with this variant clicked on 15 Key Questions and Answer Gists. 11 out of 12 participants with the Section Gists and Term Definitions (Sections and Terms) clicked on a Section Gist and a Term Definition. On average,

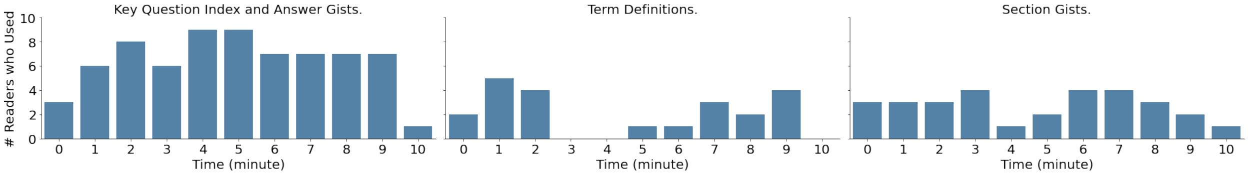


Figure 5.6: Number of readers who used each feature of Paper Plain during each minute of the ten-minute paper-reading task. Notably, all features were used throughout the reading task, rather than exclusively at the beginning or end. The Key Question Index and Answer Gists were consulted particularly frequently. Feature usage is shown only for participants in the Paper Plain condition (N=12); however, participants in other conditions (e.g., those who only had access to the Key Question Index and Answer Gists) exhibited similar behavior, with higher usage of Section Gists and Term Definitions when only those features were enabled (see Figure 5.7).

participants with this variant clicked on 18 Section Gists and 5 Term Definitions.

When participants had access to all the features they often opted for the Key Question Index and Answer Gists. 11 out of 12 participants with access to all of Paper Plain clicked on a Key Question and opened an Answer Gist, doing so on average 13 times for Key Questions and 14 for Answer Gists. In contrast, only 8 out of 12 participants clicked on a Section Gist or Term Definition. Participants that did engage with these latter features also used them much less, clicking on average only 7 Section Gists and 4 Term Definitions. Figure 5.7 plots the usage of each feature for Paper Plain and illustrates this preference for the Key Question Index and Answer Gists when all features were present.

Participants often consulted the same questions in the question index and the same answer gists multiple times. More specifically, while the Key Question Index listed only 8 questions in each condition, on average participants clicked on questions more than 10 times when the index was available. One reason participants may have clicked on questions repeatedly is that participants reported using the index as navigational support, where the questions were clicked to jump to information a participant found important. A clear example of this behavior is from P10, who used the index to jump to sections of the paper, which we discuss later in this section.

Participants used Paper Plain's features throughout the entire reading task, implying that the features continued to provide value well into the reading task. See Figure 5.6, which shows the minute-by-minute usage of the features over the course of reading task. Notably, while there is a slight 'warm-up' period for each feature—usually in the first two minutes—where participants used the features less, usage increased

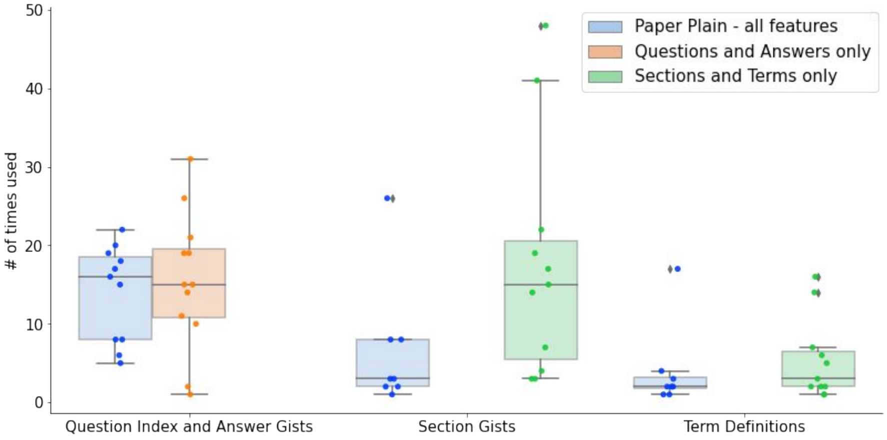


Figure 5.7: Number of readers who used each feature across the different variants of PAPER PLAIN. Points represent individual readers. For example, all but one reader who had access to all features used Term Definitions more than 5 times, shown by the single blue dot above the rest in the far right of the plot, at the 'Term Definitions' tick. Notice the drop in usage of Section Gists and Term Definitions when all features are available (the blue boxplots).

after this initial phase, and led to sustained interaction with the features for the remainder of the task time.

Along with this sustained usage we observed changes in reading strategies when participants had PAPER PLAIN's features compared to when they did not. Most participants with the baseline PDF reader read papers linearly and, similar to what we observed in Section 5.3, got stuck in dense sections with limited important information (e.g., technical backgrounds) (P2, 5, 6, 10, and 22). For example, P22 did not get to the end of one of the papers because they were focused heavily on understanding the dense methodology and background sections. When told they had a minute left, all but one of these participants (P2, 5, 10 and 22) quickly scrolled to the end of the paper to read the sections there, suggesting that they viewed these sections as important but did not have adequate time to read them.

All participants with PAPER PLAIN made it to the end of a paper; PAPER PLAIN's features supported readers in doing so in different ways. The Section Gists and Term Definitions helped participants understand dense text (P1, 3–5, 7, 15, 18), while the Key Question Index and Answer Gists allowed participants to quickly find text that was informative for them (P2, 4, 7–10, 13, 18–20).

Participants with the Section Gists and Term Definitions were able to easily make sense of dense passages (P1, 3–5, 7, 15, 18). As P18 explained, "It [the Section Gists] broke down very complicated medical text into easily understandable terms that helped me to keep up with the article and not skip over the wall

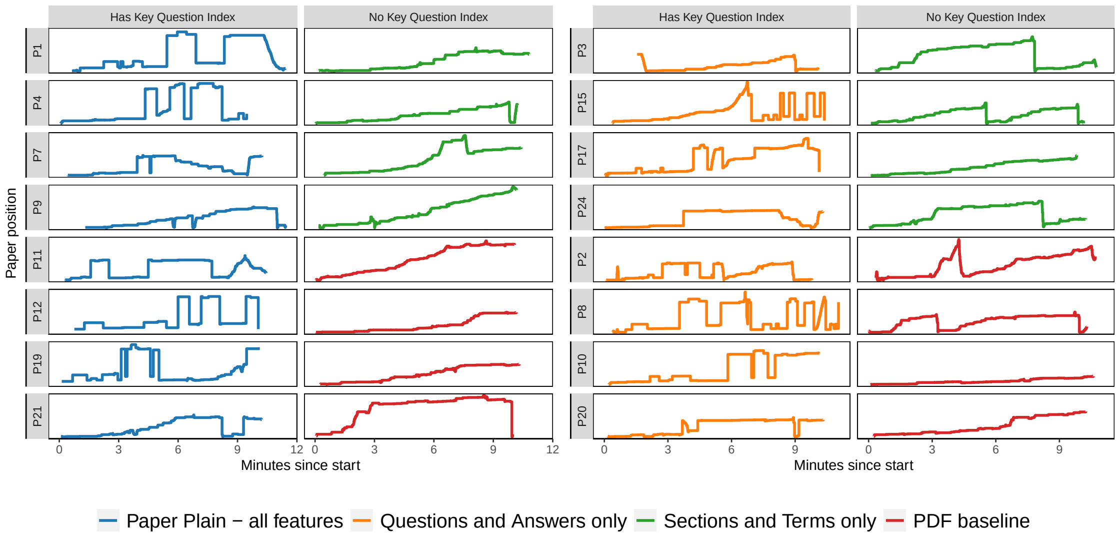


Figure 5.8: Participant's reading behavior differed when reading one paper with the Key Question Index and Answer Gists and another paper without. Each plot's y-axis calculates the vertical position of the participant's viewport relative to total paper length (e.g., the bottom and top of each graph are the beginning and end of the paper, respectively). We observe much more jumping behavior when the Key Question Index and Answer Gists are present.

of text." Participants also used the Section Gists to decide whether or not they wanted to read a section and, when they decided to read the section, as a guide for understanding the complex text (P5, 7). As P7 reported, they "liked having a brief summary of what to expect so I don't walk in completely clueless." This feedback aligns with our design goal for the Section Gists, which was to help readers avoid reading an abundance of dense text by giving them a preview of what the text is about (Section 5.4).

In contrast, participants with the Key Question Index and Answer Gists used the questions' guidance to quickly find text that was informative for them by jumping to that information (P2, 4, 7–10, 13, 18–20). A clear example of this was P10, who read through the abstract and introduction of a paper, then opted for using the Key Questions to jump through different sections of the paper. When asked why, they replied that they trusted that the questions would provide them the information they were looking for.

How the Key Question Index and Answer Gists encouraged a nonlinear reading strategy is also reflected in where participants spent their time in a paper. Participants with the Key Question Index and Answer Gists (with both Questions and Answers and the full PAPER PLAIN) jumped back and forth through a paper, shown by the consistent usage of jumping to answers in Figure 5.6, while participants without the Key Question

Index and Answer Gists often read papers top to bottom, once through. This behavior is exemplified in Figure 5.8, which plots participants' position in a paper over the course of reading with and without the Key Question Index and Answer Gists.

The key question index influenced reading behavior in several observable ways. First, readers who had access to the key question index dwelled significantly longer on the sections that they encountered while reading. When readers had access to the key question index, their dwell time in any one position in the paper lasted an average of 5.19 seconds ($\sigma = 7.72$), compared to 3.34 seconds ($\sigma = 10.99$) for those without the key question index (paired samples t-test, $t_{19} = 4.14, p < 0.001$).

Second, participants with the key question index tended to read papers piecemeal and non-linearly, in contrast to linear reading behavior of those without the feature. See Figure 5.8, where it can be observed that readers with the key question index jumped from one section of a paper to another often in a reading session. Participants jumped on average over 10 times per session, based on the number of times they used the key question index, and usually within a few minutes of starting the reading task, shown by the number of readers who used the key question index within the first 2 minutes of the study in Figure 5.6.

Third, readers with the key question index tended to fixate on the beginning and end of the paper, rather than the middle matter. These areas often contained the introduction and discussion sections. Participants in our formative studies often felt that these sections contained the most important high-level takeaways. In contrast, readers without the key question index tended to distribute their attention more uniformly across a paper, spending considerable time on the middle matter of a paper. When readers had access to the key question index, their average total time spent on pages containing either the abstract, introduction, discussion or conclusion was 9 minutes and 8.86 seconds (out of a total of 10 minutes of reading) ($\sigma = 3$ minutes and 44.60 seconds), compared to 6 minutes and 48.99 seconds ($\sigma = 3$ minutes and 6.44 seconds) for those without the key question index. This difference was significant (paired samples t-test, $t_{19} = 4.84, p < 0.05$). While we cannot say that there was no information of interest in the middle sections, the reading patterns suggest that the presence of the key question index led to a more selective reading concentrating on many sections that contain important information for non-expert readers.

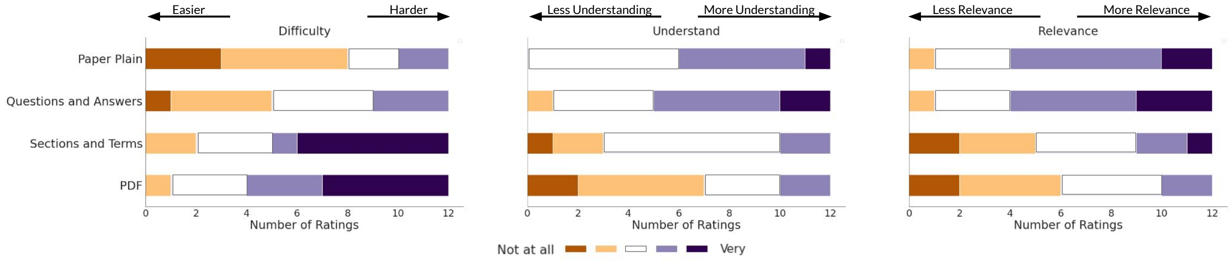


Figure 5.9: Readers' subjective reading difficulty, confidence in their understanding of the paper and ability to get all relevant information from the paper for different variants of Paper Plain.

5.7.2 How does Paper Plain affect participants' self-reported reading difficulty, understanding, and ability to identify relevant information?

Figure 5.9 plots the reading difficulty, understanding, and relevance scores for both papers across each system variant, and we observe significant differences between them. This is also reflected in our mixed-effects model F-test ($p < 0.001$ for all three measurements after Holm-Bonferroni [Holm, 1979] correction). We report estimated fixed-effect coefficients in Appendix B.4 and instead discuss more interpretable results comparing system variants in this section. We report here on medians (denoted $\tilde{x}$) for each subjective rating because ratings were scored on Likert-style scales.

Table 5.3 presents the differences in the fixed-effects between all pairs of interface variants. Participants with Paper Plain were significantly more confident that they got all relevant information from the papers ($\tilde{x} = 4.00$, $\sigma = 0.87$, with 5.00 being the most confident) and understood the papers ($\tilde{x} = 3.50$, $\sigma = 0.69$), compared to the PDF reader baseline ($\tilde{x} = 2.50$, $\sigma = 1.00$ and $\tilde{x} = 2.00$, $\sigma = 1.00$). Participants with Paper Plain also rated their reading difficulty significantly lower ($\tilde{x} = 2.00$, $\sigma = 1.06$, with 5.00 being hardest) compared to participants who had the PDF reader baseline ($\tilde{x} = 4.00$, $\sigma = 1.04$).

Building on our qualitative findings in RQ1, we saw that participants' use of Paper Plain's features made them more confident in their ability to find information important to them in the papers. This support manifested differently based on the Paper Plain features available to a participant. There were two major ways we saw Paper Plain improving participants' reading experience: providing in-situ support with the Section Gists and Term Definitions and a high-level overview with the Key Question Index and Answer Gists.

	$PP - QA$	p	$PP - SD$	p	$PP - PDF$	p
Reading Difficulty (1–5)	-0.344	0.7481	-1.485	**0.0011**	-1.983	<**.0001**
Understand (1–5)	-0.104	0.9842	0.719	0.0866	1.177	**0.0020**
Relevance (1–5)	-0.193	0.9133	0.752	0.0772	1.167	**0.0030**

	$QA - SD$	p	$QA - PDF$	p	$SD - PDF$	p
Reading Difficulty (1–5)	-1.141	**0.0132**	-1.639	**0.0003**	-0.498	0.4786
Understand (1–5)	0.823	**0.0401**	1.281	**0.0008**	0.457	0.4106
Relevance (1–5)	0.946	**0.0183**	1.361	**0.0006**	0.415	0.5093

Table 5.3: Post-hoc (two-sided) tests for pairwise differences in fixed-effects estimates between interfaces. This table reports the difference in fixed-effects estimates $i - j$ and Holm-Bonferroni-corrected p-values [Holm, 1979] under our mixed-effects model, where i and j correspond to interface options — PP = Paper Plain, QA = Key Question Index and Answer Gists, SD = Both Section Gists and Term Definitions, and PDF = PDF baseline. For example, in the column for $PP - PDF$ and row for "Reading Difficulty," we can interpret the result as Paper Plain is associated with, on average, 1.983 points lower rating of reading difficulty than a PDF baseline when controlling for participant and paper. Statistically significant p-values are bold. More details about this analysis are in Appendix B.4.

The in-situ nature of the Section Gists and Term Definitions helped participants understand the paper without switching contexts (P2, 6, 7, 11, 16–17, 19). For example, P19 found the Term Definitions useful for understanding the paper and the more specific medications it mentioned. P2 reported that the Section Gists were helpful to understand the paper text in a language they understood and P17 found the Section Gists broke "down complicated medical text into layman's terms that are easily understandable and helped to keep up with the flow of the article." Our observations suggest that Paper Plain's in-situ support successfully provided information to participants with minimal context switching.

Participants also used the Key Question Index and Answer Gists to get an overview of a paper quickly and easily, boosting their confidence to then dive in to the paper text (P2–3, 9–11, 20). P9 reported that "with so many sample sizes, numbers, and information to go through, it was helpful to get a summary to direct my reading and understanding." P20 mirrored this sentiment, explaining that the simplified answers gave them the gist of the entire paper quickly, so they had more time to get into its details. P3 illustrated these benefits well, explaining that the Key Question Index and Answer Gists were "beneficial because. . . I could have a baseline of what to expect and my mind would not have to pull in many random parts of information and could easily block what I did not need when I only needed a couple bits while I was reading." Similar to how the Key Question Index and Answer Gists supported a non-linear reading strategy (described in Section 5.7.1), it seemed that the Key Question Index and Answer Gists allowed participants to get a general sense of a paper early and focus their reading to sections they found most important.

The Key Question Index and Answer Gists provided useful guidance for readers in this reading context. As shown in Table 5.3, readers that only had the Key Question Index and Answer Gists rated their reading difficulty significantly lower ($\tilde{x} = 3.00$, $\sigma = 0.97$) than participants with the baseline PDF reader ($\tilde{x} = 4.00$, $\sigma = 1.04$). Participants with the Key Question Index and Answer Gists also rated their confidence that they got all relevant information in a paper ($\tilde{x} = 4.00$, $\sigma = 0.94$) and that they understood the paper ($\tilde{x} = 4.00$, $\sigma = 0.89$) significantly higher compared to the PDF baseline ($\tilde{x} = 2.50$, $\sigma = 1.00$ $\tilde{x} = 2.00$, $\sigma = 1.00$).

The preference for the Key Question Index and Answer Gists illustrates the importance of the novel guidance technique in PAPER PLAIN. 18 out of 20 readers who had the Key Question Index and Answer Gists in at least one condition selected the Key Question index, not the Answer Gists, as the most helpful feature. P18, who selected the Key Question index as the most helpful feature, said they would absolutely use the questions, because "...medical papers are difficult to follow and understand without guidance." Participants reported liking the Key Questions for quickly finding and understanding relevant information (P2, 4, 7-10, 13, 18-20). P4 reported not having any idea how to approach the research papers, and the Key Questions helped guide them to questions they should have. P7 used the Key Questions because "It answered questions that I would have had if it was me in the scenario … it helped highlight directly to the passage instead of having to sift through all of the information." These findings support the insight of this chapter that novel guidance-offering features are important for supporting readers in approaching medical research papers.

5.7.3 Do we observe any difference in paper comprehension when participants use PAPER PLAIN?

Participants on average answered 3.73 ($\sigma = 1.51$) out of 7 multiple choice questions correctly. This is well above a random chance baseline of 1.75 questions correct (i.e., 25% of 7 questions correct, with each question having four answers). This suggests that to some degree, the questions were able to measure an understanding of the paper that arose from completing the reading task. There was no significant effect of

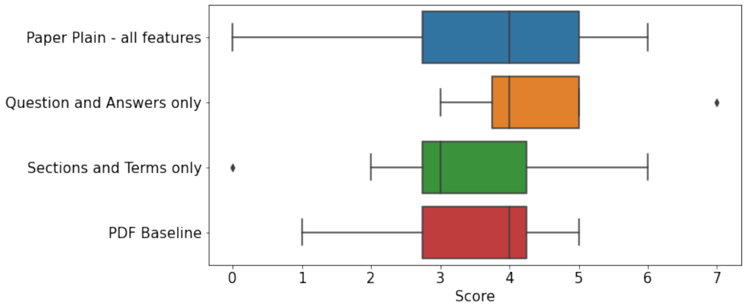


Figure 5.10: Comprehension scores for variants of PAPER PLAIN. Score is number of comprehension questions answered correctly out of seven for each paper.

either interface or paper on multiple choice scores under the mixed effects model F-test ($F_{4,20} = 1.38, p = 0.2672$).

According to a follow-up non-inferiority t-test, participants scored no worse on the multiple choice questions with PAPER PLAIN ($\mu = 3.67, \sigma = 1.78$) compared to the PDF reader ($\mu = 3.50, \sigma = 1.31$, $t_{28} = 1.82$, $p < 0.05$). Figure 5.10 compares the scores of participants on the multiple choice instrument, grouped by interface variant. High variability in scores σ ($\sigma = 1.31$ for the PDF reader and $\sigma = 1.78$ for PAPER PLAIN) makes it difficult to determine what effect PAPER PLAIN might have had. August et al. [2022b] discusses our instrument for measuring comprehension and future systems for improving paper comprehension.

Post-hoc analysis suggests that some multiple choice questions were answered correctly more often with the key question index than without. While no questions could be answered by consulting the key question sidebar (see Section 5.6.1), some questions were answerable by reading a passage that was highlighted by clicking a question in the question index (for example, the first and third questions in Table 5.2). Participants answered these questions correctly more often when they had the key question index than when they did not ($\mu = 3.00, \sigma = 1.48$ vs. $\mu = 2.50, \sigma = 1.38$ for 5 such questions in the Disc Herniation paper; $\mu = 2.17, \sigma = 0.94$ vs. $\mu = 1.58, \sigma = 0.67$ for 3 such questions in the Lupus paper). This trend can be seen in Figure 5.11. While the trend is not statistically significant (paired samples t-test $t_{26} = 1.89, p = .07$), it suggests the interesting possibility that features of PAPER PLAIN may lead to better understanding of some aspects of a paper more than others.

Participants generally found the generated gists useful, and when confronted with vague system pre-

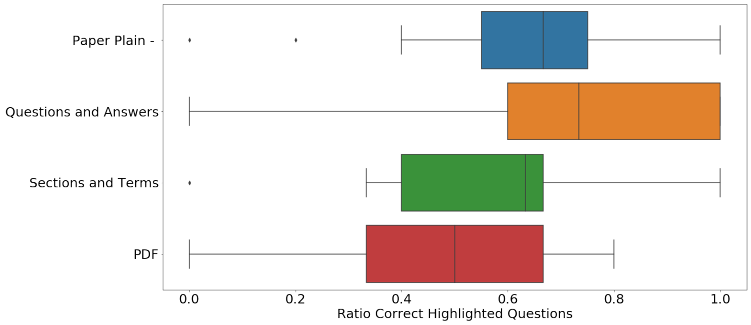


Figure 5.11: For a subset of questions that were highlighted through interaction with the key question index, scores on multiple choice questions appeared to improve when readers had access to the key question index. This effect was more pronounced when participants only had access to the Key Question Index and Answer Gists, suggesting that other features might have distracted from answering these multiple choice questions.

dictions and generations, participants usually, though not always, used the original text to fill in missing information. We observed one participant, P11, who read only the Answer Gists for a paper and rated their confidence for understanding the paper at a 5 (the highest) while rating their reading difficulty at a 1 (the easiest). However, this participant got only 2 out of 7 comprehension questions correct, well below the average of 3.73 for all participants, suggesting that the gists were not sufficient for answering many of the comprehension questions. In contrast to this participant, other participants reported that the gists (both Answer and Section) were helpful as a starting point for understanding, but looked at the underlying text, too. Some participants also reported that information in the gists was vague or missed information in the original text, necessitating reading the original (P10, 22, 24). P24 made sure to double check all the information in the gists with the original sections because the gists were automatically generated. While they did not find incorrect information in the gists, they did report that the Sections Gists sometimes were vague or reported on details less important to them while leaving out details that were more important to them (e.g., the percent of people who recovered from a surgery was reported in a section but not the Section Gist). P10 also noticed that the area surrounding some of the highlighted answers contained useful information, and so made sure to go back through the answering passages to read the surrounding text in addition to the Answer Gist and passage.

There was also some indication that participants wanted different levels of specificity from the gists (P3, 14). For example, P3 reported that the gists were “too far simplified” while P14 found the gists “Only marginally more helpful than reading the paper itself.” In these cases participants often opted for the original text.

5.8 Summary

This chapter illustrated how interactive information interfaces can redesign scientific language for healthcare consumers that need it. In particular, we develop PAPER PLAIN, an interactive system inspired by our writing strategies that augments the paper itself with new affordances to make the paper more approachable to healthcare consumers.

PAPER PLAIN uses features like plain language summaries and a key question index that leverage many of the strategies identified in Chapter 4. The plain language gists use strategies like JARGON and EXPLANATION to overcome complex language. The key questions emphasize the main findings and real world impacts of a paper (drawing on the MAIN and IMPACT strategies).

Participants used and appreciated PAPER PLAIN's features throughout reading a paper. Readers used the Section Gists to easily make sense of dense passages while reading a paper and leveraged the guidance of the Key Question Index and Answer Gists to quickly find text that was informative for them. All but one participant said they would use PAPER PLAIN to read medical papers.

Participants who used PAPER PLAIN also rated their reading difficulty significantly lower and rated their confidence they got all relevant information from a paper significantly higher. Participants found it easier to read with PAPER PLAIN because it gave them an approachable overview of a paper with the Key Question Index and Answer Gists and helped them understand dense text in the context of the paper with the in-situ Term Definitions and Section Gists.

In summary, we take these results to indicate the promise of PAPER PLAIN for assisting healthcare consumers in making sense of medical research papers. Our formative study revealed that non-expert readers, although motivated, have limited time and energy to engage with medical literature. Reading medical papers can be overwhelming and demoralizing, limiting people's ability to collect relevant information from multiple papers. The sustained usage of PAPER PLAIN's features and positive response from participants in our usability study suggest that such a tool would be a welcome addition to healthcare consumers' information seeking toolkit.

Chapter 6

The Benefit and Feasibility of Reader-Sensitive Scientific Language Complexity

PAPER PLAIN redesigned medical research papers to be approachable to a general audience, but the system provided only a single version of text for all potential readers, limiting its ability to communicate with the diverse stakeholders of science. We observed some indication of this from the user study, where some participants commented that the gists were too simplified, while others considered them barely more intelligible than the paper itself. In this chapter we determine the benefit and feasibility of adjusting language beyond a one-size-fits-all version. In a within-subjects study ($N = 200$) we presented participants with expert-authored summaries of scientific articles for different education levels. We found that the least complex summaries were easier to understand and perceived as more interesting by participants with low topic familiarity. In contrast, those with high topic familiarity gained no such benefit from low complexity text. To understand the feasibility of generating such summaries, we developed a prototype leveraging NLP advances in summarization and controllable generation. We observed highly similar impacts of complexity using generated summaries curated for factuality in this follow-up study ($N = 194$), establishing the feasibility of generating summaries at varying complexities. This chapter includes materials originally published in August et al. [2022a] and August et al. [2023].

6.1 Introduction

People have different knowledge that can impact how they respond to scientific information [Nisbet and Scheufele, 2009b; Forzani, 2016; Bliss, 2019]. Past work has shown that personalizing science communication to a readers' interests and background (e.g., a patient's health history [Skinner et al., 1994]) can improve engagement and understanding [Strecher et al., 1994; Skinner et al., 1994; Marco et al., 2006]. The models PAPER PLAIN used for generating gists can generate simplified summaries given target specifications (e.g., complexity level) [Guo et al., 2021; August et al., 2022a]. However, it remains unknown whether adapting the language of scientific summaries to different levels of language complexity could be beneficial and what manner of simplification is best suited for a given reader. Is there a single best complexity for all potential readers, or does it depend on a readers background? There are many automated measures that capture different notions of language complexity (for a review, see [Pitler and Nenkova, 2008; Collins-Thompson, 2014]) but little evidence of how adjusting these complexity measures might affect communication with different people.

In this chapter we measure whether novice readers can benefit more from simpler summaries and evaluate the possibility of automatic, adaptive scientific language. In a within-subjects experiment ($N = 200$) where participants with varying familiarity of a topic read expert-written plain language summaries of scientific papers at three levels of complexity. We found that language complexity significantly impacted self-reported and quantitatively measured reading ease and understanding measures. Topic familiarity also mattered for determining the ideal summary for a reader based on reading experience measures: as we would expect, participants who were not familiar with a summary's topic responded more positively to the lowest complexity, while participants with more topic familiarity did not benefit from lower complexity summaries.

We then ran an additional user study using machine-generated summaries curated for factuality in place of expert-written ones. We develop a novel decoding-time controllable generation technique for generating summaries at different levels of complexity. We operationalize complexity based on readability and science communication research [Pitler and Nenkova, 2008; Gardner and Davies, 2013; Leroy et al., 2010]. We envision models helping experts to quickly create multiple versions of a plain language summary, thereby enabling more scalable text flexible to different audiences. In our follow-up study, we found that generated summaries led to similar benefits in self-reported reading ease and understanding for readers of different topic familiarities. Our results suggest that such augmentation of scientific communication is feasible with expert-in-the-loop support. We discuss the implications of our findings, specifically on the importance of designing language to communicate with an intended audience, and how automated methods can assist human efforts to do so.

6.2 Study 1: Effect of Expert-Written Summaries at Different Complexities

We conducted a within-subjects experiment to determine if changing language complexity matters when communicating scientific information and how a reader's background knowledge influences their response to language complexity. Specifically, our study answered the following questions:

1. How does language complexity affect participants' self-reported reading ease, perception of valuable information, understanding, and interest?
2. How does language complexity affect reader comprehension?
3. Does reader familiarity with a summary's topic impact the effects of language complexity?

6.2.1 Methods

Our experiment had participants answer questions about their scientific background, read summaries of scientific papers written at three different levels of complexity, and answer questions about the summaries.

Materials

Article selection We selected research papers that had public appeal by sampling papers posted and widely discussed in the large subreddit *r/science* in 2019. We randomly sampled 10 papers posted on *r/science* that contained a link to a research paper (as opposed to a press release or news article), and that had a score within the top 10% of posts containing research papers. We used the `PSAW` Python PushShift API for accessing *r/science*.[1] The papers ranged in topics from public policy to nanotechnology, reflecting the breadth of research papers posted and discussed on *r/science*.

[1] `https://psaw.readthedocs.io/en/latest/`

Authoring the summaries In order to test how summaries written at different complexities affected communicating scientific information, we developed multiple summaries of the research papers. We broke down the summaries into sections answering key questions about the paper following prior work showing that summaries broken down with headers were preferred by general audience readers [Santesso et al., 2015]. The key questions were based on our prior work studying the key information that science communicators focus on in a paper (Chapter 4) and from our study of questions general audience readers found useful to determine relevant information in research papers (Chapter 5). The questions were:

1. What did the paper want to find out?
2. What did the paper do?
3. What did the paper find?
4. What are the limitations of the findings?
5. What is the real world impact of this work?

An expert science writer with over 5 years of science communication experience crafted three versions of each summary. Each version was written for a different audience of a certain education level: a high school student, a college educated adult, and a researcher. We define these three complexity levels as Low, Medium, and High. For the Low and Medium versions, the writer summarized information from the paper that answered each of the key questions. For the High version, the writer selected sentences from the original paper that answered the questions. The High version assumed that the reader has expert knowledge in the paper's topics. In this case, the most faithful and detailed language (i.e., language from the original paper) was preferred. Because the original paper text used a different voice than the other two versions, we lightly edited the High version by changing "we" to "the researchers."

We validated that the different versions were distinct by having one author review each summary version and provide feedback to the writer on language complexity between the three versions in four weekly meetings, as well as asynchronously with Google Docs. The rest of the authors reviewed the completed summaries to determine that each version was distinct from the others in language complexity while reporting on the same information. The writer was paid $17.22 USD per hour. Crafting all 30 summaries took

Complexity Level	Summary
High	These results demonstrate an unprecedented opportunity for **development of these nanorgs as renewable sugar-free microbial factories** for the production of biofuels and chemicals using sunlight in a scalable process, but also as a means of externally regulating the cellular function of living cells using electromagnetic stimuli such as light, sound, or magnetic field.
Medium	This work is some of the first to examine the feasibility of **interfacing nanoscale materials with living cells to create "living factories"** capable of producing large biochemical yields. This work also paves the way for medicine to investigate the potential for wireless cellular monitoring by triggering selective cell functions, which could have broader implications for diagnostic and therapeutic applications of this technology.
Low	This work is some of the first to be done investigating **the possibility of using nanoscale materials inside living cells to create "living factories",** which has implications for the renewable energy industry as well as furthering the boundaries of microbial science. It also investigates how we can selectively trigger certain cell functions by using light or other electromagnetic radiation, which has far-ranging applications for medicine.

Table 6.1: Examples of the expert-written summaries. These summaries were under the heading "What are the real world impact of the findings?" for the same paper. Bolded text are example of changes in complexity across the three versions while reporting on roughly the same information.

Complexity Level	# words$_{\text{std}}$ ↑	# sentences ↑	AVL ↑	TE ↑	Function Words ↓	GPT ppl. ↑
High	$483.10_{107.02}$	$16.70_{4.03}$	$0.39_{0.03}$	$0.55_{0.04}$	$0.27_{0.03}$	$94.60_{30.91}$
Medium	$369.60_{82.56}$	$12.10_{1.97}$	$0.38_{0.05}$	$0.47_{0.05}$	$0.31_{0.02}$	$60.08_{14.84}$
Low	$358.90_{98.92}$	$11.50_{3.21}$	$0.37_{0.05}$	$0.43_{0.05}$	$0.32_{0.02}$	$53.68_{9.96}$

Table 6.2: Differences in automated complexity measures between expert-written summary versions. Arrows denote expected increase (↑) or decrease (↓) in measure as complexity increases. **TE** = Thing Explainer out-of-vocabulary words, **GPT ppl** = GPT language model perplexity.

the writer and authors two months of asynchronous work, pointing to the difficulty in scaling the writing of multiple summary versions.

Analyzing the summaries We collected a total of 30 expert written summaries, 3 for each of the 10 articles. The summaries had on average 404 words (std=109). Table 6.1 provides examples of the summaries. As an additional validation that the summary versions differ in language complexity, we report on five automated measures of language complexity. These measures are not meant to be an exhaustive list (for a review, see [Pitler and Nenkova, 2008]), but a selection of measures that capture different elements of complexity important to definitions.

Academic Vocabulary List (AVL) occurrences The AVL is a list of academic vocabulary drawn from corpora spanning many scientific disciplines [Gardner and Davies, 2013]. We measure the fraction of AVL words in a generated definition.

Thing Explainer out-of-vocabulary (TE): We count the ratio of words outside the top 1,000 most common words in English. The words are based on Wiktionary's contemporary fiction frequency list.[2] This method was popularized by the book *Thing Explainer*, which explains scientific concepts using only the 1,000 most frequent words in English [Munroe, 2017].

Sentence length Sentence length is a commonly used metric for document level complexity and is part of many classic readability measures [Pitler and Nenkova, 2008]. Longer sentences are often considered more complex.

Function words In medical communication, the proportion of function words (e.g., prepositions, auxiliary or verbs) was found to be positively correlated with perceived and actual readability [Leroy et al., 2008, 2010]. We measure the proportion of function words in a sentence using `scispacy` [Neumann et al., 2019].

Language model perplexity (GPT ppl.) Language models are systems for predicting words in a sequence. The perplexity of the model is a measure of how different a sequence of text is from the language the model was trained on. Perplexity has been found to correlate with perceived and actual reading difficulty [Pitler and Nenkova, 2008; Collins-Thompson, 2014]. We use the GPT model [Radford and Narasimhan, 2018] to measure language model perplexity, as it was trained on common English (as opposed to scientific text).

Table 6.2 reports on the complexity measures for the summaries. The High versions were on average longer (mean=483.1, std=107.0 words) than either the Low (mean=358.9, std=98.9) or Medium versions (mean=369.6, std=82.6). The High versions also used more academic vocabulary, more words outside of the top 1,000 most common English words, had a lower proportion of function words, and higher language model perplexity. The Medium versions also had more words outside the top 1,000 compared to the Low versions, though differences in function word proportion and language model perplexity were much smaller

[2] `https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Contemporary_fiction`

between these two versions. While overall the measures show a continuum of complexity from High to Low, the small differences in some measures suggest that writing for different audiences can lead to nuanced differences in language not well captured by automated measures.

Participants

Recruitment We recruited participants on Amazon Mechanical Turk. Participants were paid $2.50 for the study, in line with local minimum wage. Participants were required to have completed over 1,000 HITs with a minimum approval rating of 95%, be US-based, and be master Turkers. This study was approved by our institution's IRB. The slogan for the study was, "Read about interesting scientific findings and answer questions about your experience."

		# Participants
Age	0-19	0
	20-29	14
	30-39	68
	40-49	71
	50-59	30
	60-69	14
	70-79	3
	80+	0
Gender	Male	99
	Female	99
	Prefer not to answer	2
Education	High school	59
	College	117
	Graduate school	19
	Professional school	5
# STEM courses after high school	0	36
	1–3	90
	4–6	41
	7–9	11
	≥10	22

(a) Participant demographics

Topic Familiarity	# Ratings
1	361
2	116
3	97
4	26
5	0
Total	600

(b) Topic familiarity based on question "How familiar are you with the topic of this article?" 1="I have never heard about this topic before", and 5="I have written research papers on this topic."

Table 6.3: Participant demographics (a) and topic familiarity (b) for study on expert-written summary versions.

Demographics A total of 200 participants completed the study. Table 6.3a details participant demographics. The majority of participants had completed at least college (117), 59 had only completed high school. The average age of participants was 44.54 years ($\sigma = 10.43$). Participants displayed a diversity of STEM educations: 22 participants had completed over 10 STEM courses since high school, while 36 had completed 0.

Topic familiarity After each summary, participants rated their familiarity with the article's topic on a 1—5 Likert-style scale based on the question: "How familiar are you with the topic of this article?"[3] with 1 being "I have never heard about this topic before" and 5 being "I have written research papers on this topic." Table 6.3b details the 600 topic familiarity ratings (200 participants each rated 3 summaries). Participants generally had little familiarity with the topic of the summarized articles. Out of 600 total ratings, 361 (60.1%) were rated as '1', the lowest level of familiarity.

Measures

Reading experience ratings We collected subjective ratings to understand how the different complexity levels affected participants' reading experience. Participants completed the ratings after reading each summary. These included:

1. **Reading ease**: Participants rated their reading difficulty on a 1–5 Likert-style scale based on the question: "How easy was it for you to read the article?"

2. **Understanding**: Participants rated their understanding of the paper on a 1–5 Likert-style scale based on the question: "How confident do you feel in your understanding of the article?"

3. **Interest**: Participants rated how interesting they found a summary on a 1–5 Likert-style scale based on the question: "How interesting did you find the article?"

4. **Value**: Participants rated how valuable they found the information in the summary on a 1–5 Likert-style scale based on the question: "How much would you agree that this article contained valuable information?"

[3] Because participants were only ever presented summaries, not the original paper, in the study the summaries were referred to as 'articles.'

Complexity	Question	Answer	Associated text
High	What differences in children were associated with higher maternal intake of nuts in early pregnancy?	The children with higher maternal nut intake had improved neuropsychological development	In this longitudinal cohort study, the researchers found that higher maternal intake of nuts in early pregnancy was associated with enhanced neuropsychological development in offspring at 1.5, 5, and 8 years old.
Medium	What happened to children of mothers who had a high nut intake during pregnancy?	The children of mothers who had a high nut intake during pregnancy performed better on tests of memory and attention	Researchers found that pregnant mothers who ate a diet rich in nuts gave birth to children who performed better in tests of attention, memory, and executive function.
Low	What happened to children of mothers who ate the highest amount of nuts while pregnant?	The children of mothers who ate the highest amount of nuts while pregnant had the highest scores on tests of memory and attention	The researchers found that the children of women who ate the highest amount of nuts while pregnant had the highest scores in tests of attention and memory.

Table 6.4: Examples of comprehension questions for each complexity level with answering excerpts from the summaries.

Multiple-choice comprehension scores To measure retention of specific facts in the article, we developed two multiple choice questions for each article. Because the questions asked about information in the summaries, there was a risk that the questions' language would be more aligned with one complexity version over another. To mitigate this risk, for every multiple choice question, we developed a version aligned with each complexity level while still ensuring all versions of a given question tested participants on the same information from the article. For each article, participants were assigned the corresponding question version that aligned with their assigned summary complexity level. For example, the Low question might include the answer "asking about eating habits," while in the High version the answer would be "administering a dietary questionnaire." Each question had four possible answers. Answer choices were also edited to have similar complexity as the question and summary. One author wrote these questions and all authors reviewed them. Table 6.4 shows an example multiple choice question and its versions for each complexity level. Each question was scored as True or False. We report the accuracy of participants, or how often they answered a given question correctly.

Procedure

At the start of the experiment participants filled out a demographics questionnaire, including questions on their education, STEM experience and interest in different scientific subjects. After the demographic questionnaire, participants read three article summaries, each with a different complexity. The summaries and complexity levels were randomized. Each participant saw one of each complexity in random order. Summaries were displayed as a title and a dropdown list of sections. Participants could open multiple dropdowns at once.

Participants were asked to read the summaries for at least 30 seconds, though they could read for as long as they wanted. If participants clicked the continue button before 30 seconds, they were prompted to read for at least 30 seconds. They could ignore this prompt by clicking the continue button again. Participants on average took 142 seconds per article (std=103 seconds). Participants then answered questions on their topic familiarity, reading experience and comprehension.

Analysis

We compared measures across the complexity versions using mixed-effects linear models [Lindstrom and Bates, 1990]. To accurately identify the effect complexity has on our measures and its interaction with topic familiarity, we define two models for each measure.

1. LMM_{full}: Containing fixed effects for the complexity version, topic familiarity, an interaction term for familiarity and complexity, and random effects for paper and participant IDs.

2. LMM_{none}: Containing a fixed effect for topic familiarity and random effects for paper and participant IDs.

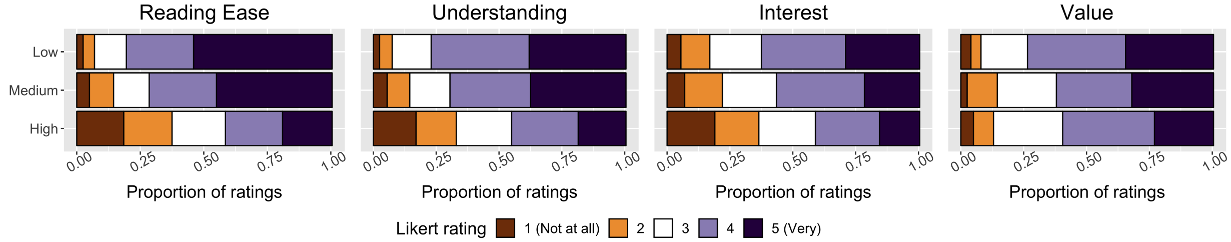


Figure 6.1: Distribution of ratings for each subjective reading experience measure across complexity levels. The ratings were based on the following questions: Reading ease: "How easy was it for you to read the article?", Value: "How much would you agree that this article contained valuable information?", Understand: "How confident do you feel in your understanding of the article?", Interest: "How interesting did you find the article?" Notice the greater number of high ratings (blue) and fewer low ratings (orange) as participants are presented with less complex summaries.

With these models we evaluate how complexity affects reading measures (i.e, experience ratings, like reading ease, or multiple choice comprehension score) by comparing the model goodness-of-fit between LMM_{full} and LMM_{none} using the χ^2 likelihood-ratio test. If LMM_{full} has a significantly stronger fit, this suggests that complexity has a significant effect on that reading measure. We fit these models using a Gaussian link function for reading experience ratings and a Logistic link function for the comprehension scores.

We quantify the differences between the complexity versions and their interaction with topic familiarity by conducting post-hoc two-sided t-tests for pairwise comparisons. For these analyses we use the PYMER4 Python package for fitting the models and pairwise comparisons, the LMERTEST R package for the likelihood-ratio tests, and STATSMODELS Python package for multiple hypotheses corrections.

6.2.2 Results

How does language complexity affect participants' self-reported reading ease, perception of valuable information, understanding, and interest?

We first investigated our question on how language complexity impacted participants' reading experience measures (reading ease, perception of value, interest, and understanding). We fit LMM_{full} and LMM_{none} for each of the reading experience measures. The likelihood-ratio test showed that LMM_{full} had a significantly better fit than LMM_{none} ($p < 0.005$) for all reading experience measures. Table C.2 in the Appendix lists all χ^2 and (Holm-Bonferroni corrected) p values.

Figure 6.1 plots participants' ratings across summary complexities. We observe that, across all measures, there is a greater number of high ratings and fewer low ratings as participants are presented with less complex summaries. We conducted post-hoc tests to examine the differences d in subjective ratings between pairs of complexity levels estimated by LMM_{full}. These differences allow us to the see the effects of complexity levels on participant ratings while controlling for participant and paper. Overall participants found the Low summaries most appealing. Compared to the High summaries, participants rated Low summaries as significantly easier to read ($d_{ease} = 0.894$, $p < 0.0001$), understand ($d_{understand} = 0.593$, $p < 0.0001$), and more interesting ($d_{interest} = 0.373$, $p = 0.042$). Participants also rated the Medium summaries as significantly easier to read and understand compared to the High summaries ($d_{ease} = 0.655$, $p < 0.0001$; $d_{understand} = 0.390$, $p = 0.006$). See the "All" rows in Table 6.5 for pairwise contrasts d and their p-values.

How does language complexity affect reader comprehension?

We fit LMM_{full} and LMM_{none} for multiple choice comprehension scores. The likelihood-ratio test showed that LMM_{full} had a significantly better fit than LMM_{none} ($p < 0.01$) for multiple choice scores.[4]

Participants generally performed worse on the comprehension test when presented High summaries ($acc = 71\%$). Post-hoc tests showed that participants answered significantly more multiple choice questions correctly when they read the Low ($acc = 79.75\%, p = 0.022$) and Medium ($acc = 79.50\%, p = 0.029$) summaries. See the "All" row in Table 6.5 pairwise contrasts and their p-values.

Does reader familiarity with a summary's topic impact the effects of language complexity?

Post-hoc tests of pairwise differences in reading experience measures across complexity and topic familiarity revealed two results.

First, topic familiarity was a strong indicator of reading experience measures. Looking at Figure 6.2, as familiarity increases, ratings across all metrics and complexity levels generally go up (i.e, the orange bars shrink while the dark blue bars grow).

Second, topic familiarity interacted with complexity to equalize reading measures. Also apparent from

[4] We also tested the fit of a variant of LMM_{full} that additionally included a question-specific random effect, since participants answered multiple questions per article-summary. We found model fit did not significantly improve over LMM_{full} ($\chi^2 = 2.319, p = 0.2556$) and so proceeded with LMM_{full}.

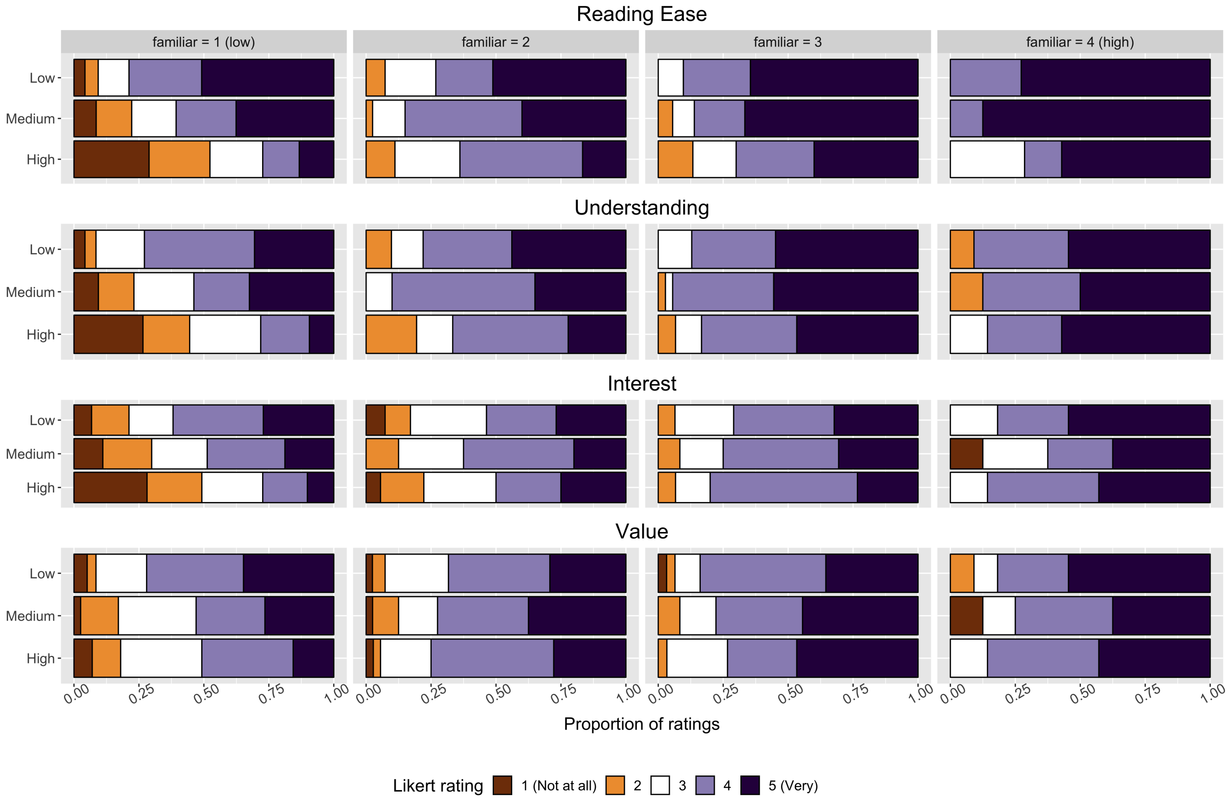


Figure 6.2: Distribution of ratings for each reading experience measure across complexity and participant topic familiarity. (a) As familiarity increases, the rating levels across all metrics generally increases. (b) As familiarity increases, the distribution over ratings become more similar across complexity levels.

Figure 6.2: at low familiarity, rating distribution are most different across the complexity levels. As familiarity increases, though, there were fewer low ratings and more high ratings for all complexity levels. For reading ease and understanding, this change is gradual, with differences between complexity levels still perceivable at familiarity level 3; for interest and value, this equalization happens as early as familiarity level 2. This is further supported by Table 6.5, where we see significant differences between complexity levels for lower familiarities, but no significant difference at the higher familiarities.

6.3 Study 2: Feasibility of Generating Summaries at Different Complexities

Our results contribute to research showing that tailoring content to different audiences can lead to more effective science communication [Strecher et al., 1994; Skinner et al., 1994; Marco et al., 2006]. We extend this research by identifying reader topic familiarity as an important element to determine the ideal complexity to improve reader interest in a summary. Our first study, like prior work exploring alternative content [Adar et al., 2017], required each version to be manually created. This involved meeting regularly with a writer to create alternate versions of a scientific summary. We next explore the feasibility of partially automating this process with an additional user study where instead of using expert-written summaries, we use NLP methods to generate summaries. Our study was guided by similar questions as Section 6.2, with the nuance that in this case, we were curious if automated generations could replicate the findings we observed initially.

6.3.1 Methods

Our second study closely followed the design of our first study (Section 6.2). It was a within-subjects experiment where participants read one of three versions of a paper summary. Rather than the Low and Medium summaries being written by an expert, in this second study they were generated using NLP models.

In our second study we compared across generated text, rather than comparing generated text with expert-written text, because our goal was to achieve the same variation in language complexity, and resulting differences in reader response based on topic familiarity, with generated text as we observed with expert-written text.

	Familiarity	d^{Lo-Me}	p	d^{Lo-Hi}	p	d^{Me-Hi}	p
Reading Ease	1	0.554	**<0.0001**	1.490	**<0.0001**	0.936	**<0.0001**
	2	0.103	0.621	0.782	**0.001**	0.679	**0.003**
	3	0.197	0.391	0.695	**0.013**	0.498	0.059
	4	0.101	0.817	0.609	0.544	0.508	0.588
	All	0.238	0.069	0.894	**<0.0001**	0.655	**<0.0001**
Understanding	1	0.458	**<0.0001**	1.160	**<0.0001**	0.701	**<0.0001**
	2	0.022	0.910	0.693	**0.002**	0.671	**0.002**
	3	0.172	0.597	0.391	0.240	0.219	0.597
	4	0.160	1.000	0.127	1.000	-0.033	1.000
	All	0.203	0.094	0.593	**<0.0001**	0.390	**0.006**
Interest	1	0.296	**0.021**	0.943	**<0.0001**	0.647	**<0.0001**
	2	-0.007	0.975	0.298	0.593	0.305	0.593
	3	0.024	1.000	-0.009	1.000	-0.033	1.000
	4	0.864	0.220	0.261	0.603	-0.603	0.520
	All	0.294	0.085	0.373	**0.042**	0.079	0.613
Value	1	0.314	**0.020**	0.509	**<0.0001**	0.195	0.104
	2	-0.012	1.000	0.009	1.000	0.021	1.000
	3	-0.087	1.000	-0.099	1.000	-0.012	1.000
	4	0.329	1.000	-0.123	1.000	-0.451	1.000
	All	0.136	0.996	0.074	1.00	-0.062	1.00
Comprehension	1	0.000	0.990	0.047	0.575	0.047	0.575
	2	0.054	0.406	0.160	0.053	0.106	0.234
	3	-0.015	0.836	0.137	0.142	0.152	0.105
	4	-0.014	1.000	0.122	1.000	0.137	1.000
	All	0.006	0.884	0.117	**0.022**	0.111	**0.029**

Table 6.5: Post-hoc (two-sided) tests for pairwise differences in fixed-effects estimates between complexity versions and across all participant topic familiarities. ‘All’ topic familiarity refers to pairwise differences across complexity levels without a topic familiarity subgroup (e.g., average difference across complexity levels.) This table reports the difference in fixed-effects estimates $i - j$ and Holm-Bonferroni-corrected p-values [Holm, 1979] under our mixed-effects model, where i and j correspond to complexity options. — Lo = Low, Me = Medium, and Hi = High. Statistically significant p-values are bold. For example, in the column for d^{Lo-Hi} and row for “Reading Ease,” and “1” topic familiarity we can interpret the result as participants with a 1 topic familiarity rated the Low complexity, on average, 0.894 points higher for reading ease (out of 5) compared to the High complexity when controlling for participant and paper.

Participants

		# Participants
Age	0-19	0
	20-29	50
	30-39	88
	40-49	33
	50-59	18
	60-69	4
	70-79	1
	80+	0
Gender	Male	97
	Female	96
	Prefer not to answer	1
Education	Pre-high school	1
	High school	31
	College	115
	Graduate school	42
	Professional school	6
# STEM courses after high school	0	21
	1–3	94
	4–6	57
	7–9	10
	≥10	13

(a) Participant demographics

Topic Familiarity	# Ratings
1	152
2	76
3	134
4	165
5	54
Total	581

(b) Topic familiarity based on question "How familiar are you with the topic of this article?" 1="I have never heard about this topic before", and 5="I have written research papers on this topic."

Table 6.6: Study 2: Participant demographics (a) and topic familiarity (b) for study on generated summary versions.

We recruited participants on Amazon Mechanical Turk with the same requirements as in Section 6.2. The only difference was that we did not require master Turkers, in order to increase recruitment. In a pilot study we observed comparable results between master and non-master workers.

After removing 12 participants who indicated technical difficulties or cheating on the study, a total of 194 participants completed the study. Similar to participants in Section 6.2, the majority of participants had completed at least college (115). 13 participants had completed over 10 STEM courses since high school, while 21 had completed 0. Compared to participants in Section 6.2, participants in our second study rated their familiarity higher overall. Table 6.6 provides participant demographic and familiarity rating details.

Materials

We introduce a new, lightweight method to generate summaries with different complexity via reranking. Past work has explored selecting candidate generations based discriminator scores to control for specific topics or discourse structure but found that it did not provide strong control [Dathathri et al., 2020; Gabriel et al., 2021a]. Because our generation task does not require topic shifts, we adapt this method by scoring and selecting candidates based on complexity discriminators. Below we describe the novel method underpinning our summaries.

Generating alternative plain language summaries To generate summary versions at different complexities, we start by generating candidate generations using GPT-3. GPT-3 is a language model commonly used in generation tasks, including plain language summarization [Brown et al., 2020]. We adapt a preset prompt for GPT-3 to generate summaries with varying complexity. The original prompt is "Summarize this for a second-grade student: [TEXT]" Our adapted prompts for GPT-3 were 14 alternate prompts, from "first-grade student" to "twelfth-grade student", along with "college student" and "college-educated adult."

GPT-3 was not designed to explicitly vary text complexity, so while generations might vary naturally in complexity due to the changes in prompt, there is no guarantee that prompts will align with complexity (i.e., prompting GPT-3 with "Summarize for a first grade student" will not necessarily lead to lower complexity than prompting with "tenth grade student"). In a preliminary analysis of the summaries, we found that the summaries, while trending toward simpler with lower grades, could still be quite complex in the first grade prompted version and much simpler at higher grade levels. Table C.3 in the Appendix provides examples of generations and associated prompts.

To align generations along a gradient of complexity, we score candidate generations based on logits from a discriminator trained to distinguish scientific journal text from science news text. While this method requires multiple generations to work properly, it does not require gradient or probability distribution updating during generation or any prior pretraining, allowing for much greater flexibility during generation (e.g., generating from a language model without access to vocabulary logits during generation).

We use a linear SVM discriminator, with features drawn from science communication and readability literature, discussed in Section 6.2.1. We train the discriminator on classifying sentences from either scientific journal abstracts or scientific news articles. Journal abstracts are sampled from the ArXiv dataset [Clement et al., 2019] and scientific news articles are sampled from our corpus of science news articles from Chapter 4.

We use this discriminator to rank the complexity of the GPT-3 generations. After scoring each generation for complexity, we select the generation with the highest and lowest complexity scores for the Low and Medium versions. For the High summaries, we use the original sentences extracted from the paper by the writer in Section 6.2.

There are automatic methods for scientific information extraction [Cohan et al., 2019] and PDF parsing [Lopez, 2009; Shen et al., 2021] that could in the future be used extract information directly from a research paper PDF. We leave such extensions to future work, as our goal was to explore the feasibility of automatically adjusting language complexity. Any errors introduced by other automated methods (e.g., incorrect text from PDF parsing) could muddy our ability to identify how alternate complexity levels perform in our envisioned context.

Assessing factuality in generated summaries A major limitation of current generative models are that they can generate text with meaning that was not part of the original information provided to the model [Maynez et al., 2020]. These errors are referred to as hallucinations, and can lead to factually incorrect generations [Maynez et al., 2020; Goyal and Durrett, 2021]. While there are methods for reducing hallucinations or encouraging factuality [Gabriel et al., 2021b; Lu et al., 2021; Laban et al., 2021], no automated method guarantees factual accuracy or fidelity to original text. In the context of science communication, such hallucinations can risk confusing, or worse, misinforming readers. A reader might trust a hallucinated result opposite to what was reported in the original paper [Devaraj et al., 2022], or be so confused by the contradictory evidence as to lose trust in the research.

Because of these risks, we advocate for NLP systems to be used in conjunction with experts. Plain language summaries are often written by researchers, editors, or science writers [Stoll et al., 2022; Shailes, 2017]. Authors could generate multiple versions of a summary and verify the factual accuracy. In this way, we could greatly lessen the workload of writing plain language summaries, make summaries adaptable to different audiences, and protect against factually incorrect generations.

In the context of our study, one author selected generations that did not contain factually incorrect

Model	# words$_{\text{std}}$ ↑	# sentences ↑	AVL ↑	TE ↑	Function Words ↓	GPT ppl. ↑
High	$483.10_{107.02}$	$16.70_{4.03}$	$0.39_{0.03}$	$0.55_{0.04}$	$0.27_{0.03}$	$94.60_{30.91}$
Medium	$529.20_{182.48}$	$20.80_{7.05}$	$0.40_{0.03}$	$0.51_{0.04}$	$0.31_{0.03}$	$64.14_{24.16}$
Low	$259.00_{43.42}$	$13.20_{1.75}$	$0.27_{0.04}$	$0.28_{0.05}$	$0.36_{0.03}$	$23.92_{5.71}$

Table 6.7: Study 2: Differences in automated complexity measures between generated summary versions. Arrows denote expected increase (↑) or decrease (↓) in measure as complexity increases. Note that here the High version is the same as from Section 6.2.

information, acting as the expert for checking generated summaries before publishing. Out of 120 generated summaries (6 questions × 10 papers × 2 complexities), 14 generations contained incorrect information. In all 14 cases, a replacement was found by selecting from at most 6 alternative generations. The average number of generations the author looked at to find a replacement was 2.36. Appendix C.1 contains more information on hallucinations in our generated summaries.

Automated complexity measures We report on the automated measures of complexity for each summary version in order to see how the generated summaries differ across complexity levels. Table 6.7 details the complexity measures for each generated version. The Medium and Low complexity versions had large differences in average number of words (529.20, $\sigma = 182.48$ vs. 259.00, $\sigma = 43.42$), average proportion of words outside the top 1,000 Thing Explainer words (0.51, $\sigma = 0.04$ vs. 0.28, $\sigma = 0.05$), average proportion of function words (0.31, $\sigma = 0.03$ vs. 0.36, $\sigma = 0.03$) and language model perplexity (64.14, $\sigma = 24.16$ vs. 23.92, $\sigma = 5.71$). In the case of academic vocabulary (AVL), the Medium summaries were scored as slightly more complex than the High summaries (0.40, $\sigma = 0.03$ vs. 0.39, $\sigma = 0.03$). Compared to the expert written summaries from Section 6.2.1, the generated summaries had more distinct differences in the automated complexity measures. This is not surprising, since the complexity scorer we use to rank generations uses these features to select different complexity versions.

Procedure

The study followed closely the procedure of our first study (Section 6.2). There were the same controls for reading for at least 30 seconds. Participants on average took 100 seconds per article (std=83 seconds). After reading the text, participants answered the same questions about their reading experience and comprehension questions (Section 6.2.1). For the comprehension questions, we use the same questions, even when the generated summaries did not have explicit answers to the questions. When reporting on comprehension, we ignore the 6 (out of 40) questions which did not have answers in a generated summary at a given complexity. 5 out of the 6 ignored questions were in the Low summaries, suggesting that the more simplified text had a higher risk of removing information from the summary.

Analysis

We conduct the same analysis as Section 6.2.1. We fit the same two linear mixed effects models [Lindstrom and Bates, 1990] — LMM_{full} and LMM_{none} — for each reading measure. We conduct likelihood-ratio tests between the models to determine if complexity affects each reading measure. We quantify the differences between the generated complexity versions and their interaction with topic familiarity by conducting post-hoc two-sided t-tests for pairwise comparisons.

6.3.2 Results

The likelihood-ratio tests between our fitted models showed that LMM_{full} had a significantly better fit than LMM_{none} ($p < 0.05$) for all reading experience measures and for the multiple choice comprehension scores. This indicates that complexity had a significant effect on reading experience measures and multiple choice scores. Table C.2 in the Appendix lists all χ^2 and (Holm-Bonferroni corrected) p values. In the following section we report on the results for our second study. Because our questions guiding this analysis are similar to our first study, we focus on the most interesting comparisons to the findings reported in Section 6.2.2.

Medium summaries became more difficult to read and understand

Similar to our first study, participants rated Low summaries as significantly easier to read ($d_{ease} = 0.548$, $p < 0.0001$) and understand ($d_{understand} = 0.330$, $p = 0.001$) than the High summaries. However, we observed two different results in this second study. First, while in our first study, participants found Medium summaries significantly easier to read and understand than High summaries, the two complexities were not significantly different in our second study. Second, while in our first study, participants did not rate the Low and Medium summaries as significantly different, in our second study participants rated Low summaries as significantly easier to read and understand than Medium summaries ($d_{\textit{ease}} = 0.481$, $p < 0.0001$; $d_{\textit{understand}} = 0.286$, $p = 0.003$).

To understand these differences in results between the two studies, we draw attention to the complexity measures on generated summaries (Table 6.7) compared to expert-written summaries (Table 6.2). Notice that the generated Low and Medium summaries had larger differences in complexity measures compared to the expert-written Low and Medium summaries. We postulate the difference in results between the two studies are largely explained by this shift in the Medium summaries to be further in complexity to Low summaries and closer in complexity to High summaries. This is further evidence of complexity's effect on reading experience measures: as the Medium summaries shifted to higher complexity (as measured by our automatic complexity measures) between our first and second study, we see a resulting shift in participant reading experience measures.

Participants with the highest topic familiarity answered more multiple choice questions correctly with the High summaries

Participants in general answered multiple choice questions correctly more often when they read the High summaries (58.24%) compared to Low (53.35%) and Medium summaries (50.14%). Interestingly, this effect was driven solely by participants with the highest topic familiarity. Participants most familiar with a topic answered multiple choice questions correctly significantly more often when reading the High summaries compared to both Low ($p < 0.0001$) and Medium summaries ($p < 0.0001$), as shown in Table 6.8. There were no significant differences across complexity levels for any other topic familiarity.

These findings are in contrast to our findings from the first study, where high topic familiarity participants didn't score significantly differently across complexity levels. One possible explanation is that the different versions of multiple choice questions were aligned to the complexity levels of the summaries in the first study, whereas we did not re-align questions to generated summaries in this second study.[5] This didn't affect the High summaries, which were the same between the two studies, but not re-aligning the questions might have made it harder for readers to find the correct answers on the Low and Medium summaries. That we observe significantly higher scores with the High summaries over other complexity levels only among

[5] This decision is because we did not want the multiple choice questions to be designed around the generations, instead seeing how the generations included (or failed to include) information deemed important in the expert-written summaries.

participants with the highest topic familiarity suggests that topic familiarity may have allowed participants to better take advantage of the aligned language between the questions and summaries.

Generated summaries achieve similar effects of topic familiarity and complexity on reading experience measures

We again show the strength of topic familiarity on reading experience measures and the diminishing benefits of lower complexity as topic familiarity increased. Figure C.2 in the Appendix shows that as familiarity increased, ratings generally increased across metrics and complexity levels, low complexity text provided diminishing benefits. Only participants with the lowest topic familiarity found the summaries at the lowest complexity more interesting and valuable, as shown in Table 6.8. For reading ease and understanding, the same trend holds, though participants with slightly more familiarity also retained some benefit in reading ease and understanding from Low summaries compared to High summaries.

6.4 Summary

In this chapter we explored whether readers benefit from a scientific text complexity sensitive to their background knowledge and if creating alternative text could be partially automated to support broad online uptake. In our first study ($N = 200$) using expert-written summaries, we found that language complexity mattered for improving reading experience and comprehension. In addition, we showed that the ideal complexity depends on a reader's background knowledge: readers with the lowest topic familiarity responded most positively to the lowest complexity summaries, rating them as easier to read, understand, more interesting, and valuable. As a reader's familiarity with a topic increased, the benefits of simplified text disappeared.

Simplifying language necessarily abstracts some details, but general audience readers often express an interest in reading research papers in their original form, even when such papers are complex and difficult to understand (Chapter 5). It is in the original research paper a reader can find all detailed information on a study, provided they can parse the complex language. Our findings show that some readers with high familiarity in a topic, even if they are not researchers in the field, can parse original paper content. At the same time, readers with less familiarity in a topic will generally benefit from plain language summaries more than from the original scientific content. This suggests that there is no one-size-fits-all complexity for

	Familiarity	d^{Lo-Me}	p	d^{Lo-Hi}	p	d^{Me-Hi}	p
Reading Ease	1	1.385	**<0.0001**	1.642	**<0.0001**	0.257	0.125
	2	0.355	0.281	0.703	**0.013**	0.348	0.281
	3	0.389	0.106	0.346	0.120	-0.044	0.801
	4	0.057	1.000	-0.046	1.000	-0.104	1.000
	5	0.215	1.000	0.094	1.000	-0.122	1.000
	All	0.481	**<0.0001**	0.548	**<0.0001**	0.067	0.455
Understanding	1	0.835	**<0.0001**	1.101	**<0.0001**	0.266	0.112
	2	0.405	0.192	0.665	**0.021**	0.260	0.270
	3	0.034	0.856	0.229	0.633	0.196	0.633
	4	0.030	1.000	-0.078	1.000	-0.108	1.000
	5	0.127	0.703	-0.266	0.703	-0.393	0.516
	All	0.286	**0.003**	0.330	**0.001**	0.044	0.622
Interest	1	0.589	**0.001**	0.907	**<0.0001**	0.317	0.056
	2	0.159	1.000	0.148	1.000	-0.011	1.000
	3	-0.048	1.000	-0.046	1.000	0.002	1.000
	4	-0.004	1.000	-0.011	1.000	-0.007	1.000
	5	0.252	1.000	0.053	1.000	-0.199	1.000
	All	0.190	**0.062**	0.21	0.061	0.02	0.818
Value	1	0.068	0.675	0.407	**0.031**	0.338	0.078
	2	-0.211	0.969	0.018	0.969	0.229	0.969
	3	-0.173	1.000	-0.157	1.000	0.016	1.000
	4	0.151	0.664	-0.086	0.664	-0.237	0.424
	5	0.271	0.615	-0.101	0.720	-0.371	0.568
	All	0.021	1.0	0.016	1.0	-0.005	1.0
Comprehension	1	0.054	0.499	0.130	0.153	0.076	0.499
	2	0.033	0.793	-0.090	0.793	-0.123	0.563
	3	0.024	0.760	-0.117	0.253	-0.141	0.111
	4	0.025	1.000	0.044	1.000	0.019	1.000
	5	0.136	0.223	-0.428	**<0.0001**	-0.564	**<0.0001**
	All	0.049	0.330	-0.026	0.461	-0.075	0.062

Table 6.8: Study 2: Post-hoc (two-sided) tests for pairwise differences in fixed-effects estimates between generated complexity versions and across participant topic familiarities. For example, in the column for d^{Lo-Me} and row for "Interest" and "1" topic familiarity we can interpret the result as participants who rated their topic familiarity for a summary as a 1 rated the generated Low complexity, on average, 0.589 points higher for interest (out of 5) compared to the Medium complexity when controlling for participant and paper.

summaries. Adjusting complexity to reader topic familiarity can improve communication of and engagement with scientific information.

Altogether, the results of our first study support our assumption that a single plain language summary for all potential readers is insufficient. Providing versions of different complexity, however, likely requires some level of automation to broadly support diverse readers in various contexts. We therefore ran a follow up study ($N = 194$) using generated summaries curated for factuality to see if they were able to support different readers based on their topic familiarity, similar to expert-written summaries. Participants with the least topic familiarity again rated the least complex generated text as significantly easier to read, understand, and as more interesting compared to the most complex text, while participants with the highest topic familiarity again gained no significant benefit from the least complex text. From these findings, we conclude that it is feasible to automatically generate scientific summaries at different complexities as a way of making such text flexible to different audiences.

Chapter 7

Discussion

We have unprecedented access to information online. Someone with an internet connection can learn about cutting edge research in any subject, read more books, blog posts or news articles than could fit in any physical library, and share information with people around the globe instantly. The vast majority of this information is encoded in language. While information online is quickly spreading to new audiences, language is lagging behind. Text containing this information was written with a specific audience in mind (e.g., journal papers assume their readers are other researchers), but as audiences grow, more people fall outside this text's intended audience.

This thesis shows how a mismatch in language and audience can impact on who can truly access information. This recognition is well-studied in sociolinguistics (e.g., Bell [1984]'s theory of audience design), but rarely is it applied to interactive systems. This thesis is the first to use experimental techniques to draw out empirically driven design recommendations for language design.

We show how scientific language can discourage engagement on *r/science* for millions of people and can pose a near-impenetrable barrier to understanding for patients and caregivers who need the information to make informed decisions about their health. People who are historically underrepresented in science are often the ones who are most likely to be discouraged by scientific language. This bias in how science is communicated online can further alienate stakeholders already left out of the scientific enterprise. Our findings that automated tools for changing language complexity can alleviate these communication barriers point to the promise of intelligent systems that improve communication between broader and more diverse audiences.

7.1 Implications of Language as Design

Below we outline the implications of this work for community builders, science communicators and writers, designers, and intelligent system builders.

Community builders Our large-scale analysis of posts and comments on *r/science* contributes to research on community norms and can provide insights for community designers for socializing new members. A key issue for communities is welcoming newcomers. Newcomers represent growth and interest in a community, but can also risk overwhelming or alienating current members by (perhaps unknowingly) violating community norms. Past work has explored ways of socializing newcomers by pairing them with experienced members [Ford et al., 2018] or providing just-in-time guidelines on new posts [Matias, 2019]. Seering et al. [2017] also found that positive example setting, especially by high profile members, was effective at encouraging new members to adopt norms.

Our work provides an additional avenue of socializing new members through language norms. Matching the language of a community is an important element for receiving positive responses [Sharma and De Choudhury, 2018]. Our results from the analysis of *r/science* (Chapter 3), identify words and phrases common on *r/science* that also represent many norms of the community. For example, the lack of personal pronouns is paralleled by the norm of not providing personal anecdotes. Community designers, such as moderators, could make these language norms explicit as a way of welcoming new users. Leveraging insights gained from an analysis of community language, tools could help new members rewrite their posts or comments to better align with community language. Our method for identifying writing strategies in science communication (Chapter 4) also suggest that these tools could do more than just identify common words or phrases; they could highlight discourse-level strategies (e.g., telling stories) that are well-received in the community, providing newcomers positive examples of norms to follow.

Science communicators and writers Our computational approaches to science communication strategies can also assist in training new science communicators or help them write to a new audience. Many expert science communicators already do, or are encouraged to, keep their audience in mind when writing (Chapter 4), but guidelines are usually general advice with little empirical evidence supporting their claims. Rather than only having general guidelines, science communicators can use our classifiers from Chapter 4 to see how their language leverages specific strategies. Inspired by prior systems that support writers in understanding the rhetorical features of their language [Ishizaki and Kaufer, 2012], tools could assist writers by highlighting the strategies they use and provide suggestions based on their intended audience. For example, when writing for journalists, science communicators could check how their writing emphasizes the impact of the findings.

Interface designers Interface designers can leverage our insights to design language more appropriate for a given user group. In this thesis we identified what features of language are important to change by using computational methods to identify strategies and language patterns in real world text (Chapters 3 and 4) and by running user studies (Chapters 5 and 6). While the findings reported in this dissertation focus on science communication, the methodologies employed for identifying language changes and their effect on user behavior can inform, for example, AB testing for interfaces with different language style variants.

Intelligent system builders As AI technology advances, new interfaces integrating this technology can provide tremendous value to users. Developers of these intelligent interfaces can take this thesis as an illustration of one path towards tools integrating AI to augment communication. Our findings show that generative modelling (e.g., GPT-3) is reaching the point where generated text can be integrated into end-user tools. Most prior work integrating generative text models into tools have focused on writing support (e.g., [Gero et al., 2021; Chung et al., 2022]), and often in a secondary role. For example, Gero et al. [2021] showed how generative models could help inspire writers when communicating about scientific topics. In their use case, generative models provided ‘sparks’, single sentences meant to inspire longer explanations. The success of our automated methods for reading in PAPER PLAIN (Chapter 5) and in generating scientific summaries in Chapter 6 suggest that AI can also take a more active role in the writing and reading process. The major cavaet to these findings is that model hallucinations are still a major risk for readers, though our preliminary findings in Chapter 6 suggest that such hallucinations are easily identified by experts. We hope that our techniques to make machine output useful for readers can provide useful insight for future tools integrating machine intelligence.

7.2 Limitations & Risks

One major hurdle in deploying systems using generative models is the risk of hallucinations. This thesis argues that such hallucinations necessitate human-in-the-loop systems. Rather than expert curation being a limitation, though, we envision such systems improving human-human communication. Science communication is ideally a conversation, not a transmission of information [Nisbet and Scheufele, 2009b]. A fully automated system would allow researchers to publish plain language summaries without thinking about their potential audiences. Our hope is that requiring experts to read generations, while potentially limiting the speed at which such summaries are published, will encourage researchers to think more deeply about the audiences they are reaching with their work.

While the goal of this thesis to make scientific information accessible to a wider audience, it could also further exacerbate disparities between stakeholders. This thesis exclusively focused on the English language, but this could disproportionately support native English speakers in accessing and understanding scientific information. Current science communication efforts suffer from similar issues [Stoll et al., 2022; Meneghini and Packer, 2007], and the findings in this thesis take a first step towards personalizing communication in this context. In future steps, we are excited about the possibility of exploring how different languages and dialects can further adjust language to the individual.

7.3 Future Work

This thesis opens up exciting avenues of future work on augmenting human communication online.

7.3.1 Constructing Generalized Reader Models

In Chapter 6 we provide the first step in making language flexible to a reader by showing how changing language complexity can impact reader experience. There is a wealth of prior work on how to design adaptive user interfaces [Gajos et al., 2008a; Schneider-Hufschmidt et al., 1993], but rarely is language considered a feature for personalization. Future work could explore what the best ways to personalize or adapt language are. For example, given a reader profile or short initial quiz, a system could recommend a complexity level for a reader [Wallace et al., 2022; Murthy et al., 2021]. Incorporating reading history into a user model could also help systems adjust language across the web for a reader. Alternatively, a reader could have full control over the complexity levels, using a knob or dial to scan through possible versions to select their preferred version. This work would have exciting implications for interface personalizing by showing how readers respond to language adjusting in real time to suit their needs.

7.3.2 Adapting Beyond Science Communication

Scientific text is not the only context where non-expert readers wrestle with highly technical documents; this thesis can inspire efforts in addressing similar barriers in these other contexts. Some aspects of these contexts merit new design efforts, while others might benefit from similar systems as PAPER PLAIN. For example, the important questions to ask while reading a medical paper are different than those for a legal contract or privacy statement. This necessitates a re-crafting of PAPER PLAIN's Key Question Index and Answer Gists for other domains. Furthermore, some documents may need to be read in a particular order (e.g., a software tutorial), and providing an alternative index, as PAPER PLAIN does, could confuse readers. In these cases, any key question indexes into the document would need to be aligned with the document's original structure. Providing in-situ sections summaries and definitions could help address needs in these domains, as these features can help readers understand what they are reading within the flow of a document. In some contexts, language complexity might also not be the right dimension to vary. For example, our prior work has shown that in online experiments [August and Reinecke, 2019] or security statements [Stokes et al., 2023], formality might be an important language dimension that can impact engagement and compliance. Future work could identify other communication contexts and associated dimensions of language to adapt.

7.3.3 Adapting Beyond the Reader

People read text for different reasons [Britt et al., 2014b], which might impact their preferred language just as much as their background. For example, someone reading a new biomedicine paper out of interest might have a different tolerance or desire for complexity compared to a patient who is making a medical decision in part based on the information they find. In the first case, language complexity might be seen as a nuisance, and the majority of readers only want a high level takeaway. In the second case, complexity might be required to have fidelity to the original paper. Building systems that are cognizant of these different contexts is equally important to being flexible to the readers themselves. Future work could evaluate how reader goals, rather than background knowledge (Chapter 6), influence their response to dimensions of language like complexity.

7.3.4 Supporting Writers

Communication is two-way, so while automatically adjusting language to the individual augments part of communication, it is equally important to help people communicate better with their intended audience. We envision tools that can “crash-test” writing by modeling how a potential reader will respond to it—-anything from a tweet to a research paper. Future work could build models representing a reader with a particular background and make predictions about that reader’s comprehension and general response to presented text. Our current tools for adapting language automatically collect responses to language choices from real readers, which can inform the development and continued improvement of such reader models.

7.3.5 Supporting Collective Sense-Making

The success of online forums like *r/science* shows that communicating scientific information to an individual does not happen in a vacuum. People share and discuss papers together. Many of the methods introduced in this thesis focus on adapting language to the individual. An additional direction to support communication is enabling people to collectively make sense of complex language. For example, tools could support social annotations on research papers that allow people to define words or explain concepts for other people. Alternatively, future tools could could make it easier for people to see how language was shared, and potentially changed, through a network, which could help people follow information flowing through social media.

7.4 Conclusion

As an illustration of how the methods described in this thesis can be used in adjusting scientific language, the following conclusions are generated and lightly edited using the techniques described in Chapter 6:

Low Complexity This thesis is about how changes in language can change people's behavior and how automated systems can be used to change language for different people. For example, in the subreddit *r/science* language changes are associated with changes in people's behavior. To understand what language is important to change, the author investigated how science writers write for a general audience. The author then takes inspiration from this to build a reading tool for making medical research papers approachable to more people.

Medium Complexity This thesis looks at how language choices can influence user behavior, and how automated systems can be used to design language specifically for different people. The thesis begins by focusing on the subreddit *r/science* and how language choices there can restrict access to scientific information. The author then looks at how experts design scientific language for a general audience, and takes inspiration from these strategies to build a reading interface that makes medical research papers approachable to a general audience. Finally, the thesis looks at how people respond to levels of language complexity based on their background knowledge, and how this can be used to generate summaries with varying complexities.

High Complexity In this thesis, the author shows that language choices, similar to other interface design choices, influence user behavior and introduces automated systems that design language for different people. The author begins by focusing the study of language choices to the subreddit *r/science* and shows how language choices in it are associated with changes in people's behavior, potentially restricting access to scientific information. To understand what language is important to change when adapting to different people, the author investigates how experts design scientific language for a general audience. The author takes inspiration from these expert strategies to build a reading interface for making medical research papers approachable to a general audience. To adjust language to finer-grained audiences, the author investigates how people respond to levels of language complexity based on their background knowledge and develop a novel controllable generation method to adjust the complexity of generated summaries. In two user studies the author observed that generated summaries lead to similar reader responses as with expert summaries, establishing the feasibility of generating summaries with varying complexities. The work provides guidance on designing language for specific audiences and adaptable communication at scale. The author concludes with a summary of the contributions and a discussion of future research on designing language to encourage better communication online.

Bibliography

American Association for the Advancement of Science AAAS. Communication fundamentals.

Takeshi Abekawa and Akiko Aizawa. 2016. Sidenoter: Scholarly paper browsing system based on pdf restructuring and text annotation. In *COLING*.

@access Working Group. 2021. Who needs access? you need access!

Eytan Adar, Carolyn Gearig, A. Balasubramanian, and Jessica R. Hullman. 2017. Persalog: Personalization of news article content. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*.

Tal August, Dallas Card, Gary Hsieh, Noah A Smith, and Katharina Reinecke. 2020a. Explain like i am a scientist: The linguistic barriers of entry to r/science. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM.

Tal August, Lauren Kim, Katharina Reinecke, and Noah A Smith. 2020b. Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 5327–5344.

Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2023. Know your audience: The benefit and feasibility of reader-sensitive scientific language complexity. In *In Submission*.

Tal August, Nigini Oliveira, Chenhao Tan, Noah Smith, and Katharina Reinecke. 2018. Framing effects: Choice of slogans used to advertise online experiments can boost recruitment and lead to sample biases. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19.

Tal August and Katharina Reinecke. 2019. Pay attention, please: Formal language improves attention in volunteer and paid online experiments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11. ACM.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022a. Generating scientific definitions with controllable complexity. In *ACL*.

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2022b. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ArXiv*, abs/2203.00130.

Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. 2019. Elastic documents: Coupling text and tables through contextual visualizations for enhanced document reading. *IEEE Transactions on Visualization and Computer Graphics*, 25:661–671.

Xiaoliang Bai, Yong sheng Lian, Jie Wang, Hongxing Zhang, Meichao Jiang, Hao Zhang, Bo Pei, Changqing Hu, and Qiang Yang. 2021. Percutaneous endoscopic lumbar discectomy compared with other surgeries for lumbar disc herniation. *Medicine*, 100.

Alison S. Baskin, Ton Wang, Nicole M Mott, Sarah T. Hawley, Reshma Jagsi, and Lesly Dossett. 2020. Gaps in online breast cancer treatment information for older women. *Annals of Surgical Oncology*, 28:950–957.

Douglas M. Bates, Benjamin M. Bolker, and Steven C. Walker. 2014. fitting linear mixed effects models using lme 4 arxiv.

Amanda Baughan, Tal August, Naomi Yamashita, and Katharina Reinecke. 2020. Keep it simple: How visual complexity and preferences impact search efficiency on websites. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Allan Bell. 1984. Language style as audience design. *Language in Society*, 13:145 – 204.

Antonio Bella, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana. 2010. Quantification via probability estimators. In *2010 IEEE International Conference on Data Mining*, pages 737–742. IEEE.

Emily M Bender and Alex Lascarides. 2019. Linguistic fundamentals for natural language processing ii: 100 essentials from semantics and pragmatics. *Synthesis Lectures on Human Language Technologies*.

Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative hci research: Going behind the scenes. In *Synthesis Lectures on Human-Centered Informatics*.

Angela Collier Bliss. 2019. Adult Science-Based Learning: The Intersection of Digital, Science, and Information Literacies. *Adult Learning*, 30(3):128–137.

Deborah Blum, Mary Knudson, Robin Marantz Henig, et al. 2006. *A field guide for science writers*. Oxford University Press, USA.

Tanner A. Bohn and Charles X. Ling. 2021. Hone as you read: A practical type of interactive summarization. *ArXiv*, abs/2105.02923.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258.

Luke Bratton, Rachel C Adams, Aimée Challenger, Jacky Boivin, Lewis Bott, Christopher D Chambers, and Petroc Sumner. 2019. The association between exaggeration in health-related science news and academic press releases: a replication study. *British Medical Journal*, 4.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3:101 – 77.

Belinda Bray, Bev France, and John K Gilbert. 2012. Identifying the essential elements of effective science communication: What do the experts say? *International Journal of Science Education, Part B*, 2(1):23–41.

M. Anne Britt, Tobias Richter, and Jean François Rouet. 2014a. Scientific Literacy: The Role of Goal-Directed Reading and Evaluation in Understanding Scientific Information. *Educational Psychologist*, 49(2):104–122.

Mary Anne Britt, Tobias Richter, and Jean-François Rouet. 2014b. Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist*, 49:104 – 122.

Rainer Bromme and Susan R. Goldman. 2014. The public's bounded understanding of science. *Educational Psychologist*, 49:59 – 69.

Alex Broom. 2005. Virtually he@lthy: The impact of internet use on disease experience and the doctor-patient relationship. *Qualitative Health Research*, 15:325 – 345.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Peter Brusilovsky and Leonid Pesin. 1998. Adaptive navigation support in educational hypermedia: An evaluation of the isis-tutor. In *CIT 2015*.

Anita Bruzzese. 2018. How to explain technical information to non-techies.

Tania Bubela, Matthew C Nisbet, Rick Borchelt, Fern Brunger, Cristine Critchley, Edna Einsiedel, Gail Geller, Anil Gupta, Jürgen Hampel, Robyn Hyde-Lay, et al. 2009. Science communication reconsidered. *Nature Biotechnology*, 27(6):514–518.

Katie Burke. 2018. 12 tips for scientists writing for the general public.

Moira Burke and Robert Kraut. 2008. Mind your ps and qs: the impact of politeness and rudeness in online communities. In *Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work*, pages 281–284. ACM.

Terry W. Burns, D. John O'Connor, and Susan M. Stocklmayer. 2003. Science communication: a contemporary definition. *Public Understanding of Science*, 12(2):183–202.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. *ArXiv*, abs/2004.15011.

Mengyao Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *ACL*.

Dallas Card and Noah A Smith. 2018. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1636–1646.

Marc-Allen Cartright, Ryen W. White, and Eric Horvitz. 2011. Intentions and attention in exploratory health search. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*.

Justine Cassell and Dona Tversky. 2005. The language of online intercultural community formation. *Journal of Computer-Mediated Communication*, 10(2).

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. In *Proceedings of the ACM on Human-Computer Interaction*, 32.

Vinay K. Chaudhri, Britte Haugan Cheng, Adam Overholtzer, Jeremy Roschelle, Aaron Spaulding, Peter E. Clark, Mark T. Greaves, and David Gunning. 2013. Inquire biology: A textbook that answers questions. *AI Mag.*, 34:55–72.

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Dennis Chong and James N Druckman. 2007. Framing Theory. *Annual Review Political Science*, 10:103–26.

Munmun De Choudhury, Meredith Ringel Morris, and Ryen W. White. 2014. Seeking and sharing health information online: comparing search engines and social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Rune Christensen. 2018. Cumulative link models for ordinal regression with the r package ordinal.

John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. Talebrush: Sketching stories with generative pretrained language models. *CHI Conference on Human Factors in Computing Systems*.

Robert B. Cialdini, Raymond R. Reno, and Carl A. Kallgren. 1990. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6).

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *ACL*.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset.

Avital Cnaan, Nan M. Laird, and Peter Slasor. 1997. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, 16:2349–2380.

Anthony M Cocco, Rachel D. Zordan, David Taylor, Tracey J Weiland, Stuart J Dilley, Joyce A Kant, Mahesha Hk Dombagolla, Andreas Hendarto, Fiona Wy Lai, and Jennie Hutton. 2018. Dr google in the ed: searching for online health information by adult emergency department patients. *Medical Journal of Australia*, 209.

Cochrane. 2021. New standards for plain language summaries.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S. Weld. 2019. Pretrained language models for sequential sentence classification. *ArXiv*, abs/1909.04054.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, pages 97–135.

Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David A. Sontag. 2011. Personalizing web search results by reading level. In *CIKM '11*.

Nikolas Coupland. 2002. Style and sociolinguistic variation: Language, situation, and the relational self: theorizing dialect-style in sociolinguistics.

Curtis Crawford. *Writing for a General Audience: Science Journalism*, page 251–258.

Robert Cudeck. 1996. Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate behavioral research*, 31 3:371–403.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 5636–5646.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon M. Kleinberg. 2012. Echoes of power: language effects and power differences in social interaction. *Proceedings of the 21st international conference on World Wide Web*.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 307–318.

Rumen Dangovski, Michelle Shen, Dawson Byrd, Li Jing, Desislava Tsvetkova, Preslav Nakov, and Marin Soljacic. 2020. We can explain your research in layman's terms: Towards automating science journalism at scale.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Suzanne Day, Stuart Rennie, Danyang Luo, and Joseph D. Tucker. 2020. Open to the public: paywalls and the public rationale for open access medical research publishing. *Research Involvement and Engagement*, 6.

Ashwin Devaraj, William Sheffield, Byron C. Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Kristina Dzara and Ariel S Frey-Vogel. 2019. Medical education journal club for the millennial resident: An interactive, no-prep approach. *Academic pediatrics*.

Penelope Eckert. 2004. The meaning of style.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Saul Epstein. 1997. Impure science: Aids, activism, and the politics of knowledge. *Nature Medicine*, 3:242–243.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *ArXiv*, abs/2007.12626.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Fermat's Library. 2021. Fermat's library.

Victor S. Ferreira. 2019. A mechanistic framework for explaining audience design in language production. *Annual review of psychology*, 70:29–51.

Casey Fiesler, Jialun “Aaron” Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of International AAAI Conference on Web and Social Media (ICWSM)*.

Joseph L Fleiss. 1994. Measures of effect size for categorical data.

Denae Ford, Kristina Lustig, Jeremy Banks, and Chris Parnin. 2018. We don’t do that here: How collaborative editing with mentors improves engagement in social q&a communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Elena Forzani. 2016. Individual differences in evaluating the credibility of online information in science: Contributions of prior knowledge, gender, socioeconomic status, and offline reading ability.

Riccardo Fusaroli, Bahador Bahrami, Karsten Olsen, Andreas Roepstorff, Geraint Rees, Chris Frith, and Kristian Tylén. 2012. Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological science*, 23(8):931–939.

Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021a. Discourse understanding and factual consistency in abstractive summarization. In *Proceedings of the 16th Annual Meeting of the European Chapter of the Association for Computational Linguistics*.

Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Asli Çelikyilmaz, and Yejin Choi. 2021b. Discourse understanding and factual consistency in abstractive summarization. In *EACL*.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2020. Go figure! a meta evaluation of factuality in summarization. *ArXiv*, abs/2010.12834.

Kristin Gagnier and Kelly Fisher. 2017. Communicating science to non-scientific audiences.

Krzysztof Z Gajos, Katherine Everitt, Desney S. Tan, Mary Czerwinski, and Daniel S. Weld. 2008a. Predictability and accuracy in adaptive user interfaces. In *CHI*.

Krzysztof Z Gajos, Jacob O. Wobbrock, and Daniel S. Weld. 2008b. Improving the performance of motor-impaired users with automatically-generated, ability-based interfaces. In *CHI*.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *ACL*.

Dee Gardner and Mark Davies. 2013. A new academic vocabulary list. *Applied Linguistics*, 35(3):305–327.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *ArXiv*, abs/2009.11462.

Katy I. Gero, Vivian Liu, and Lydia B. Chilton. 2021. Sparks: Inspiration for science writing using language models. *ArXiv*, abs/2110.07640.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence.

Maya J. Goldenberg. 2016. Public misunderstanding of science? reframing the problem of vaccine hesitancy. *Perspectives on Science*, 24:552–581.

Amy L. Gonzales, Jeffrey T. Hancock, and James W. Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37:19 – 3.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. *ArXiv*, abs/2104.04302.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Yue Guo, Weijian Qiu, Yizhong Wang, and Trevor A. Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. *ArXiv*, abs/2012.12573.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of the BioNLP Workshop*, page 99–107. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 2020 Annual Meeting of the Association for Computational Linguistics*.

Donald P. Haider-Markel and Mark R Joslyn. 2001. Gun policy, opinion, tragedy, and blame attribution: The conditional influence of issue frames.

Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5):664–688.

Aaron Halfaker, Aniket Kittur, Robert Kraut, and John Riedl. 2009. A jury of your peers: quality, experience and ownership in wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. ACM.

Thomas Hayden, Michelle Nijhuis, et al. 2013. *The Science Writers' Handbook: Everything You Need to Know to Pitch, Publish, and Prosper in the Digital Age*. Da Capo Lifelong Books.

Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Marti A. Hearst, Emily Pedersen, Lekha Priya Patil, Elsie Lee, Paul Laskowski, and Steven L. Franconeri. 2020. An evaluation of semantically grouped word cloud designs. *IEEE Transactions on Visualization and Computer Graphics*, 26:2748–2761.

Per Hetland. 2014. Models in science communication: formatting public engagement and expertise. *Nordic Journal of Science and Technology Studies*, 2(2):5–17.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Judith A Holton. 2007. The coding process and its challenges. *The Sage handbook of grounded theory*, (Part III):265–89.

Daniel J Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

Hypothes.is. 2021. Hypothes.is: Annotate the web, with anyone, anywhere.

Suguru Ishizaki and David Kaufer. 2012. Computer-aided rhetorical analysis.

Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 2026–2031.

Abhinav Jain, Nitin Gupta, Shashank Mujumdar, Sameep Mehta, and Rishi Madhok. 2018. Content driven enrichment of formal text using concept definitions and applications. *Proceedings of the 29th on Hypertext and Social Media*.

Victoria Johansson, Anna Sigridur Islind, Tomas Lindroth, Eva Angenete, and Martin Gellerstedt. 2021. Online communities as a driver for patient empowerment: Systematic review. *Journal of Medical Internet Research*, 23.

Ridley Jones, Lucas Colusso, Katharina Reinecke, and Gary Hsieh. 2019. r/science: Challenges and opportunities for online science communication. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Meghana Kalavar, Sasha Hubschman, Julia L Hudson, Ajay E. Kuriyan, and Jayanth Sridhar. 2021. Evaluation of available online information regarding treatment for vitreous floaters. *Seminars in Ophthalmology*, 36:58 – 63.

Anna Kernder, Elena Elefante, Gamal Chehab, Chiara Tani, Marta Mosca, and Matthias Schneider. 2020. The patient's perspective: are quality of life and disease burden a possible treatment target in systemic lupus erythematosus? *Rheumatology (Oxford, England)*, 59:v63 – v68.

Martin Kerwer, Anita Chasiotis, Johannes Stricker, Armin Günther, and Tom Rosman. 2021. Straight from the scientist's mouth—plain language summaries promote laypeople's comprehension and knowledge acquisition when reading about individual research findings in psychology. *Collabra: Psychology*.

Joëlle Kivits. 2006. Informed patients and the internet. *Journal of Health Psychology*, 11:269 – 282.

Robert E. Kraut and Paul Resnick. 2012. *Building Successful Online Communities: Evidence-based Social Design*. MIT Press.

Mark Kröll, Gunnar Schulze, and Roman Kern. 2014. A study of scientific writing: Comparing theoretical guidelines with practical implementation. In *COLING Workshop on Synchronic and Diachronic Approaches to Analyzing Technical Language*, pages 48–55.

Christoph Kueffer and Brendon MH Larson. 2014. Responsible use of language in scientific writing and science communication. *BioScience*, 64(8):719–724.

Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Christensen. 2017. lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82:1–26.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *ArXiv*, abs/2111.09525.

William Labov. 1973. The linguistic consequences of being a lame. *Language in Society*, 2(1):81–115.

William Labov. 2006. *The Social Stratification of English in New York City*. Cambridge University Press.

Talia Lavie and Noam Tractinsky. 2004. Assessing dimensions of perceived visual aesthetics of web sites. *Int. J. Hum. Comput. Stud.*, 60:269–298.

Falk Leichsenring, Allan A. Abbass, Ellen Driessen, Mark Hilsenroth, Patrick Luyten, Sven Rabung, and Christiane Steinert. 2018. Equivalence and non-inferiority testing in psychotherapy research. *Psychological Medicine*, 48:1917 – 1919.

Gondy Leroy, Stephen Helmreich, and James R. Cowie. 2010. The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics*, pages 438–449.

Gondy Leroy, Stephen Helmreich, James R. Cowie, Trudi Miller, and Wei Zheng. 2008. Evaluating online health information: Beyond readability formulas. In *AMIA Annual Symposium Proceedings*, page 394. American Medical Informatics Association.

Julie Letchford, Hazel R. Corradi, and Trevor Day. 2017. A flexible e-learning resource promoting the critical reading of scientific papers for science undergraduates. *Biochemistry and Molecular Biology Education*, 45.

Daniel Li, Thomas Chen, Albert Tung, and Lydia B Chilton. 2021. Hierarchical summarization for longform spoken dialog. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 582–597.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Gitte Lindgaard, Cathy Dudek, Devjani Sen, Livia Sumegi, and Patrick S. Noonan. 2011. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Trans. Comput. Hum. Interact.*, 18:1:1–1:30.

Magnus Lindstrom and Douglas M. Bates. 1990. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46 3:673–87.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ArXiv*, abs/2107.13586.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *ECDL*.

Annie Louis and Ani Nenkova. 2013a. A corpus of science journalism for analyzing writing quality. *Dialogue & Discourse*, 4(2):87–117.

Annie Louis and Ani Nenkova. 2013b. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Karina Lovell, Deborah Cox, Gillian Haddock, Christopher Jones, David Raines, Rachel Garvey, Chris Roberts, and Sarah Hadley. 2006. Telephone administered cognitive behaviour therapy for treatment of obsessive compulsive disorder: randomised controlled non-inferiority trial. *BMJ : British Medical Journal*, 333:883.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Neurologic decoding:(un) supervised neural text generation with predicate logic constraints. *ArXiv*, abs/2010.12884.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Neurologic decoding: (un)supervised neural text generation with predicate logic constraints. In *NAACL*.

María José Luzón. 2013. Public communication of science in blogs: Recontextualizing scientific discourse for a diversified audience. *Written Communication*, 30(4):428–457.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 2014 Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Chrysanne Di Marco, Peter Bray, H. Dominic Covvey, Donald D. Cowan, Vic Di Ciccio, Eduard H. Hovy, Joan Lipa, and C. Yang. 2006. Authoring and generation of individualized patient education materials. *American Medical Informatics Association Annual Symposium proceedings*, pages 195–9.

Iain James Marshall, Joël Kuiper, and Byron C. Wallace. 2016. Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association : JAMIA*, 23:193 – 201.

Trevor Martin. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 27–31.

Alejandro Martínez and Stefano Mammola. 2021. Specialized terminology reduces the number of citations of scientific papers. *Proceedings of the Royal Society B*, 288.

J. Nathan Matias and Merry Mou. 2018. Civilservant: Community-led experiments in platform governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.

Jorge Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116:9785 – 9789.

Andrew Maynard and Dietram A. Scheufele. 2020. What does research say about how to effectively communicate about science?

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. *ArXiv*, abs/2005.00661.

Lisa McCorkell, Gina S. Assaf, Hannah E. Davis, Hannah Wei, and Athena Akrami. 2021. Patient-led research collaborative: embedding patients in the long covid narrative. *Pain Reports*, 6.

Rogerio Meneghini and Abel Laerte Packer. 2007. Is there science beyond english? *EMBO reports*, 8.

James Milroy and Lesley Milroy. 1978. Belfast: Change and variation in an urban vernacular. *Sociolinguistic Patterns in British English*, 19:19–36.

Aliaksei Miniukovich, Antonella De Angeli, Simone Sulpizio, and Paola Venuti. 2017. Design guidelines for web readability. *Proceedings of the 2017 Conference on Designing Interactive Systems*.

Jonathan T. Morgan and Anna Filippova. 2018. 'welcome' changes?: Descriptive and injunctive norms in a wikipedia sub-community. 52. ACM.

Morten Moshagen and Meinald T. Thielsch. 2010. Facets of visual aesthetics. *Int. J. Hum. Comput. Stud.*, 68:689–709.

Mati Mõttus and David Jose Ribeiro Lamas. 2015. Aesthetics of interaction design: A literature review. In *MIDI '15*.

H. Münchow, T. Richter, and S. Schmid. 2020. *What Does It Take to Deal with Academic Literature?*, pages 241–260. Springer Fachmedien Wiesbaden, Wiesbaden.

Randall Munroe. 2017. *Thing explainer complicated stuff in simple words*. John Murray.

Sonia Krishna Murthy, Daniel King, Tom Hope, Daniel Weld, and Doug Downey. 2021. Towards personalized descriptions of scientific concepts. In *Proceedings of The Fifth Widening Natural Language Processing Workshop*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vinay Nair, Shahab Khan, and Kenar D. Jhaveri. 2012. Interactive journals and the future of medical publications. *The American journal of medicine*, 125 10:1038–42.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cícero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. *ArXiv*, abs/2102.09130.

National Institutes of Health. 2005. Policy on enhancing public access to archived publications resulting from nih-funded research.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Dong Nguyen and Carolyn P. Rosé. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media*, pages 76–85.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural*

Language Processing and International Joint Conference on Natural Language Processing, pages 1587–1598. Association for Computational Linguistics.

Kristian Hvidtfelt Nielsen and Rikke Schmidt Kjærgaard. 2011. News coverage of climate change in nature news and science now during 2007. *Environmental Communication*, 5(1):25–44.

Matthew C. Nisbet and Dietram A. Scheufele. 2009a. What's next for science communication? promising directions and lingering distractions. *American journal of botany*, 96 10:1767–78.

Matthew C Nisbet and Dietram A Scheufele. 2009b. What's next for science communication? promising directions and lingering distractions. *American journal of botany*, 96(10):1767–1778.

Emily Nunn and Stephen Pinfield. 2014. Lay summaries of open access journal articles: engaging with the general public on medical research. *Learned Publishing*, 27.

Nigini Oliveira, Michael Muller, Nazareno Andrade, and Katharina Reinecke. 2018. The exchange in stack-exchange: Divergences between stack overflow and its culturally diverse participants. *Proceedings of the ACM on Human-Computer Interaction*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

George Philipp and Ryen W. White. 2014. Interactions between health searchers and search engines. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Pontus Plavén-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. The readability of scientific texts is decreasing over time. *eLife*, 6.

Horst Pöttker. 2003. News and its communicative quality: The inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511.

Eufemia S Putortì, Simona Sciara, Norman U Larocca, Massimo P Crippa, and Giuseppe Pantaleo. 2020. Communicating science effectively: When an optimised video communication enhances comprehension, pleasantness, and people's interest in knowing more about scientific findings. *Applied Psychology*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Tzipora Rakedzon, Elad Segev, Noam Chapnik, Roy Yosef, and Ayelet Baram-Tsabari. 2017. Automatic jargon identifier for scientists engaging with the public and science communication educators. *PloS one*, 12(8).

Mathieu Ranger and Karen Bultitude. 2016. 'the kind of mildly curious sort of science interested person like me': Science bloggers' practices relating to audience recruitment. *Public Understanding of Science*, 25(3):361–378.

Katharina Reinecke and Abraham Bernstein. 2013. Knowing what a user likes: A design science approach to interfaces that automatically adapt to culture. *MIS Q.*, 37:427–453.

Katharina Reinecke and Krzysztof Z Gajos. 2014. Quantifying visual preferences around the world. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Katharina Reinecke, Patrick Minder, and Abraham Bernstein. 2011. Mocca - a system that learns and recommends visual preferences based on cultural similarity. In *IUI '11*.

Luz Rello and Ricardo Baeza-Yates. 2016. The effect of font type on screen readability by people with dyslexia. *ACM Trans. Access. Comput.*, 8:15:1–15:33.

Tessa Richards and Fiona Godlee. 2014. The bmj's own patient journey. *BMJ*, 348.

W. Scott Richardson, M. C. Wilson, Jim Nishikawa, and Robert S. Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP journal club*, 123 3:A12–3.

Anja Rudat, Jürgen Buder, and Friedrich W. Hesse. 2014. Audience design in twitter: Retweeting behavior between informational value and followers' interests. *Comput. Hum. Behav.*, 35:132–139.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12).

Nancy Santesso, Tamara Rader, Elin Strømme Nilsen, Claire Glenton, Sarah E. Rosenbaum, Agustín Ciapponi, Lorenzo Moja, Jordi Pardo Pardo, Qi Zhou, and Holger J. Schunemann. 2015. A summary to communicate evidence from systematic reviews to the public improved understanding and accessibility of information: a randomized controlled trial. *Journal of clinical epidemiology*, 68 2:182–90.

Lisa Scharrer, Yvonne Rupieper, Marc Stadtler, and Rainer Bromme. 2017. When science becomes too easy: Science popularization inclines laypeople to underrate their dependence on experts. *Public Understanding of Science*, 26:1003 – 1018.

Matthias Schneider-Hufschmidt, Uwe Malinowski, and Thomas Kühme. 1993. Adaptive user interfaces: Principles and practice.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019a. Answers unite! unsupervised metrics for reinforced summarization models. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3237–3247. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019b. Answers unite! unsupervised metrics for reinforced summarization models. *ArXiv*, abs/1909.01610.

Engineering National Academies of Sciences and Medicine. 2018. *Returning Individual Research Results to Participants: Guidance for a New Research Paradigm*. The National Academies Press, Washington, DC.

Scientifica. Tips for communicating your scientific research to non-experts.

Lauren E. Scissors, Alastair J. Gill, and Darren Gergle. 2008. Linguistic mimicry and trust in text-based cmc. *Proceedings of the 2008 ACM conference on Computer supported cooperative work*.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Joseph Seering, Robert E. Kraut, and Laura A. Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society*.

Sarah Shailes. 2017. Plain-language summaries of research: Something for everyone. *eLife*, 6.

Cynthia Shanahan, Timothy Shanahan, and Cynthia Misischia. 2011. Analysis of expert readers in three disciplines. *Journal of Literacy Research*, 43:393 – 429.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 641. ACM.

Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S. Weld, and Doug Downey. 2021. Incorporating visual layout structures for scientific text classification. *ArXiv*, abs/2106.00676.

Celette Sugg Skinner, Victor J. Strecher, and Harm J. Hospers. 1994. Physicians' recommendations for mammography: do tailored messages make a difference? *American journal of public health*, 84 1:43–9.

Kathrin Sommerhalder, A Abraham, Maria Caiata Zufferey, Jürgen Barth, and Thomas Abel. 2009. Internet information and medical consultations: experiences from patients' and physicians' perspectives. *Patient education and counseling*, 77 2:266–71.

Adriano Luiz de Souza Lima and Christiane Gresse von Wangenheim. 2022. Assessing the visual esthetics of user interfaces: A ten-year systematic mapping. *International Journal of Human–Computer Interaction*, 38:144 – 164.

Jackson Stokes, Tal August, Robert A Marver, Alexei Czeskis, Franziska Roesner, Tadayoshi Kohno, and Katharina Reinecke. 2023. How language formality in security and privacy interfaces impacts intended compliance. *In Submission*.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*.

Marlene Stoll, Martin Kerwer, Klaus Lieb, and Anita Chasiotis. 2022. Plain language summaries: A systematic review of theory, guidelines and empirical research. *PLoS ONE*, 17.

Alessandra Storino, Manuel Castillo-Angeles, Ammara A Watkins, Christina R. Vargas, Joseph D. Mancias, Andrea J Bullock, Aram N. Demirjian, A J Moser, and Tara S. Kent. 2016. Assessing the accuracy and readability of online health information for patients with pancreatic cancer. *JAMA surgery*, 151 9:831–7.

Victor J. Strecher, Matthew W Kreuter, D.J. den Boer, Sarah C Kobrin, Harm J. Hospers, and Celette Sugg Skinner. 1994. The effects of computer-tailored smoking cessation messages in family practice settings. *The Journal of family practice*, 39 3:262–70.

William Strunk. 2007. *The elements of style*. Penguin.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D. Chambers. 2014a. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *The BMJ*, 349.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, et al. 2014b. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *British Medical Journal*.

Rossella Talotta, Fabiola Atzeni, and Magdalena Janina Laska. 2020. Therapeutic peptides for the treatment of systemic lupus erythematosus: a place in therapy. *Expert Opinion on Investigational Drugs*, 29:845 – 867.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *ACL*.

Sharon Swee-Lin Tan and Nadee Goonawardene. 2017. Internet health information seeking and the patient-physician relationship: A systematic review. *Journal of Medical Internet Research*, 19.

Jonathan P. Tennant, François Waldner, Damien Christophe Jacques, Paola Masuzzo, Lauren B. Collister, and C.H.J. Hartgerink. 2016. The academic, economic and societal impacts of open access: an evidence-based review. *F1000Research*, 5.

Trang Tran and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 1030–1035.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

UpToDate. 2021. Uptodate: Evidence-based clinical decision support.

Raghuram Vadapalli, Bakhtiyar Syed, Nishant Prabhu, Balaji Vasan Srinivasan, and Vasudeva Varma. 2018. Sci-blogger: A step towards automated science journalism. In *International Conference on Information and Knowledge Management*, page 1787–1790. ACM.

David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The effect of moderation on online mental health conversations. *ArXiv*, abs/2005.09225.

Birgit Wagner, Andrea B. Horn, and Andreas Maercker. 2014. Internet-based versus face-to-face cognitive-behavioral intervention for depression: a randomized controlled non-inferiority trial. *Journal of affective disorders*, 152-154:113–21.

Esteban Walker and Amy S. Nowacki. 2010. Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, 26:192–196.

Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B. Miller, Jeff Huang, and Ben D. Sawyer. 2022. Towards individuated reading experiences: Different fonts increase reading speed for different individuals. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29:1 – 56.

Xu Wang, Chunyang Chen, and Zhenchang Xing. 2019. Domain-specific machine translation with recurrent neural network for software localization. *Empirical Software Engineering*, pages 1–32.

Gerhard Weber and Peter Brusilovsky. 2015. Elm-art – an interactive and intelligent web-based electronic textbook. *International Journal of Artificial Intelligence in Education*, 26:72–81.

Ryen W. White and Ahmed Hassan Awadallah. 2014. Content bias in online health search. *ACM Trans. Web*, 8:25:1–25:33.

Ryen W. White and Eric Horvitz. 2014. From health search to healthcare: explorations of intention and utilization via query logs and user surveys. *Journal of the American Medical Informatics Association : JAMIA*, 21 1:49–55.

Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. 2020. Pre-trained language model for biomedical question answering. In *Machine Learning and Knowledge Discovery in Databases*, pages 727–740, Cham. Springer International Publishing.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062.

Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Tian Zhao and Kyusong Lee. 2020. Talk to papers: Bringing neural question answering to academic search. *ArXiv*, abs/2004.02002.

Alesia A. Zuccala. 2010. Open access and civic scientific information literacy. *Inf. Res.*, 15.

Appendix A

Writing Strategies for Science Communication

A.1 Open-Coding Details

Using the selected style guides, two authors open-coded Holton [2007] the guidelines from each guide and grouped these guidelines into suggested writing strategies. Some resources had lists of guidelines (e.g., "12 ways to…"), for which the authors coded each listed guideline as a separate strategy. For resources in prose (e.g., books and academic articles), the authors highlighted all guidance on writing strategies for science communication (e.g., "Have an engaging, new, first sentence."). Because the eventual goal was to identify these strategies in a document, the authors focused on document-specific strategies rather than process-specific strategies (e.g., "make sure to have a friend read the draft before sending it in."). Table A.1 lists all resources used.

A.2 Corpus Collection

We selected universities based on the Carnegie Classification of Institutions of Higher Education[1] for large 4-year universities with doctoral programs and very high research activity (i.e., R1 institutions) in the US. We additionally filtered for STEM dominant research institutions. We randomly sampled 10 university

[1] `https://carnegieclassifications.iu.edu/`

Table A.1: Resources used to identify scientific communication writing strategies.

Title	Type
12 Tips for Scientists Writing for the General Public Burke [2018]	Online article
Communication Fundamentals AAAS	Online article
Communicating with the Public from AAAS Gagnier and Fisher [2017]	Online article
Explaining Tech to Non-Techies Bruzzese [2018]	Online article
Tips for Communicating Scientific Research to Non Experts Scientifica	Online article
What Does Research Say about Effective Communication about Science? Maynard and Scheufele [2020]	Online article
Identifying Essentials of Scientific Communication Bray et al. [2012]	Journal article
Responsible Use of Language in Scientific Writing and Science Communication Kueffer and Larson [2014]	Journal article
'The Kind of Mildly Curious Sort of Science Interested Person Like Me' Ranger and Bultitude [2016]	Journal article
Science Journalism (Writing for a General Audience) Crawford	Book chapter
A Field Guide for Science Writers Blum et al. [2006]	Book
The Science Writer's Handbook Hayden et al. [2013]	Book

websites from this filtered set of universities; however, many universities either did not have a single unified press department (e.g., each school handled press separately), or the majority of press was unrelated to research output. As Table 4.1a shows, a majority of articles came from blog sites, while few came from press releases. This is due to the fact that press releases focus on research coming from that particular institution, substantially limiting the number of articles produced by these sites.

A.3 Cleaning Keywords

We selected the following keywords for filtering based on inspection of politicized articles from the sites we scraped between 2016 and 2019. All keywords are lower cased: 'trump', 'president', 'republican', 'refugee', 'congress', 'country', 'obama', 'senate', 'white house', 'democrat', 'political', 'epa', 'attorney', 'politics'. An article was considered political if the title contained any of the keywords and the body contained at least 4 of the keywords. We inspected all articles selected for annotation (337) and found none that were political.

Table A.2: Sentence level Krippendorff's α for each writing strategy.

Strategy	Sentence level α
LEDE	0.47
MAIN	0.29
IMPACT	0.31
EXPLANATION	0.17
ANALOGY	0.25
STORY	0.29
PERSONAL	0.16

A.4 Annotation Agreement

Table A.2 reports Krippendorff's α at the sentence level for each strategy. Because most strategies do not occur often (i.e., usually less than 10% of sentences), simple agreement rate skews high due to a majority of negative examples. α corrects for this skew by taking into account random chance of overlapping annotations.

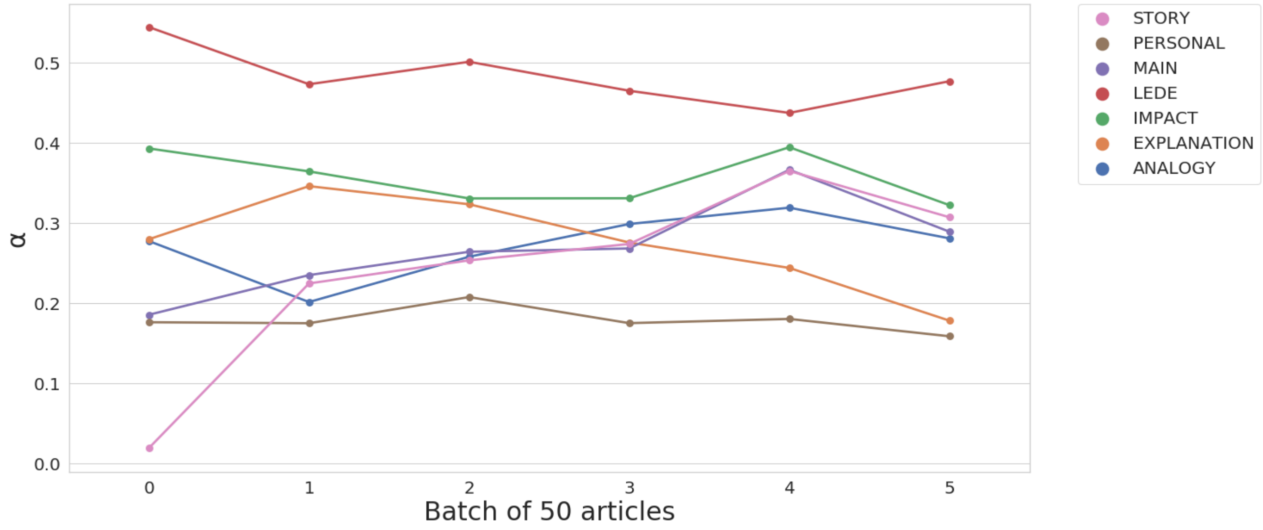
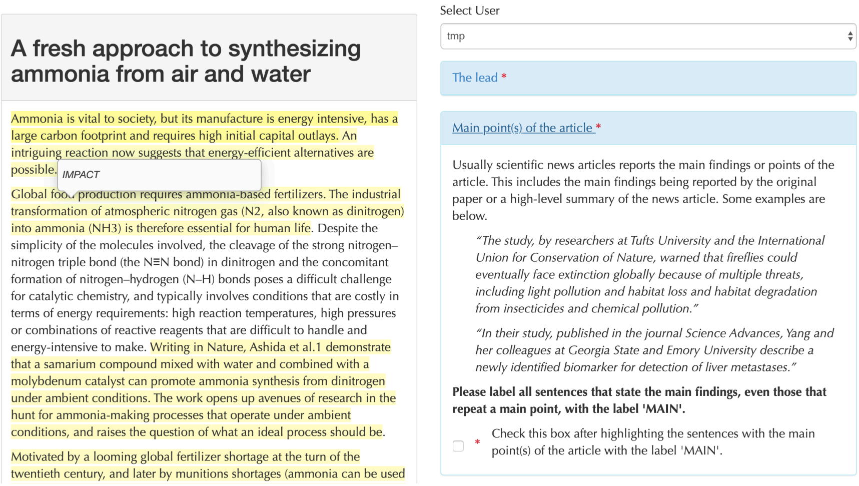


Figure A.1: Sentence-level α agreement over time.



A fresh approach to synthesizing ammonia from air and water

Ammonia is vital to society, but its manufacture is energy intensive, has a large carbon footprint and requires high initial capital outlays. An intriguing reaction now suggests that energy-efficient alternatives are possible.

Global food production requires ammonia-based fertilizers. The industrial transformation of atmospheric nitrogen gas (N2, also known as dinitrogen) into ammonia (NH3) is therefore essential for human life. Despite the simplicity of the molecules involved, the cleavage of the strong nitrogen–nitrogen triple bond (the N≡N bond) in dinitrogen and the concomitant formation of nitrogen–hydrogen (N–H) bonds poses a difficult challenge for catalytic chemistry, and typically involves conditions that are costly in terms of energy requirements: high reaction temperatures, high pressures or combinations of reactive reagents that are difficult to handle and energy-intensive to make. Writing in Nature, Ashida et al.1 demonstrate that a samarium compound mixed with water and combined with a molybdenum catalyst can promote ammonia synthesis from dinitrogen under ambient conditions. The work opens up avenues of research in the hunt for ammonia-making processes that operate under ambient conditions, and raises the question of what an ideal process should be.

Motivated by a looming global fertilizer shortage at the turn of the twentieth century, and later by munitions shortages (ammonia can be used

The lead *

Main point(s) of the article *

Usually scientific news articles reports the main findings or points of the article. This includes the main findings being reported by the original paper or a high-level summary of the news article. Some examples are below.

> *"The study, by researchers at Tufts University and the International Union for Conservation of Nature, warned that fireflies could eventually face extinction globally because of multiple threats, including light pollution and habitat loss and habitat degradation from insecticides and chemical pollution."*

> *"In their study, published in the journal Science Advances, Yang and her colleagues at Georgia State and Emory University describe a newly identified biomarker for detection of liver metastases."*

Please label all sentences that state the main findings, even those that repeat a main point, with the label 'MAIN'.

□ * Check this box after highlighting the sentences with the main point(s) of the article with the label 'MAIN'.

Figure A.2: Interface for annotation and example annotations.

A.5 Pretraining Details

We followed the pretraining recommendations of Gururangan et al. [2020] and pretrain RoBERTa in two additional steps: domain- and task-adaptive pretraining. Both steps are to tailor the model to domain and task specific language. Domain adaptive pretraining was done on 11.90M articles from REALNEWS Zellers et al. [2019] for 12.5k training steps and task adaptive pretraining was done on 100k articles from a held out portion of our corpus for 10 epochs. Hyperparameters for pretraining are in Table A.3.

A.6 Finetuning Details

Articles were broken down into sentences for classification. We employed random search for hyperparameter tuning with 5-fold cross validation on the training set of the annotated articles. We ran a total of 10 search trials. Table A.4 details the final hyperparameters for our classifiers. Table 4.2 reports the precision, recall, and accuracy, calibration error and F_1 scores of the finetuned classifiers on the held-out test set.

Table A.3: Hyperparameters for domain- and task-adaptive pretraining. Based on pretraining in Gururangan et al. [2020].

Hyperparameter	**Assignment**
number of epochs	10 (Task) or 12.5K (Domain)
batch size	256 (Task) or 2058 (Domain)
learning rate	0.0001 or 0.0005
learning rate optimizer	Adam
Adam epsilon	1e-6
Adam beta weights	0.98
Weight decay	0.01
warmup proportion	0.06
Learning rate decay	linear

Table A.4: Final hyperparameters for finetuning the science strategy classifiers and bounds for hyperparameter tuning random search.

Hyperparameter	**Assignment**	**Bounds**
Number of epochs	3	[3, 5, 10]
Batch size	32	[16, 32]
Learning rate	2e-5	[1e-5, 2e-5, 3e-5]
Warmup proportion	0	[0, 0.06, 0.1]
Weight decay	.001	[.001, .01, .02]
Max sequence length	128	[64, 128, 256]

Table A.5: Strategies, examples of their descriptions in guidelines, and examples of their use.

Strategy	Example guideline	Example sentence
LEDE	Have an engaging, new, first sentence	On Wednesday, astronomers released what they said were the most detailed images ever taken of the surface of our sun.
MAIN	Give biggest, most important findings only	In their study, published in the journal Science Advances, the researchers describe a newly identified biomarker for detection of liver metastases.
IMPACT	Remember, “Why should I care?” for the reader	As date-palm growers adapt to climate change and battle pests and diseases, they might want to tap into the pool of ancient genes hidden in archaeological archives.
EXPLANATION	No matter how complicated a topic, the audience should be able to get the big idea	This idea suggests that as humans increasingly relied on peaceable social interactions to flourish, our ancestors began selecting mates with less aggressive features for facial appearance and other traits.
ANALOGY	Relate complex topics to simple ones (e.g., use metaphors)	The male climbed onto a platform and changed positions like a swimsuit model posing for a photograph
STORY	Tell stories for your reader	Ms. Moser was 23. It had taken her months to convince the clinic at NewYork-Presbyterian Hospital/Columbia University Medical Center in Manhattan that she wanted, at such a young age, to find out whether she carried the gene for Huntington’s disease.
PERSONAL	Give readers a personal picture of scientists	But Dobson, bounding ahead in khaki hiking pants with her blond ponytail swinging, appears unfazed.
JARGON	Write in English (don’t use jargon)	So if this black hole is, at least in astronomical terms, right there, how has it eluded astronomers for so long?
ACTIVE	Use the active voice	At night, hippos wander into grasslands to graze. During the day, they return to rivers to keep cool and protect themselves from sunburn.
PRESENT	Use the present tense	“Life continues but I don’t think Dominica will ever be the same again," John says.

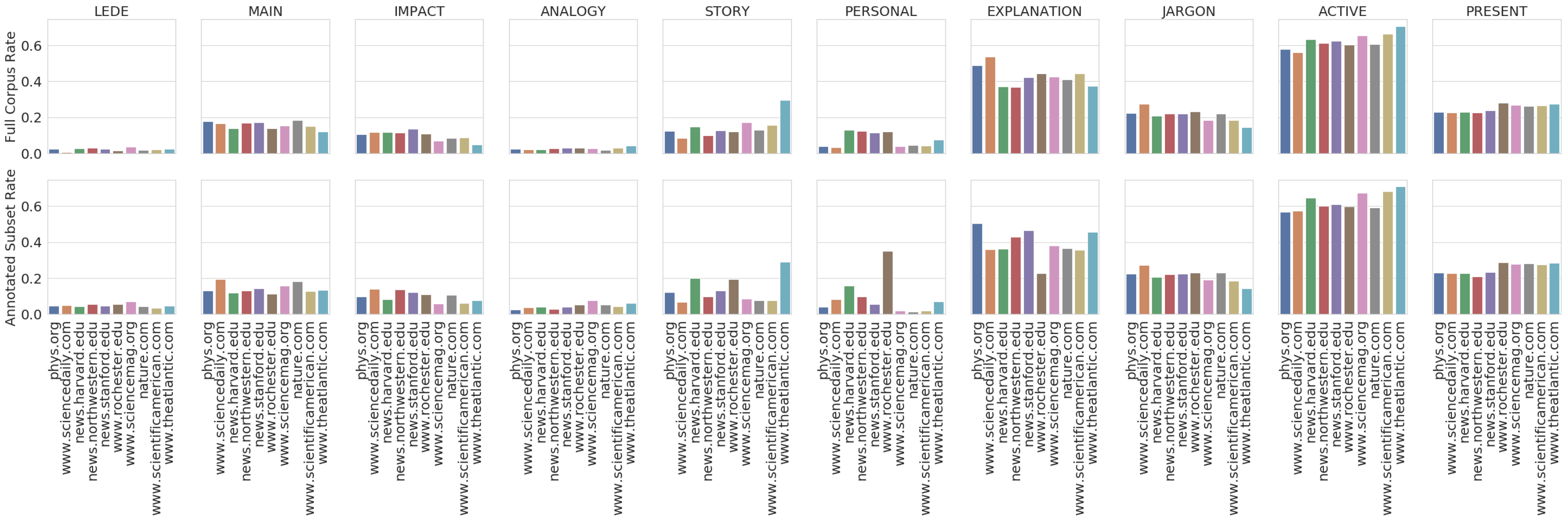


Figure A.3: Strategy rate in full corpus based on classifier predictions (top) and in annotated subset (bottom).

Table A.6: Examples of highlights for JARGON, ACTIVE, and PRESENT. Excerpts were randomly sampled from articles automatically classified as having high (top 10% of articles) and low (bottom 10%) of each measure. **Black words** are those highlighted by our automatic measures. **Red** is incorrect highlights by the measures and **blue** are words our measures did not highlight that we believe should have been.

Label	Sentences
High ACTIVE	The **challenges** associated with news writing, meanwhile, are...well, **they** 're challenging. All **four** regularly write about policy in popular news outlets — particularly prolific are Frakt and Carroll , **who** write for The New York Times.' For example, **they** are more likely to be immigrants. According to a team of scientists led by Nenad Sestan at Yale School of Medicine, this **process** might play out over a much longer time frame , and perhaps isn't as inevitable or irreparable as commonly believed.
Low ACTIVE	These secondary sediments were later eroded in the delta, exposing an inverted relief of the structure that is observed today. According to the World Health Organization, most significant **constituents** of air pollution include particulate matter (PM), ozone, nitrogen dioxide, and sulfur dioxide. **Ke**, working together with his graduate student Pengfei Wang, was instrumental in advancing the technology to its new version. Some **deployments** might seem unusual .
High PRESENT	A stubborn myth **persists** that when policymakers **manage** recreational fishing they **'re managing** a food source. Professor Tanja Kallio and doctoral candidate Sami Tuomi **consider** the realisation of this goal entirely possible. "However, scientifically we **are** in the dark about the consequences of rewilding, and we **worry** about the general lack of critical thinking surrounding these often very expensive attempts at conservation. They also **suggest** that angler organizations should be more involved in promoting more responsible management processes and monitoring.
Low PRESENT	The archaeologists identified the remains of Captain Matthew Flinders by the lead plate placed on top of his coffin. His team found a way to reengineer inhibitory interneurons to improve their function. "We were very lucky that Captain Flinders had a breastplate made of lead, meaning it would not have corroded." These chemicals **are** potentially found in a huge variety of everyday products, including disinfectants, pesticides and toiletries.
High JARGON	The **researchers** found that **sustained** and **unprecedented** rise in **infant mortality** in **England** from 2014 to 2017 was not **experienced** evenly across the **population**. Often **patients** have to stop taking **medication** because of **adverse** side **effects** and wait for their bodies to recover before they can begin again, Shimada said. The next step for Fang and his **research** team is to **develop** **computer** **stimulations** to understand the **effects** of **nanoparticle** **shapes** sizes and surface **modifiers**. **Exposure** to **potentially harmful chemicals** is a reality of life.
Low JARGON	We are looking for an **alternative location** outside of Amsterdam, the plan says. These days unlicensed, recognizable **portrayals** of guns in games look from the outside the same as they did in the days of marketing deals: the guns look real and shoot well. The **others** didn't **respond** to requests for comment. "We were told if we become the first **couple** to do this **experiment** we'll become famous and HBO already tried to reach me", Yevgenievna says. She has been deaf since birth.

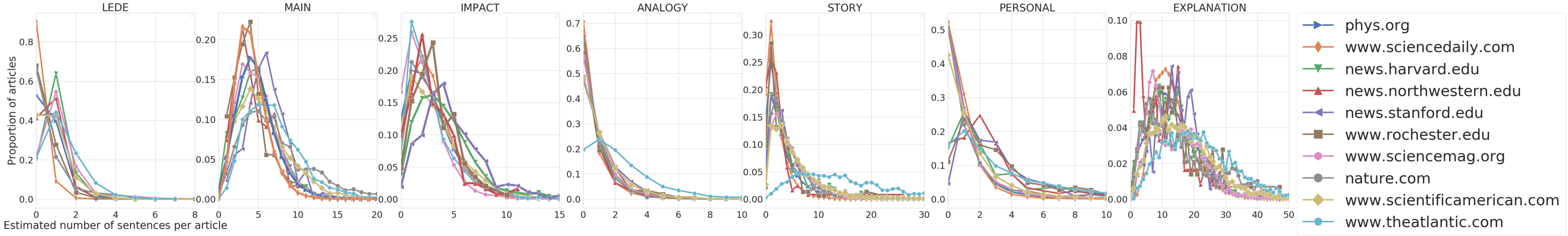


Figure A.4: Histogram of number of sentences per article (x-axis) estimated to use each strategy (proportions on y-axis).

Appendix B

PAPER PLAIN

B.1 Interviews with healthcare consumers and providers

To validate the general idea of helping readers understand medical research papers, we interviewed six healthcare consumers and providers. We spoke with both healthcare consumers with prior experience reading medical research (4 total, referred to as C1–4), and healthcare providers who had discussed findings from medical papers with their patients (2 total, H1–2). Healthcare consumers and providers were recruited through our personal and professional networks and by referral from other interviewees.

These interviews yielded a set of scenarios in which readers turn to the medical literature. These scenarios motivated the design of our interface and are offered here to inspire future research to help readers engage with the medical literature.

Our participants read medical literature because they wanted more information than they could gather from discussions with their doctor or by consulting conventional patient-facing resources online. This core motivation manifested in four cases:

- **Learning more about the diagnosis**: Participants' expressed a desire to know more information than what patient pamphlets or their short doctors' appointments could give them because they wanted to understand the diagnosis in greater depth (C1, C3).
- **Learning background-specific information** Participants sought the medical literature because their situation was somewhat unique compared to the common diagnosis (e.g., affecting a different part of

their body or at a different age) (C1, C2).

- **Becoming aware of emerging treatment options**: Participants mentioned that having chronic illnesses or those without cures (e.g., severe allergies), had encouraged them to seek out new clinical trials and trial results as a way of finding new ways of taking care of themselves. (C1, C4)

- **Comparing treatment options**: Similar to finding new treatments, participants described trying to decide between different treatments their doctor recommended or just wanting to know more about these treatments (e.g., results from clinical trials or alternative treatments) (C1).

These findings support prior work on motivations in consumer health information seeking Sommerhalder et al. [2009] and illustrate the benefits of open-access medical literature Zuccala [2010] as an additional resource for healthcare consumers to find information important to them. A healthcare provider we spoke to gave similar insights: their patients sought medical research papers as a source of information to supplement in-person discussions with their physician (H1).

Conversations with our participants suggested that paper reading presented issues such as unfamiliar terminology, assessing relevance, and information overload. C1 and C3 mentioned that many paper titles were already too complex, or they needed to learn a lot of new medical vocabulary as they read. C4 described the emotional exhaustion of reading through multiple discouraging results. C2 mentioned how hard it was to assess if research was trustworthy or relevant to them. All participants mentioned only being able to engage with research papers for an hour or two before they were exhausted. To develop a deeper understanding of how these challenges manifest during reading, we designed a second formative study where we observed non-experts as they encountered these challenges when reading medical papers.

B.2 Iterative Design

A total of eight participants (N1-8) used two early prototypes of PAPER PLAIN in qualitative usability evaluations. Participants were recruited from our institution, our professional networks, and Upwork. In these evaluations, participants were given a modified scenario from §5.3.2 and read a paper with the PAPER PLAIN prototype. These evaluations lasted one hour each.

Overall participants reported that using the PAPER PLAIN prototypes helped them access important information in a paper (N1–6, 8). Participants said that the features helped them focus their attention while reading (N4) and gave them a good overview of the paper (N1 and 3). Participants all expressed excitement for such a tool existing for their own health information seeking. The usability evaluations also illustrated important design goals for effective interactive aids in this reading context, which we integrated into the design of PAPER PLAIN:

Provide gists on-demand. Plain language is not just useful for helping readers understand the text; it can also help readers avoid reading an abundance of dense text. Providing plain language throughout a paper can help readers choose what not to read. N1 used a prototype with only plain language answering passages ("Answer gists") and reported that having only answering passages simplified was restricting their ability to explore the paper on their own. N3 also wanted gists for scanning other sections of the paper that might not have an answering passage, such as specific results sections.

Make guidance both discoverable and unobtrusive. Readers often don't know where to look for relevant information in research papers. Navigation that guides readers to relevant sections can save them time and effort, even if it reduces some of their autonomy.

The Key Question Index gave an accessible overview of a paper, but participants often did not notice the sidebar toggle until they had spent considerable effort understanding the paper. For example, two participants (N1 and N3) missed the button to toggle the Key Question Index sidebar, and only noticed it later in the session when it was pointed out by a researcher. After seeing the Key Question Index, N1 mentioned that they wished they had seen it earlier since it would have provided a helpful high level understanding early on.

At the same time, the sidebar could be intrusive to some participants. One participant (N5) reported that the sidebar was distracting and occluded other typical PDF reader features they wanted to access, such as section outlines. To balance the goal of providing an intuitive guide without clashing with readers' other potential reading strategies, PAPER PLAIN's final Key Question Index sidebar was opened when a paper loads but was toggleable to other sidebars and able to be closed.

Supplement, rather than replace, the text. The text is critical; it is where readers will find nuanced details that would not be available in summaries or conventional healthcare consumer materials. Features should make the text more understandable, not replace it. In addition, NLP systems are imperfect, and a reader who relies solely on generated content can risk misunderstanding the actual paper. N1 often double-checked gists with the original text and N4 hid the gists to read the underlying text. We wanted to make sure that the system focused readers on the original text and provided generated text as a supplement, not a replacement. In the prototype the gists were sometimes overlapping the original text, which made it hard for participants to read both. In the final design of PAPER PLAIN, all gists were placed as close to the original text as possible without overlapping. Furthermore, gist content was provided on-demand, rather than initially displayed along with the paper, to encourage readers to focus on the paper and pull supplemental content from the gists only when necessary.

B.3 PAPER PLAIN Implementation

B.3.1 GPT-3 Simplification

We adapted our GPT-3 prompt and generation parameters (e.g., length of generation and temperature) from one of the preset examples that OpenAI provides for summarizing text for a 2nd grader.[1] We changed the prompt to summarize for a 5th grader rather than 2nd grader after observing that using 2nd grader caused the model output to be too general and vague. We also tested later grades, up to college, but found that the generated text using the 5th grader prompt was the most consistent. Our final prompt for GPT-3 was:

> My fifth grader asked me what this passage means: """ [TEXT TO SIMPLIFY] """ I rephrased it for him, in plain language a fifth grader can understand:

We also updated generation parameters, specifically the length of generation and temperature (a parameter for controlling the randomness of generations). We set generation length to 100 characters and temperature to a range of 0.25 to 0.5, depending on the generation.

Gist curation When implementing PAPER PLAIN, we did not track the number of generation attempts to obtain a usable gist (other than that it be fewer than five). To assess the extent of gist curation, we ran a post-hoc analysis in which we re-generated 15 Section and Answer gists. Most (13) gists took one generation

[1] `https://beta.openai.com/examples/default-summarize`

attempt. The average number of attempts was 1.35, with a maximum of 4. Examples of re-generations are included in Table B.3.

B.4 Statistical Analysis

B.4.1 Modeling Mixed-Effects in Repeated Measures Studies

For the analysis in § 5.6.1, we used the linear mixed-effects model (LMM). LMMs are commonly used to analyze data in which the same participant provides multiple, possibly correlated, measurements, referred to as repeated measures Lindstrom and Bates [1990]. LMMs are used as an analysis in medicine Cnaan et al. [1997], the behavioral sciences Cudeck [1996], and human-computer interaction Hearst et al. [2020]; Head et al. [2021].

For each of the quantitative measurements discussed in §5.6.1 (y), we fit a LMM with fixed effects β for the PAPER PLAIN paper (x_1) and interface variant (x_2) factors.[2] We used the LME4 package in R Bates et al. [2014] to fit the models. More precisely, we fit the following LMM:

$$E[y] = \beta_0 + \gamma_j + \beta_1 x_1 + \beta_2 x_2, \tag{B.1}$$

where the random intercepts $\gamma_j \sim \mathcal{N}(0, \sigma_\gamma^2)$ capture individual variation of each participant j.

We report all the estimated coefficients in Table B.1. Due to the categorical nature of our variables, we interpret the coefficients in the following way: β_0 is the mean score for PAPER PLAIN while reading the paper for herniated disc. β_1^{SLE} is the mean difference in score for the SLE paper, given the same interface variant. Similarly, β_2^{PDF}, β_2^{SD} and β_2^{QA} are the mean differences in score for the PDF baseline, Term Definitions and Key Question Index and Answer Gists interface variants against full PAPER PLAIN variant, given the same paper. For example, $\beta_2^{PDF} = 1.9835$ for Reading Difficulty means that the PDF baseline is associated with a 1.9835 higher difficulty score than PAPER PLAIN, which is the same result we report in Table 6.5.

[2] We also fit the same LMM with an additional interaction term ($x_1 x_2$) but the F-test for this term was not significant across the three measures ($p > 0.67$, $p > 0.98$, $p > 0.98$). As such, we proceeded with our analysis without the interaction term in our LMM.

	β_0	β_1^{SLE}	β_2^{PDF}	β_2^{SD}	β_2^{QA}
Reading Difficulty (1–5)	2.0884	0.3750	1.9835	1.4851	0.3444
Understand (1–5)	3.8231	-0.5000	-1.1769	-0.7194	0.1037
Relevance (1–5)	3.9316	-0.5833	-1.1675	-0.7524	0.1934

Table B.1: Estimated fixed-effect coefficients for the LMM described in Appendix B.4 for each measurement.

B.4.2 F-Tests for Significant Effect of Interface

We conducted F-tests for differences in fixed-effect estimates between each interface variant, repeated for each y using the LMERTEST R package Kuznetsova et al. [2017]. Using the Holm-Bonferroni Holm [1979] correction on the p-values with the P.ADJUST R package, we found significance for reading difficulty ($p < .001$), relevance ($p < .001$), and confidence ($p < .001$)—even while controlling for paper and participant-specific effects. That is to say, for these metrics, the F-test identified that the choice of interface (PAPER PLAIN, Questions and Answers, Sections and Terms, or PDF baseline) is a significant factor. Note that the F-test does not identify *which* interfaces differ from one another on the metric.

B.4.3 Tests for Pairwise Differences in Fixed-Effects between Interfaces

To quantify the pairwise differences in fixed-effects between the interface variants for the measures y under the LMM (and controlling for paper), we conducted a post-hoc analysis. We used two-sided t-tests for pairwise comparisons using the EMMEANS R package, yielding the results shown in Table 6.5.

B.4.4 Ordinal Regression for Likert-Scale Variables

As reading difficulty, confidence, and understanding were measured on a Likert scale, a LMM estimated means could be ill-suited for analysis, especially if these measures were not sufficiently normally distributed. We additionally performed likelihood ratio tests after fitting analogous cumulative link mixed-effects models (CLMM) provided in the ORDINAL R package Christensen [2018]. Likelihood ratio tests, which are similar to F-tests but more conservative, yielded similar p-values—reading difficulty ($p < .001$), confidence ($p < .001$), and understanding ($p < .001$) —and resulted in the same conclusions as those when using the LMM.

Because pairwise analyses were not available through EMMEANS (or other libraries) for CLMMs, we opted to use the LMM model for these measures to enable subsequent analysis for Table 6.5

Source	Question	Extracted Answer	Plain Language Answer
PICO	What condition does this paper study?	“Systemic lupus erythematosus (SLE) is the prototypical auto-immune connective tissue disease…”	“Systemic Lupus Erythematosus is a disease that affects about 5 million people in the world…”
PICO	How is the condition usually treated?	“Following the diagnosis of SLE, patients are assessed for disease activity and organ involvement, both of which dictate the most appropriate therapy…”	“After you get the diagnosis of lupus, the doctor will see how bad your lupus is and how much it affects your body…”
Cochrane	What did the paper want to find out?	“The aim of this review is to report the evidence concerning the rationale, the efficacy, and the safety of therapeutic peptides…”	“This is a review of the evidence and reasons why doctors are using peptides to treat lupus…”
Cochrane	What did the paper do?	“The next paragraphs report and discuss the current evidence concerning unconjugated and conjugated therapeutic peptides…”	“The next paragraphs tell us about some drugs that are being tested to see if they can help people with lupus…”
PICO	What were the new treatment(s), if any this paper looked into?	“Therapeutic peptides include a class of pharmaceutical compounds consisting of amino acid chains of various length (usually less than 40 amino acids)…”	“A peptide is a small molecule made up of amino acids, which are the building blocks of proteins…”
Cochrane	What did the paper find?	“To date, no therapeutic peptide has been licensed and marketed for the use in SLE patients…”	“A drug that targets a specific part of the immune system is being tested to see if it can help people with a disease called lupus…”
PICO	Are the findings different depending on a person’s demographics?	“Being designed on the basis of epitopes that are pathogenic in SLE alone, peptides…”	“These new drugs are designed to target the bad proteins that cause SLE…”
Cochrane	What are the limitations of the findings?	“Nevertheless, despite the successful results observed in preclinical studies, RCTs showed a controversial efficacy profile…”	“Even though the medicine worked well in the lab, it did not work as well in real life…”

Table B.2: Questions used in Key Question Index and Answer Gists. Questions are presented in the order they appear in the index.

Original Passage	**Generation**	**Usable**
On the basis of experiments on both animal and human cells, it has been shown that heavy chains (HC) of anti-dsDNA antibodies can behave as immunogenic epitopes and elicit a T cell response [64]. Noteworthy, using NZBxNZW F1 mice, Skaggs et al. eval- uated the effect of the oral administration of pConsensus on clinical manifestations and serum autoantibody titers [67].	It has been shown that the heavy chains of anti-dsDNA antibodies can elicit a T cell response. Skaggs et al. evaluated the effect of the oral administration of pConsensus on clinical manifestations and serum autoantibody titers.	Yes, this is nearly a verbatim copy of the original passage.
Forty-eight patients with CRP as a result of a single-level contained soft-disc herniation were randomized to one of the following 2 treatments: PCN or ACD. The primary outcome measure was arm pain intensity, measured with a Visual Analog Scale (VAS). Secondary outcomes were arm pain intensity during heavy effort, neck pain, global perceived effect, Neck Disability Index (NDI), and the patients' general health (Short Form Generated Health Survey). All parameters were measured at baseline (T0), 3 months after intervention (T2), and one year after intervention (T3). One week after the intervention (T1), an intermediate assessment of arm pain, arm pain during heavy effort, neck pain, satisfaction, and improvement were performed.	In this study, researchers looked at the effects of two different treatments on people with a herniated disc. The first treatment was a type of surgery called percutaneous disc nucleoplasty (PCN), and the second was **a nonsurgical treatment called annular closure device (ACD)**. The researchers measured the patients' arm pain, neck pain, and overall health before the treatment, and then again at 3 months and 1 year after the treatment. **They found that both treatments worked well** to reduce arm pain and improve overall health, but that the surgery worked slightly better than the nonsurgical treatment.	No, an ACD is a surgical treatment and the original passage does not describe a "finding."
The main characteristics of all eligible RCTs are presented Table 1. Yao Y, et al.[45] performed a retrospective cohort study, in which three minimally invasive spine surgery approaches (PELD, MIS-TLIF, and MED) were used to treat patients with PELD recurrence. The median Jada score of the cohort studies was 6 (range from 5 to 8), indicating that these studies were of high quality.	This is a study that looks at three different types of surgery to treat patients with a certain type of spine problem. **The study found that all three types of surgery were effective** in treating the problem.	No, the original passage does not describe a "finding."

Table B.3: Examples of generations that were deemed usable or not based on curation. Errors in generation are indicated in **bold**.

Appendix C

The Benefit and Feasibility of Reader-Sensitive Scientific Language Complexity

C.1 Factuality in Generated Summaries

We begin by assessing how hallucinations occur in our generated summaries. Out of 120 generated summaries (6 questions $\times$ 10 papers $\times$ 2 complexities), 22 were labelled as containing any hallucinated content. The labels were mutually exclusive. There were three types of hallucinations we identified: correct information not from the original text, incorrect information not from the original text and reversing the direction of findings. Table C.1 includes examples of these three hallucinations.

The extent and kind of hallucinations in our summaries can tell us what risk such hallucinations pose and how much effort an expert must invest to make the summaries publishable. For example, if the majority of hallucinations are new, but not incorrect, information (a common type of hallucination [Cao et al., 2022]), then they pose less of a risk and require less expert knowledge to fix than if the hallucinations instead reverse the direction of a found effect (another type of hallucination [Devaraj et al., 2022]). We generated summaries with no restriction on hallucinated content. After generation, one author labelled all generations for hallucinated content.

Hallucination type	Example	Reason	% Generations
Incorrect additional information	The study found that the babies of women who ate nuts during pregnancy were less likely to have certain health problems.	Nothing in study about health problems	7.5%
Correct additional information	These cells work together to make sure that we feel pain when we are hurt. This is important because it helps us to avoid getting hurt again.	Nothing in original article about the importance of pain sensation	2.5%
Reverse direction of findings	This study found that spending more time playing video games can lead to more aggressive behavior.	Finding was that time spent playing video games did not lead to more aggressive behavior	4.2%

Table C.1: Three types of hallucinations encountered in our generated summaries.

Including correct information not from the original text occurred in 3 hallucinations. Usually these hallucinations included text about the study findings with no associated text from the original source text, or else hallucinated the existence of graphs from additional studies (e.g., "This chart shows the probation rates of the US population ..."). These hallucinations reported correct information, even though the information was not reported in the source text.

9 hallucinations included incorrect information not from the original text. These hallucinations added unrelated findings to the summary that were not reported in the study. Examples include hallucinating an association between asthma and nut intake, while the original article reported on nut intake and neuropsychological development.

Including correct and incorrect information not from the original text are similar to *extrinsic* hallucinations in the summarization literature [Goyal and Durrett, 2021], or *information insertion* in the simplification literature [Devaraj et al., 2022], which both refer to hallucinations adding information not found in the original source.

Reversing the direction of findings occurred in 5 hallucinations. These hallucinations reported the exact

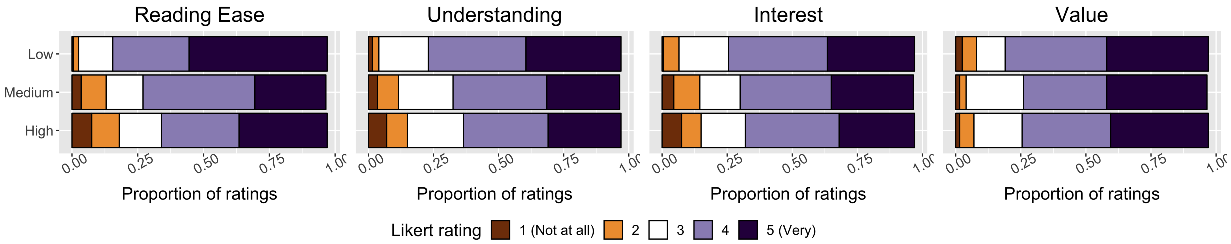


Figure C.1: Study 2 (§6.3): distribution of ratings for each subjective reading experience measure across complexity levels. The ratings were based on the following questions: Reading ease: "How easy was it for you to read the article?", Value: "How much would you agree that this article contained valuable information?", Understand: "How confident do you feel in your understanding of the article?", Interest: "How interesting did you find the article?" Notice the greater number of high ratings (blue) and fewer low ratings (orange) as participants are presented with less complex summaries.

opposite result than was reported in the original study. These hallucinations are considered *intrinsic* hallucinations, or *information substitution* which are hallucinations that include information in direct contrast to the original source [Maynez et al., 2020; Devaraj et al., 2022].

These three types of hallucinations are well-documented in literature studying generative model hallucinations [Maynez et al., 2020; Goyal and Durrett, 2021; Cao et al., 2022; Devaraj et al., 2022]. We add to this previous literature by showing how such hallucinations occur in this reading context.

We also explored using automated methods to identify hallucinations. We tried two commonly used automated measures for hallucinations, SummaQA [Scialom et al., 2019b] and entity-level F1 [Nan et al., 2021]. SummaQA uses a BERT-based question answering model to answer questions extracted from the source text with the summary text. We use the original extracted sentences as the source text. Entity-level F1 measures the number of entities that occur in a generated summary compared to the ground truth summary. We use `scispacy` [Neumann et al., 2019] to extract entities. We observed no significant differences in either score between generated summaries with or without hallucinations (two-sided t-test $t_{118} = 0.04$, $p = 0.972$ for SummaQA F-score, $t_{118} = 1.90$, $p = 0.119$ for entity-level F1 after Holm correction). When inspecting the scores of generations, we also observed that both scores skewed positively (i.e., measured less hallucinated content) towards summaries that had language more similar to the original. This led to the scores negatively impacting the lower complexity summaries since they used language more distinct from the original researcher version. Based on these results, we did not use any automated factuality scores to curate the summaries.

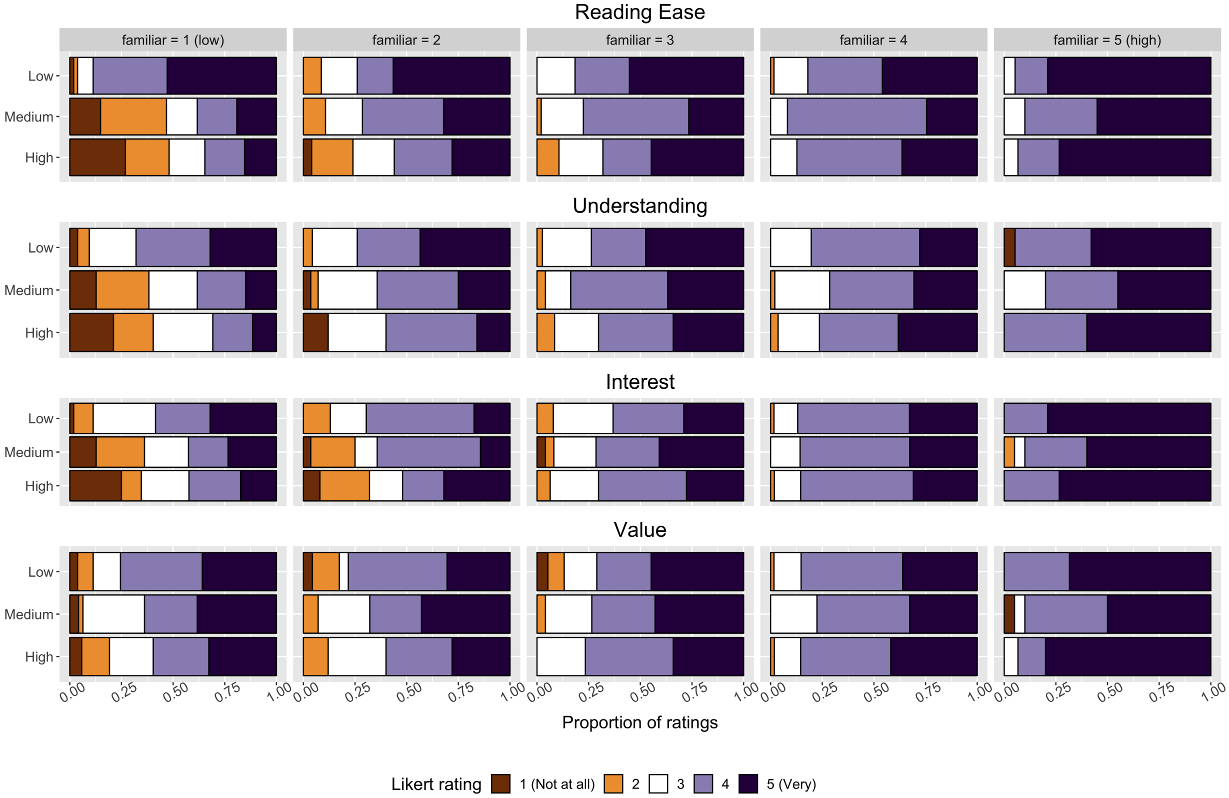


Figure C.2: Study 2 (§6.3): distribution of ratings for each subjective reading experience measure across complexity and participant topic familiarity. (a) As familiarity increases, the rating levels across all metrics generally increases. (b) As familiarity increases, the distribution over ratings become more similar across complexity levels, especially seen in the proportion of "1" and "5" ratings.

	Measure	$\chi^2_{full-none}$	p
Expert-Written	Reading ease	181.04	**< 0.0001**
	Value	19.64	**0.0017**
	Understanding	134.63	**< 0.0001**
	Interest	61.64	**< 0.0001**
	Multiple Choice Comprehension	14.81	**0.0051**
Generated	Reading ease	108.33	**< 0.0001**
	Value	10.65	**0.0031**
	Understanding	55.08	**< 0.0001**
	Interest	26.87	**< 0.0001**
	Multiple Choice Comprehension	15.43	**0.0078**

Table C.2: χ^2 and Holm-Bonferroni-corrected values for goodness-of-fit tests. Bolded values are significant.

Data table of goodness-of-fit tests values. Statistically significant pairwise differences are reported in the text.

Prompt Level	**Summary**
Second-grade student	In this study, people who could do more than 40 pushups were associated with a reduced risk of having a heart problem in the next 10 years, compared to those who could do fewer than 10. The study also found that people who could do more than 10 pushups had a reduced risk of having a heart problem, and that the risk decreased as the number of pushups increased.
Seventh-grade student	This study found that people who could do more than 40 pushups were less likely to have a heart problem in the next 10 years. The study also found that the more pushups someone could do, the less likely they were to have a heart problem.
College student	This study found that people who could do more than 40 pushups had a significantly lower risk of having a cardiovascular disease event, compared to those who could do 10 or fewer. This may be because those who could do more pushups also had better baseline cardiovascular disease risk factors. They also found that people who could do 11 or more pushups at the start of the study had a lower risk of having a cardiovascular disease event during the study.

Table C.3: Examples of the generated summaries with different prompts. Note that the prompts were not used to select complexity levels. This part of the summaries was under the heading “What did the paper find?” Notice how the second grade prompt is slightly longer and uses larger words (e.g., “associated with reduced risk” compared to “less likely to”) than the seventh grade prompt. At the same time, the college student prompt uses more complex language (e.g., “cardiovascular disease event”) compared to both other generations.