Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing

TAL AUGUST, University of Washington, USA
LUCY LU WANG, University of Washington; Allen Institute for AI, USA
JONATHAN BRAGG, Allen Institute for AI, USA
MARTI A. HEARST, University of California, Berkeley, USA
ANDREW HEAD, University of Pennsylvania, USA
KYLE LO, Allen Institute for AI, USA

When seeking information not covered in patient-friendly documents, healthcare consumers may turn to the research literature. Reading medical papers, however, can be a challenging experience. To improve access to medical papers, we explore four features enabled by natural language processing: definitions of unfamiliar terms, in-situ plain language section summaries, a collection of key questions that guides readers to answering passages, and plain language summaries of those passages. We embody these features into a prototype system, Paper Plain. We evaluate Paper Plain, finding that participants who used the prototype system had an easier time reading research papers without a loss in paper comprehension compared to those who used a typical PDF reader. Altogether, the study results suggest that guiding readers to relevant passages and providing plain language summaries alongside the original paper content can make reading medical papers easier and give readers more confidence to approach these papers.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**.

Additional Key Words and Phrases: augmented reading; plain language summaries; healthcare consumers

ACM Reference Format:
Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.* 1, 1, Article 1 (January 2023), 39 pages. https://doi.org/10.1145/3589955

1 INTRODUCTION

A robust public health system depends on the timely dissemination of medical findings to those who need them. Most often, people stay apprised of medical findings through conversation with their doctors, printed materials like pamphlets, and online resources like MedlinePlus or hospital websites [33, 57, 113]. However, these resources do not cover all medical conditions and treatments [12, 92], especially those which are the focus of emerging research [23, 88]. In many cases, the latest medical knowledge appears solely in the medical research literature [35, 39, 84, 103, 115]. For healthcare consumers such as patients, their families, and other caregivers, staying apprised of the latest research may mean becoming familiar with the literature. In the words of one patient [5]:

Authors' addresses: Tal August, taugust@cs.washington.edu, University of Washington, Seattle, Washington, USA; Lucy Lu Wang, lucylw@uw.edu, University of Washington; Allen Institute for AI, Seattle, Washington, USA; Jonathan Bragg, jbragg@allenai.org, Allen Institute for AI, Seattle, Washington, USA; Marti A. Hearst, hearst@berkeley.edu, University of California, Berkeley, Berkeley, California, USA; Andrew Head, head@seas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Kyle Lo, kylel@allenai.org, Allen Institute for AI, Seattle, Washington, USA.

1073-0516/2023/1-ART1
https://doi.org/10.1145/3589955

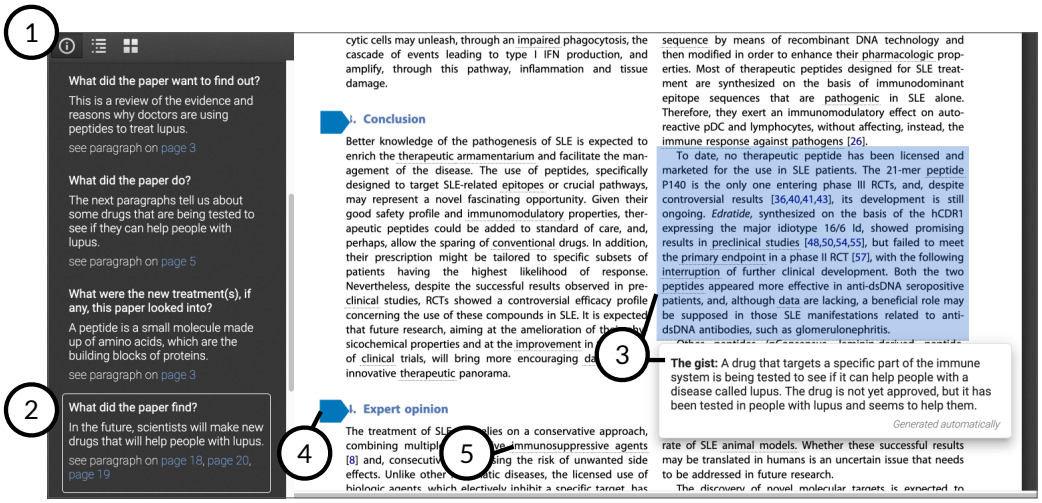

Fig. 1. Pictured is the Paper Plain user interface. When a reader opens a paper in Paper Plain, they see a side pane containing a reading guide (1), consisting of key questions the reader might ask of the paper, brief generated plain language answers, and pointers to passages in the paper where the reader 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 terminology by clicking underlined terms (5).

> *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.*

For patients like the one quoted above, research articles offer an awareness of cutting edge medical findings and the nuance of underlying studies. Patients need not fully understand articles to derive some useful information. From these articles, patients may find and share information germane to treatment options with their healthcare providers [35, 84, 115].

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 [35, 83]. This is in part due to being overwhelmed by the amount of unfamiliar terminology. It is also because healthcare consumers are unaccustomed to the norms of how research is conducted and how reports of it are structured [20, 35]. The result is that reading medical papers can be an experience that is challenging and at times demoralizing.

In this article, 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 [8, 47], hiding sections that are predicted to be irrelevant [17], and answering user-written questions [114]. 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 a paper's contents. Consider, for instance, this sentence from a paper about systemic lupus erythematosus, linked from a patient-facing MedlinePlus page [101]:

> *The most salient events include an impaired apoptosis of dying cells, a type I interferon (IFN) signature, the uncontrolled activation of T and B lymphocytes and the production of autoantibodies mainly directed against nucleic acids or ribonucleoproteins (RNP).*

This sentence is difficult not only because it contains technical terminology, but that *in combination* these words form a sentence so foreign that a reader has little chance of understanding it without learning a considerable amount of background knowledge from elsewhere. A reader not only needs to know what "autoantibodies" and "ribonucleoproteins" mean, but also how production of one implies the progression of their condition and risks to their health. A medical paper contains not one but hundreds of such sentences, making it exceedingly difficult for readers to find, let alone understand, information important to them. How can interactive interfaces make medical papers more approachable by incorporating plain language alongside original paper content?

This paper explores how future interactive aids can go beyond their typical capabilities to assist readers in understanding where to find information of interest in a paper according to the language they already know. We begin 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. An augmented reading interface for these readers will need to go beyond the capabilities of prior interfaces—that define terminology [47], provide summaries [45], or allow readers to ask questions of a paper [114]—and provide a reading experience that guides readers to useful information in the context of the paper.

To improve access to medical papers, we explore four features enabled by natural language processing (illustrated in Figure 1) and embody them in a novel interactive system, Paper Plain, through an iterative design process. First, 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. Second, when a reader clicks one of these questions, they are taken to a paragraph in the paper that answers the question along with an "answer gist," a plain language summary of that paragraph. Third, Paper Plain conveys the essence of terminology-dense passages with "section gists," in-situ plain language summaries available for each section of the paper. Finally, Paper Plain assists readers in understanding unfamiliar terms by allowing a reader to look up definitions by clicking the term. The key question index and gists are novel features in the context of reading applications for research papers; term definitions have appeared in prior reading systems, and are incorporated into Paper Plain as one of the components that make up a holistic reading support system. The design of the system is described in §4.

We envision Paper Plain as a system that can one day be enabled for any medical research paper. The system draws on active research in natural language processing for biomedical question answering [112], plain language generation [45], and term identification [82]. One limitation of current text generation capabilities is the risk of generating factually incorrect or inconsistent text, often referred to as "hallucinations." [75] Deploying any system in a medical context will require algorithmic advances and human oversight (e.g., crowdsourced fact-checking or expert review) to detect factually incorrect generations [58, 75]. For examples of current automated advances in this space, see [41, 62, 71]. In the context of this paper, we lightly curated generated text to ensure factuality and text coherence (more details in §5 and Appendix C). This allowed us to focus on developing interactions that would enable readers to meaningfully engage with medical research papers. §5 describes the implementation of Paper Plain and the manual curation of gists, while §8.3

discusses in more depth the limitations of text generation models for our application. While to date our implementation relies on some human curation, this project as a whole indicates the potential for reading experiences like Paper Plain to be deployed at scale over the scientific literature.

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 nor less accurately when they had access to Paper Plain.

The clear favorite feature was the key question index and answer gists. Participants also used and appreciated 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.

In summary, this paper contributes:

(1) A characterization of the barriers readers face when they read medical research papers. These findings both echo and extend findings from prior research about barriers to consuming medical information [35, 83, 97] by illustrating the barriers healthcare consumers face in medical papers, with important themes including readers' uncertainty about where to find relevant information in papers, and an overabundance of terminology (§3).
(2) Paper Plain, a reading interface for biomedical papers that brings together known affordances like term definition tooltips with the novel affordances of in-situ plain language summaries of paper sections and an index of key questions that guide readers to answering passages with paired plain language answers (§4).
(3) Evidence from a usability study that these new affordances helped readers quickly find passages in a paper that were informative to them. Participants using Paper Plain's key question index and answer gists, versus a typical PDF reader, reported a significantly easier time reading papers and greater confidence that they found all relevant information without a significant difference in correctness when answering questions about the paper (§7).

2 BACKGROUND AND RELATED WORK

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 [22, 33, 52, 102]. Kivits [57] 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. Cartright et al. [28] 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. Work has also studied how people search for health information online [33, 86], share through social media [30], and how online searching can lead to real-world healthcare utilization [111].

While the internet is a good source of consumer health information, it also poses challenges to searchers [97, 99]. One study found that the top search results may overapproximate the effectiveness for health interventions in comparison to the evidence in the literature [110]. Searchers might also experience information overload as they encounter unrelated search results, complex text, and contradictory guidance from multiple sources [12, 53, 97, 99]. Searchers cannot always overcome

these issues themselves, and instead may require consultation with their clinicians to make sense of the information they have found [97].

Whether it is found through web search or other means, medical literature plays an important role in providing specific, detailed, up-to-date information about health conditions and their treatments [115]. As such, there have been calls across the research community and advocacy groups alike to make literature accessible to health care consumers. In 2005, the NIH established an open access policy in part to encourage healthcare consumers to self-educate about their healthcare and related research, in consultation with their care teams [84]. Recent years have seen increasing recognition that public stakeholders, including advocacy groups and healthcare consumers, benefit from the use of primary medical research findings [35, 39]. Today, there is a movement in the medical community to involve patients more in the research process, including understanding lab reports [81], reviewing research papers [89] and leading research efforts [76, 80]. Research has shown the public benefit of this open access policy, with one such benefit being improved access to research findings for healthcare workers and consumers [103].

At the same time, medical research, and scientific research more broadly, present unique barriers to readers without research expertise [78]. Nunn and Pinfield [83] 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 [54]. Bromme and Goldman [21] 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. [35] 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. Britt et al. [20] 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. Our project illustrates how interactive reading interfaces can make medical research papers accessible to healthcare consumers through a novel interactive system, Paper Plain.

2.2 Interactive reading interfaces

Paper Plain draws inspiration from prior interactive reading systems that have used term definitions [47], question answering [29, 114], and guided reading [38]. Prior work has developed reading guides for students and researchers by constructing questions around a document. Inquire Biology [29] is a biology textbook augmented with 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 [38] introduced a new method for conducting reading groups by developing questions about a paper's methodology and findings to guide reading discussions. Zhao and Lee [114] 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 built tools for navigating concepts within a paper [8, 51] and providing reading guidance in textbooks [26, 109]. There are also interactive systems for collaborative reading of research papers, such as Fermat's Library [1], which provides community annotations on popular research papers, and Hypothes.is [2], which allows users to annotate and share annotations on any webpage.

Work has also imbued documents with summaries and definitions to assist in reading. In the context of reading research papers, Head et al. [47] 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. In the clinical context, UpToDate [4] provides expert-written summaries of current research for healthcare providers. Other work has explored tools for adaptive summarization [17] and evaluation of research literature [65, 73],

In contrast to previous reading interfaces for research papers that focus on clinicians, researchers, or students, this project 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 complex 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 [114], where to look for their answers [29] or know how to make sense of definitions of terms within a paper [47, 51] can make reading exceedingly difficult for our envisioned 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 are more likely to know. To do this, the system incorporates plain language alongside original paper content.

2.3 AI for scientific text processing

Paper Plain leverages advances in natural language processing (NLP) that have been developed to make medical information more understandable to the public, specifically healthcare consumers [36, 108]. The techniques most relevant to Paper Plain are automated term definition or replacement [105], plain language summarization [37], and consumer biomedical question answering [7]. Also relevant are writing tools to encourage plain language [44], as the underlying techniques for powering such systems are similar to those leveraged by Paper Plain (e.g., generating plain language). Paper Plain integrates these advancements in its implementation to show how such methods might support healthcare consumers in a user-facing interface and indicate the potential of scaling this reading experience across the scientific literature.

Work has introduced automated methods that define terms, simplify text, and answer biomedical questions. Veyseh et al. [105] presented a web-based system for acronym identification that works in the biomedical, scientific, and general domain and Murthy et al. [79] explored how to define scientific terminology with terms recognizable to a reader. Devaraj et al. [37] introduced a new dataset of healthcare consumer summaries for clinical topics along with and a trained model for simplifying medical text and Guo et al. [45] used plain language summaries to train a model for generating summaries of biomedical text. An alternative way of making medical language accessible to a broader public is by building question answering systems for healthcare consumers. Abacha and Demner-Fushman [7] collected a dataset of consumer health questions from NIH websites and developed methods for automated answering of these questions. Mrini et al. [77] introduced methods to improve answer recall for long and complex consumer medical questions. Other work has automatically classified the questions that healthcare consumers ask [91].

In the context of writing tools, Gero et al. [44] used generation models to help researchers author "Tweetorials," a threaded tweet meant to inform a general audience about a scientific concept on Twitter [19]. Other work has introduced writing tools to help journalists [55] or clinicians write using simpler terms [64, 87, 104], simplify text by replacing technical terminology with more common terms [14, 61, 85] and simplify e-prescription and medical instructions [27, 66].

Paper Plain draws on this active research to improve access to medical papers. §5 discusses in depth the adaptations needed to make this research provide useful output for healthcare consumers reading medical research papers.

3 OBSERVATIONS OF NON-EXPERT READERS

To gather more direct and comprehensive evidence of barriers for this population, we conducted a think-aloud reading study. Prior work on barriers has focused on consumer health information [97], scientific research in other domains [78], students [95], or searching through medical literature [35], but it is unclear how these barriers manifest for non-experts reading medical research papers.

3.1 Formative research

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.

Based on these interviews we designed four scenarios that 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. More details on these interviews and the scenarios are in Appendix A. Following the development of these scenarios, we recruited participants to walk through the scenarios in a think-aloud reading study.

3.2 Participants & recruiting

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] We discuss possible limitations to this recruiting strategy and the presence of a paid timed task in §8.4. A total of 12 participants completed the study (T1–12). Of these participants, 11 had completed college and 5 had completed professional or graduate school. 11 participants had taken 3 or fewer STEM courses since high school.

3.3 Procedure

Participants were randomly assigned into one of the four scenarios described in §3.1. Each scenario was assigned to the same number of participants. To ensure participants were equipped with some prior knowledge before approaching papers, they first read a consumer health webpage (MedlinePlus) about the medical condition in their scenario. This MedlinePlus step was meant to approximate realistic circumstances, in which a participant would receive information from their doctor about their diagnosis. After reading the MedlinePlus page, participants browsed a list of 11 research articles selected from PubMed articles linked from MedlinePlus. 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. While in real-world health information seeking, readers would undoubtedly come across irrelevant information [97], the study's focus

[1]This was above the federal minimum wage of US$7.25 and the state minimum wage of US$13.69 at the time of study.

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. Participants had enough time to read one or two papers (all were instructed to read at least one paper).

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 [15, 18] was performed on the think-aloud and questionnaire data. We followed Braun and Clarke [18]'s six phases of thematic analysis. One author familiarized 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.

3.4 Findings

Our study revealed a set of barriers readers face when reading medical research papers. Table 1 lists these barriers. Below we illustrate how these barriers manifested for non-experts reading medical papers and highlight concrete instances 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 only two participants that did not mention struggling with specific medical terminology (T4 & 9) said they instead skimmed over these terms or were able to infer them from context. 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 the usefulness of the treatments discussed in the article.

While terminology is a common barrier in scholarly communication [74], past interactions to address it present additional issues for our reading context. Past work has addressed this issue for researchers by providing definitions of terms based on earlier references in a paper [47]. However, there is no guarantee a reader in our context would understand definitions drawn from the original paper, considering that almost all text in medical papers has technical terminology. This issue suggests that a different approach to defining terminology for our envisioned readers is needed.

Barrier	Description	Quote	Readers
Unfamiliar terminology	Readers did not understand individual terms and symbols from the biomedical research domain.	*"What does this word mean?"*	T1–3, 5–8, 10–12
Overwhelmingly dense text	Readers had difficulty understanding passages that contained an overabundance of technical terminology.	*"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 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 find answers to but lacked knowledge 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 applied 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 1. Barriers readers encountered when reading medical research papers without prior experience.

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). 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)*

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 while readers do often wish to read original paper content, they might have little interest or capacity mastering the language of a particular paper, given that other papers might use different language, and that they may be pressed for time.

Not knowing what to read. While some participants read 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.

Of the 12 participants, 11 (T1–3, 5–12) had a difficult time knowing if a paper held relevant information and invested 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 they often later realized were irrelevant (T2–3, 5–8).

One such participant was T5, who reported struggling to read the entire first paper they selected because they wanted to do their due diligence to understand the results and decide if the paper was relevant to them. After getting to the discussion they realized that the section provided an accessible overview of the results. 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. (T5)*

As this quote shows, readers like T5 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 necessarily determine what the best passages were for them to read after the first paper (T2–3, 6–8).

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 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 similarly wanted 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. (T3)*

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 common 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 report study demographics or current best practices for treating an illness. Introductions would often include this 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 while trying to determine if each sentence was relevant to them.

Difficulty relating findings to personal circumstances. Some participants sought a sense of whether the findings of the paper were relevant to them personally (T2, 5, 8–9, 11). T2 and 8 wanted a better sense of 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 averages 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 information of this sort appeared in only one paper. While this personally relevant information typically does not appear in research papers, participants wished for this information nonetheless.

In summary, non-expert readers encounter a number of barriers getting oriented to and understanding biomedical research papers. Below, we discuss how novel reading interfaces might help non-expert readers overcome some of these barriers.

4 PAPER PLAIN: READING SUPPORT FOR MEDICAL RESEARCH PAPERS

We designed Paper Plain to make medical papers approachable to non-expert healthcare consumers. Unlike other systems in the augmented reading space for research papers, Paper Plain focuses on addressing the barriers of non-expert readers. To that end, Paper Plain integrates known features like term definitions with novel features like a key question index and answer gists.

Our design addresses four of the five barriers discussed in §3: unfamiliar terminology, overwhelmingly dense text, not knowing what to read, and difficulty finding answers. These were the most common barriers we observed in our formative research.

We followed an iterative design process to develop Paper Plain. Eight participants used two early prototypes of Paper Plain in preliminary usability evaluations. In our preliminary studies we observed participants double checking generated plain language (the gists) with the original text. When asked their reasons for doing so, participants mentioned generated text being vague or wanting to confirm information with the original paper. NLP systems are imperfect (e.g., by generating inconsistent information [75]) and these observations highlighted the risk of relying solely on generated content. Because of this, in Paper Plain's design all gists were placed as close to the original text as possible without overlapping, and gist content was provided on-demand, rather than initially displayed along with the paper, to encourage readers to focus on the paper and only pull from the gists for supplemental information. We discuss future designs to encourage reading original text in §8.2. The iterative design is described in more detail in Appendix B.

Paper Plain provides four main features:

(1) **Term definitions** – Tooltips contain definitions of biomedical terminology.
(2) **Section gists** – In-situ plain language summaries of sections' contents.
(3) **Key question index** – A sidebar listing likely questions a reader might have, with links into passages of the paper that answer them.
(4) **Answer gists** – Plain language summaries of the answering passages.

To illustrate the features of Paper Plain, we describe how a fictional reader, Sarah, leverages Paper Plain to learn about new treatment options from a research paper.

Sarah is a 25 year old woman (pronouns: she/her) who was recently diagnosed with Systemic Lupus Erythematosus (SLE, also called Lupus). When Sarah discusses treatment options with her doctor, she wonders if there are treatments the doctor does not mention that might benefit Sarah. In the evening after work, she looks for research papers to learn about emerging 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." [101]

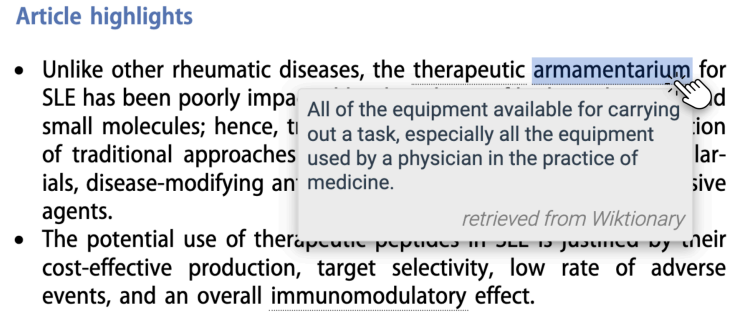

Fig. 2. Biomedical terms with definitions are underlined ("armamentarium", "immunomodulatory"). Clicking the term opens a tooltip with a definition and a reference where the definition was sourced from.

After reading the title, Sarah wonders – *what is the paper about? What are therapeutic peptides? Are they a possible new treatment for SLE?* – and begins reading.

Term definitions help Sarah resolve technical terminology. Paper Plain provides definitions for unfamiliar terms in the context of the paper so Sarah can integrate new concepts into her reading. While reading the introduction, Sarah reads a passage full of technical terminology (Figure 2). She does not know 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 [6] explaining that "armamentarium" is a certain kind of medical equipment. Sarah continues reading, using the tooltips to resolve unfamiliar terms.

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 reading 15 pages of dense text. From the introduction, Sarah had gathered that not every peptide has been equally effective as treatment, and each might be used in different circumstances, so she would prefer to only read in depth about the most promising peptides relevant to her mild case of SLE.

Paper Plain helps Sarah determine what sections are worth reading by providing in-situ plain language summaries, or "section gists." Sarah clicks on a tab indicator next to the section title, and a gist appears above the section text (Figure 3). The gist contains simple language: 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." As she reads the rest of the paper, Sarah refers to the section gists to develop a surface-level understanding of the peptide sections.

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. In

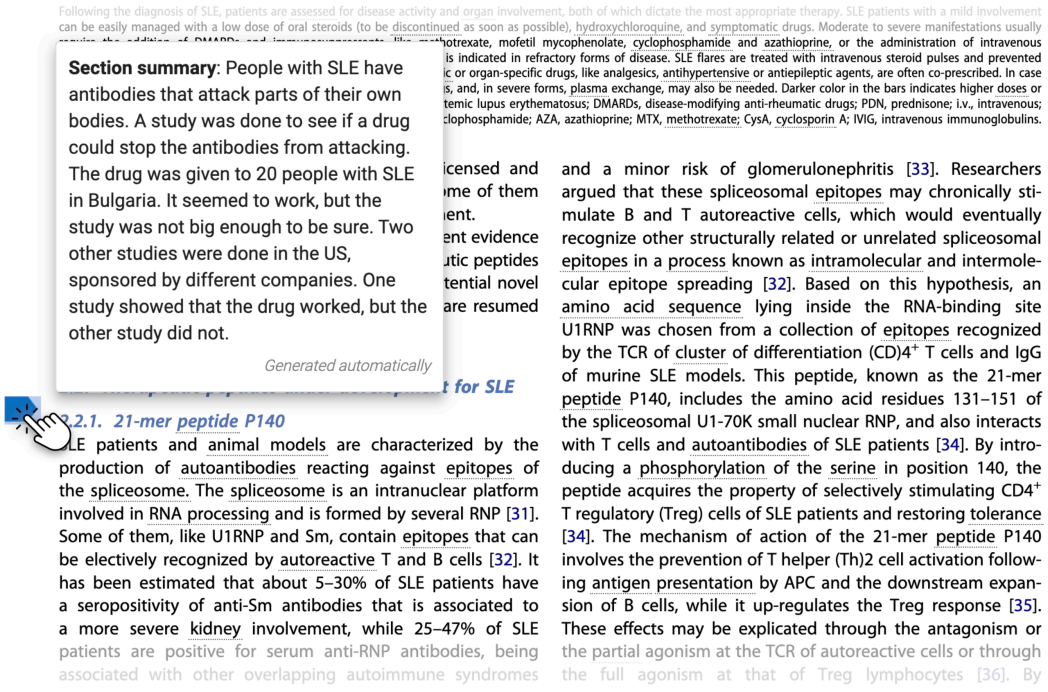

Fig. 3. A section gist for a example passage with dense text. Clicking on a tab indicator next to a section title displays a plain language summary of the section.

addition, for future papers Sarah would like a quick way of gathering the most relevant information for her first, without needing to scan the entire paper.

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?" Each question is accompanied by a one-to-two sentence plain language answer preview and hyperlinks to one or more paragraphs in the paper. Sarah sees that the question "What did the paper find?" hyperlinks to passages within the Discussion (see (1) Figure 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 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 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 terminology. 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.

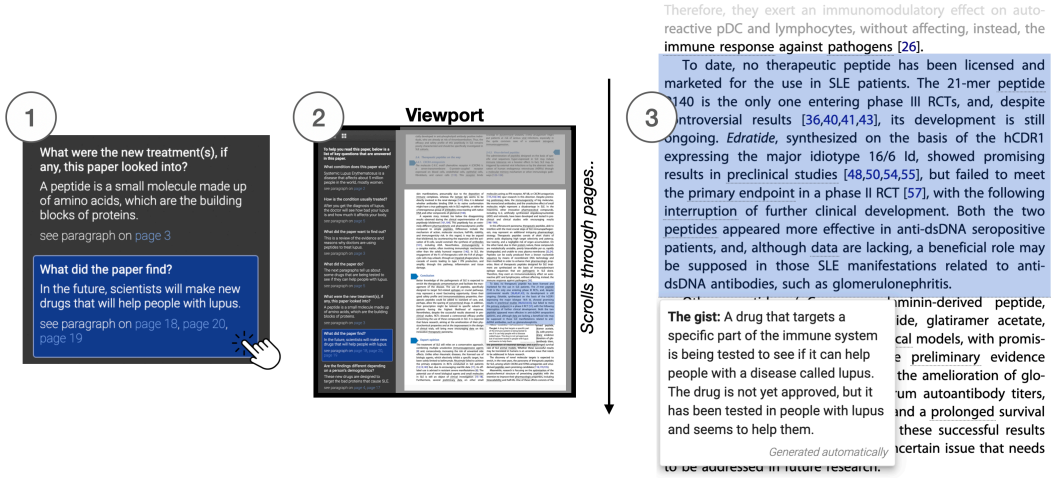

Fig. 4. The key question index guides readers to answering passages and their answer gists. When one of the questions is clicked (1), the interface will scroll (2) to the first answering passage and display a tooltip containing the answer gist. In (3), we show the simplified answer gist alongside the original paper.

The key questions remind Sarah of questions she might want to ask about a paper. Because the number of questions is small, most can be viewed without having to scroll (see (1) in Figure 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 has confirmed and deepened her understanding of the paper's limitations.

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.

5 IMPLEMENTATION

Paper Plain provides an augmented reading experience by applying NLP techniques for biomedical question answering and plain language summarization. Below we discuss how we incorporated such techniques into our prototype of Paper Plain. In §8.3 we describe how such techniques will need to be further developed to responsibly deploy tools like Paper Plain.

5.1 Term definitions

Paper Plain uses named entity recognition (NER) models to identify medical terms and entity linking (EL) models to resolve those terms against external knowledge bases containing term definitions. In our implementation, we use `scispaCy`'s [82] NER module to identify terms. We then link those terms to the Unified Medical Language System (UMLS) [16] using `scispaCy`'s EL module, and to Wiktionary [6] using string matching heuristics. For terms linked to both databases, we prioritize the definition from Wiktionary. The extraction and matching process leads to many terms for which a reader would likely not wish to see definitions because they are so well known outside of the medical literature (e.g., terms like 'expert' or 'negative'). We filter out such terms by excluding them if they are sufficiently common in general text corpora. For both Wiktionary

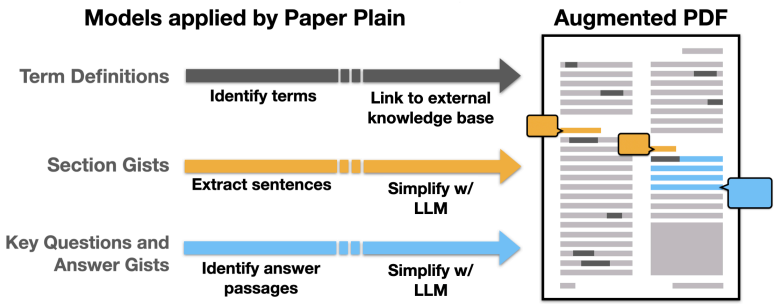

Fig. 5. Paper Plain uses machine learning models to look up term definitions, and to generate section gists and answer gists, for an interactive PDF reading experience.

and UMLS, we preserve the bottom 20% of terms based on word frequency. We use the python package `wordfeq` to identify term frequency based on the Exquisite Corpus [98]. We also remove all terms consisting of 30 or more characters because terms over 30 characters were usually ill-formed (e.g., 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.'

5.2 Section gists

Paper Plain generates section gists for the lowest-level subsections in a paper using large language models (LLMs). In our implementation, we concatenate the first sentence of every paragraph in a section and generate a plain language summary of it using GPT-3 [24].[2] GPT-3 is a pretrained large language model released by OpenAI that has obtained state-of-the-art results on many language tasks using different prompts for generation [24] and is increasingly used for many text generation tasks. Sentences were extracted manually for our prototype system, but could be automatically extracted using automated PDF parsing software [69, 96]. Using the leading sentence of each paragraph is a common competitive baseline for summarization [40]; we opt for this strategy rather than inputting the full section text because during our tests GPT-3 was prone to copying the text verbatim when given the full section. We engage in prompt engineering, a common practice for achieving fluent text for large generative models [68], to encourage fluent and specific plain language summaries. We use a GPT-3 prompt adapted from a preset example that OpenAI provides for simplifying and summarizing text for a second-grade student,[3] and we modify it to tailor texts for a fifth-grade student. We also tested later grades, up to college, but found that the generated text using the fifth-grade level prompt used plain language of the level we desired while still providing major details about the section. More details about the GPT-3 prompt appear in Appendix C.

Because of the risk of hallucinations (i.e., factual inaccuracies) in generated text, for our studies we curated gists. If the gist contained clear hallucinations (e.g., if it incorrectly referred to a peptide as a surgical procedure), or contained nonsense text (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 used that gist. Typically, it only took

[2]The best available model at the time was `text-davinci-002`, which we queried between Aug.–Sept. 2021.
[3]https://beta.openai.com/examples/default-summarize

1–2 generations to arrive at a valid gist; more details can be found in Appendix C. We discuss the risk of hallucinations and a vision of responsible technology development in §8.3.

5.3 Key question index and answer gists

Paper Plain requires the following: a predefined set of questions to form the key question index, a question answering (QA) model to extract relevant passages from the paper for each question, and an LLM to simplify the answer. In our implementation, we use questions sourced from the PICO framework [90] for clinical questions and Cochrane's guide on writing plain language summaries [3]. Both sources focus on information in medical papers that are relevant to patients and caregivers. We curate 8 questions from the two sources; these are listed in Table 5 in the appendix. For each question, we extract relevant passages from the paper using Yoon et al. [112]'s extractive QA system trained to answer questions using biomedical research papers. We follow prior work on making QA models more robust by including semantically-equivalent variations of questions [43] (e.g., what did this paper find, what are the main results of this paper?). This QA model extracts single words or phrases that answer a question rather than full passages. If the model identifies words or phrases in a passage as answering a given question, we mark that passage as an "answering passage." For our prototype system, we manually mark sentence boundaries of answering passages; such a step could be automated with tools such as [96]. Finally, we create answer gists by simplifying the extracted passages using GPT-3 [24] with the same prompt and curation we use for simplifying section gists. The sidebar shows the first 1-2 sentences of the first answer gist for each question.

6 USABILITY STUDY

We ran a partial within-subjects usability study to assess how Paper Plain's features affected non-experts' experience reading medical papers.

We were interested in the following research questions:

RQ1-How do participants use Paper Plain's features while reading a medical paper?
RQ2-How does Paper Plain affect participants' self-reported reading difficulty, understanding, and ability to identify relevant information?
RQ3 - Is there a difference in paper comprehension when using Paper Plain?

6.1 Method

6.1.1 Participants. We recruited participants from Upwork using the same recruiting materials as §3.2. 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. All participants were paid US$15 for the hour-long, remote 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 ($\mu = 35.04$, $\sigma = 13.47$). All participants had completed college, and a third had completed professional or graduate school. 19 participants (79%) had taken 3 or fewer STEM course since high school and 22 (92%) had never been involved in publishing a research paper. No participants had professional medical experience.

6.1.2 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 §3.3. For each scenario, we selected a single paper to read ([101] for SLE and [10] for a herniated disc) from the most common papers from our reading observations in §3.

Each participant underwent the following study procedure once for each scenario. First, participants read a description of the scenario, then a MedlinePlus page about the diagnosis, then

the associated research 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 the features of Paper Plain available to them (described in §6.1.4) then read the paper for 10 minutes. They were told when they had 5 minutes remaining and when they had 1 minute remaining. After each paper, participants filled out questions about the paper (§6.1.3). The duration of the reading task was set to 10 minutes following our observations from the formative study (§3) and pilot studies that this was the typical amount of time participants spent completing an initial read of a paper. At the conclusion of the study, participants completed a questionnaire where they reported their demographics, education, and research experience. Then, participants reported their experience using Paper Plain, identifying what features they found most helpful, in a questionnaire and brief interview. A researcher was present for the entire study.

6.1.3 Measures. We collected measures to assess feature usage (**RQ1**), self-reported reading experience (**RQ2**) and comprehension (**RQ3**), as described below:

Feature usage. To measure how participants used Paper Plain's features (**RQ1**) we logged all telemetry data on interactions with Paper Plain. We measured the frequency of usage of each feature, as well as the number of participants who used or spoke about a feature.

Self-reported reading experience. We collected self-report data to understand how participants felt about the support Paper Plain provided. Participants answered the following questions after each reading task on a 5-point Likert-style scale (1="Not at all," 5="Very"):

(1) "How hard did you have to work to read the paper?"
(2) "How much do you feel like you understood the paper?"
(3) "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 intended to assess understanding of the paper content without biasing in favor of Paper Plain; therefore, questions were selected that could not be answered directly from the answer gists or key question sidebar. Table 2 shows example comprehension questions and passages of the paper that contained answers to those questions.

We wrote 15–20 questions for each paper and asked two practicing physicians not involved in the study to 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 of the most interest to patients. We revised the wording on any questions or answers that were unclear following clinician feedback and two pilot studies. Ultimately, we selected 14 multiple choice questions, 7 for each paper. We measured comprehension as the proportion of questions answered correctly.

6.1.4 Interface variants. To understand the impact of Paper Plain's novel guidance-offering features on readers' experience engaging with medical papers, we evaluated variants of Paper Plain with and without these features. 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.

Question	Correct 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 2. Examples of multiple choice questions and answers from the usability study.

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.

6.1.5 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 a separate mixed-effects linear model [67] for each measurement. Paper type and system variant were fixed effects in the model and participant was a random effect. 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 D.

We note that usability studies with reading interfaces often fail to reveal significant differences in how readers answer comprehension questions with and without experimental interfaces (see, for instance, recent studies by Head et al. [47] and Badam et al. [9]). A lack of significant difference can be attributed to several reasons: there could be similar comprehension across conditions, or the instrument might not measure 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. In our context, plain language

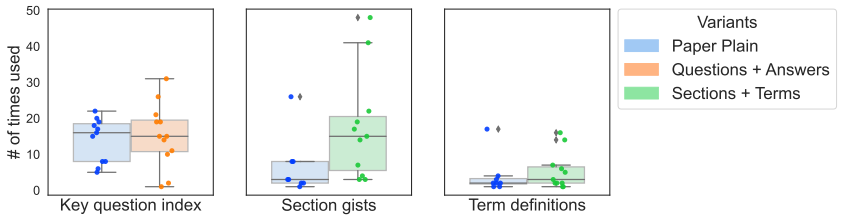

Fig. 6. Usage of features across interface variants. Each pair of side-by-side boxplots considers usage of a single feature (e.g., key question index, section gists, term definitions). Each point represents the number of times a feature was used by an individual reader. Colors represent interface variants. Overall, when a feature was available to a reader, it was used several times by that reader. Usage of section gists and term definitions was much higher when they were the only features available; they were used much less when the key question index was also available.

can overly-simplify scientific findings, and might risk leading readers to misunderstanding the material [93, 100]. In this case, a decrease in comprehension would be undesirable.

Therefore, we also conducted a non-inferiority test [107] to confirm that Paper Plain did not detract from paper comprehension. Non-inferiority tests evaluate that the experimental condition is no worse than the control (i.e., the null hypothesis is that an experimental condition is significantly worse than the control). They have been used in psychotherapy research to assess, for example, the effect of remote versus in-person interventions [63, 70, 106]. Non-inferiority tests are conducted similarly to traditional hypothesis testing, but the test evaluates if the difference between an experimental and control condition is significantly larger than an equivalence margin δ.

In our study, we set $\delta = 1$, such 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 [94] 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 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.

7 RESULTS

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

7.1 How did participants use Paper Plain's features?

Participants typically interacted with all the features of Paper Plain available to them. When participants had access to only the key question index and answer gists (Questions and Answers), they 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 in the "Sections and Terms" variant clicked on at least one section gist and term definition. On average, they clicked on 18 section gists and 5 term definitions.

When participants had access to all features they often opted for the key question index and answer gists. Rather than the section gists and term definitions, participants with access to the key questions and answer gists clicked an average of 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

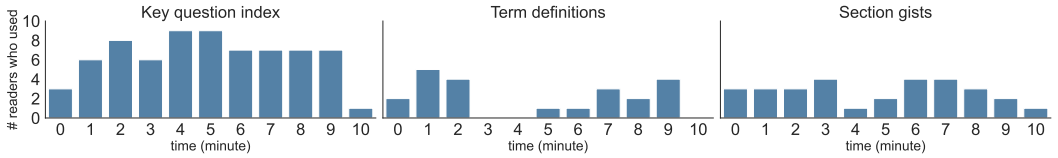

Fig. 7. Minute-by-minute usage of Paper Plain's features during the ten-minute paper-reading task. Each bar corresponds to the number of participants who used a feature in each minute of the reading task. All features were used throughout the whole reading task, and not just at the beginning or end. The key question index and answer gists saw consistent usage by a large proportion of participants. Usage is shown only for the condition where all features were available (N=12).

that did engage with these latter features also used them much less, clicking on average only 7 section gists and 4 term definitions. Figure 6 plots the usage of each feature for Paper Plain and illustrates this tendency 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. While the key question index listed only 8 questions in each condition, on average participants clicked on questions more than 10 times ($\sigma = 7.48$) 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.

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 7, which shows the minute-by-minute usage of the features over the course of reading task for readers in the Paper Plain condition. Readers 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 6). 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 after this initial phase, and led to sustained interaction with the features for the remainder of the task time.

We observed differences in reading behavior 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 §3, spent substantial time in dense sections with limited important information (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 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 more important but did not have adequate time to read them.

All participants with Paper Plain reached the end of the paper; Paper Plain's features supported participants in doing so in different ways. Participants reported that the section gists and term definitions helped them read through dense text (P1, 3–5, 7, 15, 18), while the key question index and answer gists allowed them to quickly navigate the paper (P2, 4, 7–10, 13, 18–20).

Participants with the section gists and term definitions reported that they 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 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, as a guide for understanding the complex text (P5, 7). This usage aligns with our design goal for the section gists.

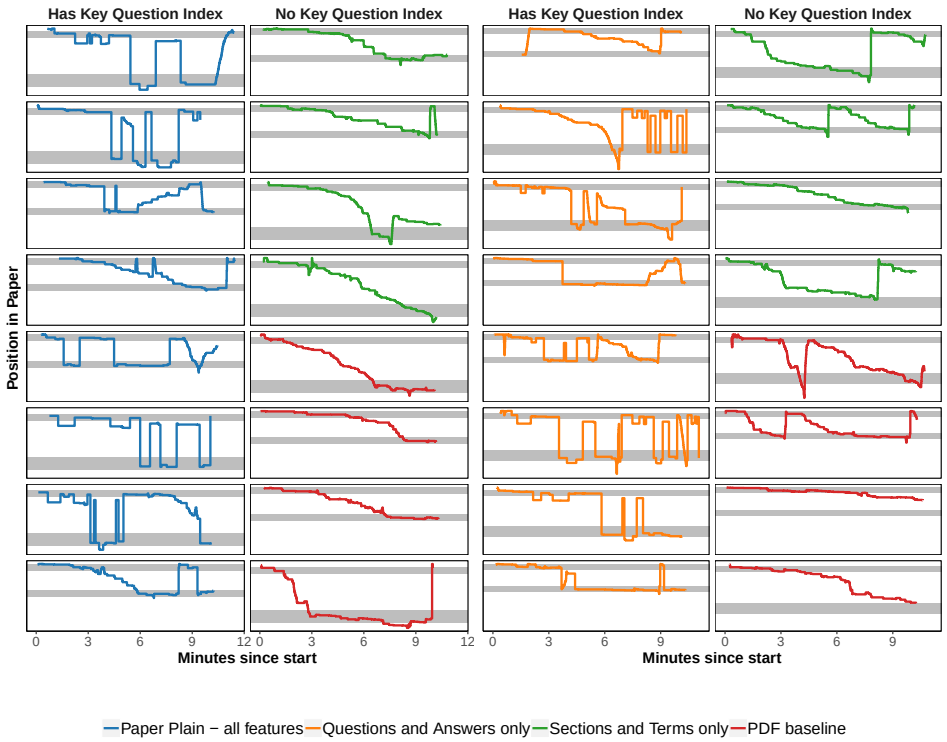

Fig. 8. Scrolling behavior with and without the key question index. Each plot shows the participant's scroll position in the paper over time, with the top of the plot corresponding to the beginning of the paper. Plots for the same participant are grouped side-by-side (e.g., the two left-most plots in the top row correspond to participant P1). Grey shaded regions correspond to the introduction, discussion, and conclusion sections of the two papers. Participants with the key question index jumped frequently from one part of the paper to the next; those without often read the paper linearly from start to end. Participants with the key question index spent more time in the introduction, discussion, and conclusion sections of the papers than those without.

Participants used the key question index to seek text that was relevant to them by jumping right to that information (P2, 4, 7–10, 13, 18–20). P10, for example, read through the abstract and introduction of a paper, then opted for using the key questions to jump through different sections of the paper. The key question index seemed to support a nonlinear reading strategy. Participants with the index (in any condition) jumped back and forth in a paper (Figure 7). Participants without the key question index often read papers top to bottom, once through (Figure 8).

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 the linear reading behavior of those without the feature. See Figure 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

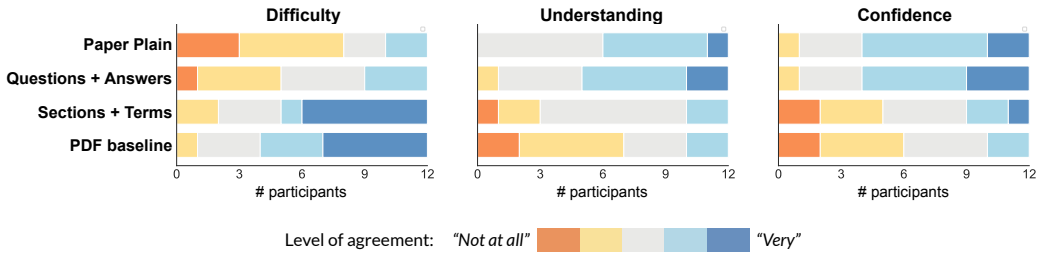

Fig. 9. Self-reported difficulty, confidence, and understanding of papers by interface. Participants reported that it was less difficult, that they felt they better understood the paper, and that they felt more confident they found all of the relevant information, while using Paper Plain with all features.

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 7.

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) (σ = 3 minutes and 44.60 seconds), compared to 6 minutes and 48.99 seconds (σ = 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.

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

Figure 9 shows an overview of participants' self-reported scores for reading difficulty, understanding, and relevance for all papers and interface variants. Our mixed-effects model F-test found a significant difference in scores across conditions ($p < 0.001$ for all three measurements following Holm-Bonferroni [49] correction). Fixed-effect coefficients appear in Appendix D. Here we discuss our interpretation of results in this section. We report medians (denoted $\tilde{x}$) for each subjective rating given the non-normal nature of Likert scale data.

The key differences were as follows (see Table 3 for all differences and significance values between pairs of interface variants). Participants with Paper Plain were significantly more confident that they found all relevant information from the papers ($\tilde{x}$ = 4.00, σ = 0.87, with 5.00 corresponding to maximum confidence) compared to the basic PDF reader ($\tilde{x}$ = 2.50, σ = 1.00). They also reported they better understood the papers ($\tilde{x}$ = 3.50, σ = 0.69 vs. $\tilde{x}$ = 2.00, σ = 1.00), and that reading was significantly less difficult ($\tilde{x}$ = 2.00, σ = 1.06 vs. $\tilde{x}$ = 4.00, σ = 1.04).

Among the features, the key question index and answer gists appeared particularly useful in reducing self-reported difficulty. Readers who experienced only the key question index and answer gists rated their reading difficulty significantly lower ($\tilde{x}$ = 3.00, σ = 0.97) than participants with

	$PP - QA$	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 3. Post-hoc (two-sided) tests for pairwise differences in fixed-effects estimates comparing interfaces. Columns show differences in fixed-effects estimates between interface variants and Holm-Bonferroni-corrected p-values [49] under the mixed-effects model. Differences are shown for pairs of interfaces including Paper Plain (*PP*), key question index and answer gists (*QA*), section gists and term definitions (*SD*), and a plain PDF reader baseline (*PDF*). For example, the cell within column "$PP - PDF$" and row "Reading Difficulty" should be interpreted to indicate that Paper Plain is associated with 1.983 points lower reading difficulty on a 5-point scale versus a PDF baseline. Statistically significant p-values are bolded. Details about this analysis appear in this section and Appendix D.

the baseline PDF reader ($\tilde{x} = 4.00$, $\sigma = 1.04$), an effect that was not observed for participants who only had access to section gists and term definitions. Participants with the key question index and answer gists also reported higher confidence ($\tilde{x} = 4.00$, $\sigma = 0.94$) and greater understanding ($\tilde{x} = 4.00$, $\sigma = 0.89$) compared to the PDF baseline ($\tilde{x} = 2.50$, $\sigma = 1.00$ $\tilde{x} = 2.00$, $\sigma = 1.00$).

From our observations, it appeared that each of Paper Plain's features played some role in making reading feel less difficult. Section gists and term definitions, for instance, seemed to help many participants seek assistive information without switching contexts (P2, 6, 7, 11, 16–17, 19). P19 found the term definitions useful to understand the medications the paper mentioned. P2 found the section gists were helpful to understand the paper text in more familiar language. P17 described that 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."

The key question index and answer gists also appeared to help participants review papers more quickly and easily (P2–3, 9–11, 20). In the words of P9, these features were useful because "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 also shared that the simplified answers helped them understand the overall story of the paper quickly, so they had more time to delve into its details. P3 elaborated that these features 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." In this way, for many participants it seemed that the key question index and answer gists helped them develop a general understanding of a paper early on and focus their reading efforts.

The key question index seemed to be a favorite feature: 18 of 20 readers who experienced the key question index in at least one condition selected the index as the most helpful feature in the final study questionnaire. Participants appreciated how the question index helped them quickly find and understand relevant information in the paper (P2, 4, 7-10, 13, 18-20). In P7's words, the question index "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." In summary,

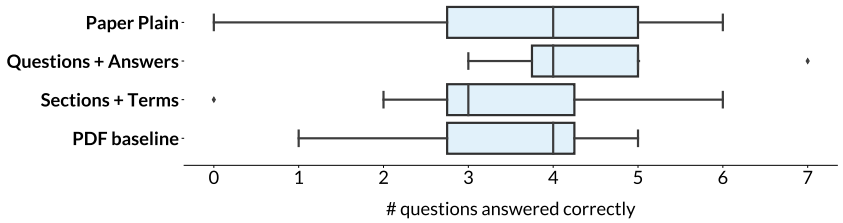

Fig. 10. The number of comprehension questions participants answered correctly, grouped by interface used.

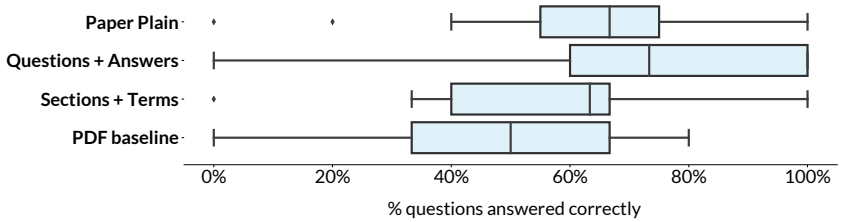

Fig. 11. For a subset of questions that were highlighted through interaction with the key question index, scores on multiple choice questions appeared to improve when readers had access to the key question index. This effect was more pronounced when participants only had access to the key question index and answer gists, suggesting that other features might have distracted from answering these multiple choice questions.

Paper Plain reduced self-reported difficulty, and increased self-reported confidence and sense of understanding, versus a plain PDF reader baseline, with evidence supporting the particular role of the key question index in these changes in readers' experiences.

7.3 Is there a difference in comprehension when using Paper Plain?

Across all conditions, participants on average answered 3.73 ($\sigma = 1.51$) of 7 multiple choice questions correctly. There was no significant effect of 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 10 compares the scores of participants on the multiple choice instrument, grouped by interface variant.

Post-hoc analysis suggests that some multiple choice questions were answered correctly more often with the key question index than without. While it was not possible to find the correct answers to questions by consulting the key question sidebar alone (see §6.1.3), some questions were answerable by reading passages highlighted by clicking a question in the question index (e.g., the first and third questions in Table 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 is shown in Figure 11. This trend is not statistically significant (paired samples t-test $t_{26} = 1.89, p = .07$); however, it does suggest the possibility that features of Paper Plain may affect how readers understand different parts of a paper depending on how they interact with the features.

8 DISCUSSION & FUTURE WORK

This paper explores how interactive information interfaces can make research papers approachable to healthcare consumers that need it. Below, we take stock of our findings, while discussing them amidst their limitations and potential to guide future reading tool development.

8.1 Summary of results

Our formative research suggests that non-expert readers, although motivated, face obstacles to reading medical literature, including overwhelmingly dense text, not knowing what to read, and difficulty finding answers to one's questions. Our evaluation of Paper Plain provided the following answers to our research questions about its effect on reading:

RQ1. How did participants use Paper Plain's features? Participants used, and found useful, each of Paper Plain's features. These features were used throughout the reading task from beginning to end. Participants used the section gists as aids for reading dense passages and used the key question index to quickly find text that was relevant to them. The key question index and answer gists were a clear favorite feature. When participants had access to all features, the key question index and answer gists were used more often than the section gists and term definitions.

RQ2. How does Paper Plain affect participants' self-reported reading difficulty, understanding, and ability to identify relevant information? Participants who used Paper Plain reported significantly lower difficulty reading, higher understanding, and higher confidence they found all information relevant to them, than those who used the baseline PDF reader. During the study sessions, participants found that their reading was facilitated by the key question index, which they believed offered an approachable overview to the paper, and with the term definitions and section gists, which helped them understand difficult passages of text.

RQ3. Is there a difference in comprehension when using Paper Plain? Accuracy answering multiple choice questions was similar for those reading with Paper Plain and with the baseline PDF reader—according to our tests, neither significantly superior or inferior. A (statistically insignificant) trend was observed where participants with the key question index answered questions correctly more often if answers to those questions were highlighted through interaction with the question index.

We note a discrepancy between the significant increase in self-reported understanding, and the absence of difference in multiple choice accuracy. One potential reason for this discrepancy could be that the two measures corresponded to different phenomena: the multiple choice questions tended to assess very specific facts from the paper (e.g., one question for the literature study paper was the inclusion criteria for candidate studies), and the subjective rating related more to one's sense of overall paper understanding. We offer the conservative interpretation that participants felt that Paper Plain helped them better understand papers as a whole, without noticeably improving or degrading their ability to answer specific questions about the paper.

8.2 Design implications

Based on this research, we offer the following guidance for future designers of related systems:

Introduce reading guidance. We believe interactive reading systems can provide more active support for guiding non-expert readers. Experts already employ strategies to gather relevant information in a paper without engaging in a deep read (e.g., skimming) [95]. In contrast, readers in our formative studies lacked strategies for reading papers, defaulting to an exhaustive linear reading. This led to readers to spend time on passages with little relevance or importance.

Incorporating features like the key question index may be able to help non-expert readers who lack fitting reading strategies. In our usability study, such an index helped readers jump to relevant sections of the paper within the first few minutes of reading. This feature was also a favorite feature

of participants. We note that there is a risk of distraction: in our pilot studies of the tool, the index distracted readers those who had their own approaches to reading papers; that said, our final design may have struck a good balance by making the index toggleable.

Incorporate plain language into the original document. Gists were frequently used by non-expert readers in the usability study. Every participant who had access to plain language features (either answer or section gists), used them during the reading task. We propose that plain language should be incorporated in a way that it serves to help understand the original document, rather than replacing it. The reality of contemporary generative models is that they often produce inconsistencies and inaccuracies [75]. Given the risk of misinforming readers, Paper Plain takes the approach of attempting to help readers focus attention on the original paper text in several ways: readers request gist content rather having it displayed by default, gists are shown directly alongside (but not occluding) paper content, and readers are made aware that gists are generated.

8.3 Ethical and Social Implications

While it is clearly of benefit to help healthcare consumers become informed about their care, systems like Paper Plain could bring about undesired consequences. First, health information can be dangerous if not understood well. A non-expert reader may be unaccustomed to important norms in the scientific process, like how a single paper does not represent scientific consensus. As such, a reader could mistake findings or interpretations in a paper as truth, which could lead them to making misinformed decisions about their care. While we note that readers are already taking such risks by turning to medical research papers [35], Paper Plain could lead to these risks being experienced for more readers, and more papers.

Furthermore, because Paper Plain incorporates generated text, it faces all of the limitations that come with contemporary text generators. Most worrying to the healthcare context is the tendency of text generators to hallucinate factually inconsistent or incorrect information [75]. On the one hand, there is reason to have optimism that accuracy will get better with time: the field of natural language generation is moving steadily towards more accountable output through measuring and encouraging factuality (e.g., by setting logical constraints generations must satisfy [71]). And well-design human-in-the-loop systems might be capable of repairing generated output by letting people to regenerate gists on command, report vague or hallucinated content, and leave annotations for future readers alerting them to possible hallucinations (e.g., using social annotation tools like Hypothes.is [2]). Such feedback simultaneously could improve models for later use while encouraging readers to play a role in evaluating the information they are accessing.

That said, as long as there is inaccuracy, the risks of hallucination are serious. They could lead a patient to make treatment decisions based on a misunderstanding, or, should a gist be overly optimistic about a treatment, lead to a loss of hope when one realizes the realities. In light of these risks, we suggest that for a system like Paper Plain to be deployed responsibly, it should augment existing sources for seeking reliable healthcare information, and clearly communicate its limitations. Healthcare consumers access information from many sources, including consumer-facing websites [102, 113], online communities [52], and research papers [11]. Paper Plain should not replace these sources, but rather fit carefully in among them.

Looking forward, we suggest that systems like Paper Plain should play a circumscribed role: they should help people find information to share with their clinician, provide a place where clinicians can direct patients to recent research, and support patient communities in developing preliminary understandings about the landscape of contemporary research about their conditions. In these settings, a tool like Paper Plain would be one component of a healthcare consumer's information diet. And in any of these settings, it would be important that the tool provides ample

messaging conveying what content is generated, perhaps with indicators conveying likely factuality (see [42, 71]), while making clear that readers should discuss their findings with healthcare providers.

8.4 Limitations

Our findings are limited in their generalizability in several regards. First, recruiting participants on Upwork might have have skewed our findings about barriers and resulting design given that participants were not reading medical papers that were personally relevant to them. These participants may have paid less attention to particular details of the paper or experienced negative findings or unclear results differently. To mitigate this limitation, we designed the tasks in the studies to closely resemble those of healthcare consumers we had interviewed who had experience reading the literature; however, such task design can only go so far.

The findings of the usability study are in part limited by the timed, abbreviated nature of the reading task. Participants may have scored differently on multiple choice questions, had a different subjective experience, and used the interface differently, if more time was given. One indicator of the influence of time on participants' experience is that some participants reported that if they had more time, they would have read the paper through again or looked for more information. For some participants, the time limit was reported to make them more anxious and to affect their ability to remember information. The time constraints could have increased participants' dependence on the key question index and answer gists, given their economy in helping a reader attain an overview of the paper, and reduced the usage of section gists given that participants may have relied on them more heavily if they were reading outside of the answer passages. This limitation should be mitigated with future studies that relax time constraints.

An additional limitation is that the participants in our study were predominantly college-educated. Our findings may not represent the usability of Paper Plain for non-college-educated readers. We note that the population of college-educated adults seeking additional medical information is significant in its own right, if unfortunately incomplete. It is important to develop and evaluate resources for those without a college education, who are often among those most marginalized by the medical system and lack access to the medical literature. Future work should focus on the ways that barriers manifest in these groups and how to make tools like Paper Plain valuable and accessible to those who are marginalized in the healthcare system.

8.5 Future directions

Our system and findings are suggestive of several interesting areas for future work.

Intelligent reading interfaces. As AI technology improves, new interfaces integrating this technology can provide tremendous value to users. This paper suggests one such kind of interface that incorporates techniques like biomedical question answering (QA) [112] and plain language summarization [24]. Other NLP techniques like machine translation [56], toxic language detection [50, 72], and news story mapping [60] might similarly enable new kinds of reading interactions.

Supporting paper comprehension. Our results suggest an effect of the interface on comprehension that is neither superior or inferior. What would it take to design reading interfaces that demonstrably improve comprehension? One challenge in designing and evaluating interfaces with this purpose is that simplifying scientific information risks over-inflating readers' sense of understanding and reduce their reliance on experts [93]. This risk needs to be kept in mind. Furthermore, one tack that may prove useful is to focus on discouraging common misunderstandings for healthcare consumers reading medical literature (see [35]) by helping readers steer clear of predatory journals without peer review and seeking findings corroborated by multiple papers.

Addressing other barriers for healthcare consumers. Extensions of Paper Plain could help readers with the barrier from §3 that we have not yet addressed—namely, relating findings to a reader's

personal circumstances. Participants in our formative studies expressed interest in patient testimonials that related to treatments in the paper, and wanting to know how patients similar to the reader responded to treatments. Future interfaces like Paper Plain could address this barrier.

Helping healthcare providers and patient advocates. Could an interface like Paper Plain benefit other stakeholders in medical research beyond patients and caregivers? Healthcare providers and patient advocates read medical research papers to apply their findings to clinical practice [25, 46, 88]. The needs and barriers faced by these groups differ from healthcare consumers, and would likely require distinct efforts to address. To give one example, providers may need to review a greater volume of research papers, relating to a broader set of patients' circumstances. Perhaps interfaces like Paper Plain could be extended to support review of collections of papers, for instance by extracting and summarizing answers for key questions across papers.

Supporting non-expert readers in other domains. Medical research is one of many contexts where non-expert readers read highly technical documents. Paper Plain's design can inspire efforts in addressing related barriers in these other contexts. Some aspects of these contexts merit new design efforts. For example, the questions that would appear in a key question index for a legal contract or privacy statement would be different than those for a medical paper. Other kinds of documents, like software tutorials, may require reading in a particular order to be sensible, to the extent that a novel indexing feature may confuse readers. In such cases, a key question index would need to be aligned to the document's original structure. We anticipate that in-situ section gists and term definitions could be similarly helpful for reading documents in many other domains, should they be appropriately tailored to the terminology familiar to the envisioned readers.

9 CONCLUSION

In this paper, we ask how interactive interfaces can make medical research papers approachable to healthcare consumers that need it. Our key insight is that medical papers can be made more approachable by incorporating plain language summaries alongside original paper content and providing guidance on the most important passages to read. We design a novel interface, Paper Plain, to provide reading support with interactive features that make use of recent developments in natural language processing. In a usability study of Paper Plain, participants who used Paper Plain report having less difficulty reading research papers compared to those who used a typical PDF reader. One feature that was particularly appreciated was the key question index, which supported question-based paper navigation. With further investment in design, AI, and careful consideration around deployment, we see tools like Paper Plain as playing a role in helping healthcare consumers become more aware of advances relevant to them in the medical field.

ACKNOWLEDGMENTS

This work would not have been possible without the contributions of others. Joseph Chang, Hyeonsu Kang, Raymond Fok, Dan Weld and Amy Zhang provided feedback on the paper. Matt Latzke taught us how to use Figma for iterative design. Sam Skjonsberg provided mentorship during the development of the prototype UI. Bailey Kuehl assisted us with recruiting study participants on Upwork. Amy L. Marsha gave us pointers on making R figures. Jay Bartot, Keshet Ronen, Irene Njuguna, and Yolanda Evans advised us on the perspectives of healthcare consumers, caregivers, and providers. Katharina Reinecke and Noah A. Smith provided general feedback on the project. This research was performed during an internship at the Allen Institute for AI with the Semantic Scholar team.

REFERENCES

[1] 2021. Fermat's Library. https://fermatslibrary.com/ Data accessed: August 1, 2021.

[2] 2021. Hypothes.is: Annotate the web, with anyone, anywhere. https://web.hypothes.is Data accessed: July 1, 2021.
[3] 2021. New Standards for Plain Language Summaries. https://consumers.cochrane.org/PLEACS Data accessed: July 1, 2021.
[4] 2021. UpToDate: Evidence-based Clinical Decision Support. https://www.wolterskluwer.com/en/solutions/uptodate Data accessed: June 15, 2021.
[5] 2021. Who needs access? You need access! https://web.archive.org/web/20220205013344/https://whoneedsaccess.org/category/patients/ Data accessed: July 1, 2021.
[6] 2021. Wiktionary, the free dictionary. https://en.wiktionary.org/wiki/Wiktionary:Main_Page Data accessed: August 1, 2021.
[7] Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics* 20, 1 (2019), 511.
[8] Takeshi Abekawa and Akiko Aizawa. 2016. SideNoter: Scholarly Paper Browsing System based on PDF Restructuring and Text Annotation. In *International Conference on Computational Linguistics*.
[9] 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, 1 (2019), 661–671.
[10] 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: A meta-analysis. *Medicine* 100, 9 (2021), e24747.
[11] Virginia Barbour, Paul Chinnock, Barbara Cohen, and Gavin Yamey. 2006. The impact of open access upon public health. *Bulletin of the World Health Organization* 84 5 (2006), 339.
[12] 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, 2 (2020), 950–957.
[13] Douglas M. Bates, Benjamin M. Bolker, and Steven C. Walker. 2015. Fitting linear mixed effects models using lme 4. *Journal of Statistical Software* 67, 1–48. Issue 1.
[14] Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *International Conference on Computational Linguistics*.
[15] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI Research: Going Behind the Scenes. In *Synthesis Lectures on Human-Centered Informatics*.
[16] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32, Database issue (2004), D267–D270.
[17] Tanner A. Bohn and Charles X. Ling. 2021. Hone as You Read: A Practical Type of Interactive Summarization. *ArXiv* abs/2105.02923 (2021).
[18] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 101 – 77.
[19] Anthony C. Breu. 2020. From Tweetstorm to Tweetorials: Threaded Tweets as a Tool for Medical Education and Knowledge Dissemination. *Seminars in nephrology* 40 3 (2020), 273–278.
[20] Mary Anne Britt, Tobias Richter, and Jean-François Rouet. 2014. Scientific Literacy: The Role of Goal-Directed Reading and Evaluation in Understanding Scientific Information. *Educational Psychologist* 49, 2 (2014), 104 – 122.
[21] Rainer Bromme and Susan R. Goldman. 2014. The Public's Bounded Understanding of Science. *Educational Psychologist* 49, 2 (2014), 59 – 69.
[22] Alex Broom. 2005. Virtually He@lthy: The Impact of Internet Use on Disease Experience and the Doctor-Patient Relationship. *Qualitative Health Research* 15, 3 (2005), 325 – 345.
[23] Phil Brown, Stephen Zavestoski, Sabrina McCormick, Brian Mayer, Rachel Morello-Frosch, and Rebecca Gasior Altman. 2004. Embodied health movements: new approaches to social movements in health. *Sociology of health & illness* 26 1 (2004), 50–80.
[24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. 33 (2020), 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
[25] Ross C. Brownson, Jonathan E. Fielding, and L. W. Green. 2018. Building Capacity for Evidence-Based Public Health: Reconciling the Pulls of Practice and the Push of Research. *Annual review of public health* 39, 1 (2018), 27–53.
[26] Peter Brusilovsky and Leonid Pesin. 2015. Adaptive Navigation Support in Educational Hypermedia: An Evaluation of the ISIS-Tutor. In *International Conference on Computer and Information Technology*.

[27] Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen. In *Association for Computational Linguistics*.

[28] 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* (2011).

[29] 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, 3 (2013), 55–72.

[30] 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* (2014).

[31] Rune Christensen. 2018. Cumulative Link Models for Ordinal Regression with the R Package ordinal. https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf R package version 2022.11-16. Data accessed: September 1, 2021.

[32] 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, 20 (1997), 2349–2380.

[33] 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, 8 (2018), 342–347.

[34] Robert Cudeck. 1996. Mixed-effects Models in the Study of Individual Differences with Repeated Measures Data. *Multivariate behavioral research* 31 3 (1996), 371–403.

[35] 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, 1 (2020), 8.

[36] Dina Demner-Fushman and Noémie Elhadad. 2016. Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing. *Yearbook of medical informatics* 1 (2016), 224–233.

[37] Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. In *NAACL*, Vol. 2021. NIH Public Access, 4972.

[38] Kristina Dzara and Ariel S Frey-Vogel. 2019. Medical Education Journal Club for the Millennial Resident: An Interactive, No-Prep Approach. *Academic pediatrics* 19 (2019), 603–607. Issue 6.

[39] Saul Epstein. 1997. Impure Science: AIDS, Activism, and the Politics of Knowledge. *Nature Medicine* 3, 2 (1997), 242–243.

[40] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research* 22, 1 (2004), 457–479.

[41] Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Asli Çelikyilmaz, and Yejin Choi. 2021. Discourse Understanding and Factual Consistency in Abstractive Summarization. In *European Chapter of the Association for Computational Linguistics*.

[42] Saadia Gabriel, Asli Çelikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A Meta Evaluation of Factuality in Summarization. In *Findings of the Association for Computational Linguistics*.

[43] Wee Chung Gan and Hwee Tou Ng. 2019. Improving the Robustness of Question Answering Systems to Question Paraphrasing. In *Association for Computational Linguistics*.

[44] Katy I. Gero, Vivian Liu, and Lydia B. Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Designing Interactive Systems Conference (DIS '22)*. Association for Computing Machinery, 1002–1019.

[45] Yue Guo, Weijian Qiu, Yizhong Wang, and Trevor A. Cohen. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 160–168.

[46] Gordon H. Guyatt, John A. Cairns, David N. Churchill, Deborah J. Cook, Brian R Haynes, Jack Hirsh, Jan Irvine, M. Levine, Mitchell A. H. Levine, Jim Nishikawa, David L. Sackett, Patrick Brill-Edwards, Hertzel C. Gerstein, Jim Gibson, Roman Z. Jaeschke, Anthony T. Kerigan, Alan J. Neville, Akbar A Panju, Allan S. Detsky, Murray W. Enkin, Pamela J Frid, Martha S. Gerrity, Andreas Laupacis, Valerie A. Lawrence, Joël Ménard, Virginia A. Moyer, Cynthia D. Mulrow, Paul S. Links, Andrew David Oxman, Jack Sinclair, and Peter Tugwell. 1992. Evidence-based medicine. A new approach to teaching the practice of medicine. *Journal of the American Medical Association* 268 17 (1992), 2420–5.

[47] 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* (2021).

[48] 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, 9 (2020), 2748–2761.
[49] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.
[50] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *ArXiv* abs/1702.08138 (2017).
[51] 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* (2018).
[52] 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, 2 (2021), e19910.
[53] 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, 1–2 (2021), 58 – 63.
[54] 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* 7, 1 (2021). 18898.
[55] Yea-Seul Kim, Jessica R. Hullman, and Eytan Adar. 2015. DeScipher : A Text Simplification Tool for Science Journalism. *Computation + Journalism*.
[56] Katrin Kirchhoff, Anne M. Turner, Amittai Axelrod, and Francisco Saavedra. 2011. Application of statistical machine translation to public health information: a feasibility study. *Journal of the American Medical Informatics Association* 18 4 (2011), 473–8.
[57] Joëlle Kivits. 2006. Informed Patients and the Internet. *Journal of Health Psychology* 11, 2 (2006), 269 – 282.
[58] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *EMNLP*. 9332–9346.
[59] Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 13 (2017), 1–26.
[60] Philippe Laban, John Canny, and Marti Hearst. 2019. A framework for a text-centric user interface for navigating complex news stories. *Computation + Journalism*.
[61] Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep It Simple: Unsupervised Simplification of Multi-Paragraph Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Online, 6365–6378.
[62] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics* 10, 163–177.
[63] 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, 11 (2018), 1917 – 1919.
[64] Gondy Leroy, James E. Endicott, David Kauchak, Obay Mouradi, and Melissa Just. 2013. User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention. *Journal of Medical Internet Research* 15, 7 (2013), e144.
[65] 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, 6 (2017), 483–490.
[66] Jiazhao Li, Corey A. Lester, Xinyan Zhao, Yuting Ding, Yun Jiang, and V. G. Vinod Vydiswaran. 2020. PharmMT: A Neural Machine Translation Approach to Simplify Prescription Directions. In *Findings of the Association for Computational Linguistics*.
[67] Magnus Lindstrom and Douglas M. Bates. 1990. Nonlinear mixed effects models for repeated measures data. *Biometrics* 46 3 (1990), 673–87.
[68] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Comput. Surveys* 55, 9 (2023), 1–35.
[69] Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Research and Advanced Technology for Digital Libraries*.
[70] 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. *British Medical Journal* 333, 7574 (2006), 883.

[71] 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 *North American Chapter of the Association for Computational Linguistics*.

[72] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLoS ONE* 14, 8 (2019), 1–16.

[73] 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* 23, 1 (2016), 193 – 201.

[74] Alejandro Martínez and Stefano Mammola. 2021. Specialized terminology reduces the number of citations of scientific papers. *Proceedings of the Royal Society B* 288, 1948 (2021), 20202581.

[75] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *ACL*. Association for Computational Linguistics, 1906–1919.

[76] 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, 1 (2021), e913.

[77] Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung H. Bui, Walter Chang, Emilia Farcas, and Ndapandula Nakashole. 2021. A Gradually Soft Multi-Task and Data-Augmented Approach to Medical Question Understanding. In *Association for Computational Linguistics*.

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

[79] Sonia K. Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl, Sophie Johnson, Jon Borchardt, Daniel S. Weld, Tom Hope, and Doug Downey. 2022. ACCoRD: A Multi-Document Approach to Generating Diverse Descriptions of Scientific Concepts. In *EMNLP: System Demonstrations*. Association for Computational Linguistics, 200–213.

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

[81] 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.

[82] 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*. Association for Computational Linguistics, 319–327.

[83] Emily Nunn and Stephen Pinfield. 2014. Lay summaries of open access journal articles: engaging with the general public on medical research. *Learned Publishing* 27, 3 (2014), 173–184.

[84] National Institutes of Health. 2005. Policy on enhancing public access to archived publications resulting from NIH-funded research. https://grants.nih.gov/grants/guide/notice-files/not-od-05-022.html Notice Number: NOT-OD-05-022. Data accessed: August 1st, 2021.

[85] Gustavo Paetzold and Lucia Specia. 2016. Anita: An Intelligent Text Adaptation Tool. In *International Conference on Computational Linguistics*.

[86] 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* (2014).

[87] Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael A. Hogarth. 2017. Text Simplification Using Consumer Health Vocabulary to Generate Patient-Centered Radiology Reporting: Translation and Evaluation. *Journal of Medical Internet Research* 19, 12 (2017), e417.

[88] Vololona Rabeharisoa, Tiago Moreira, and Madeleine Akrich. 2014. Evidence-based activism: Patients', users' and activists' groups in knowledge society. *BioSocieties* 9 (2014), 111–128.

[89] Tessa Richards and Fiona Godlee. 2014. The BMJ's own patient journey. *British Medical Journal* 348, 7962 (2014), g3726.

[90] 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 (1995), A12–3.

[91] Kirk Roberts, Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 2014. Automatically Classifying Question Types for Consumer Health Questions. In *AMIA Annual Symposium Proceedings*. AMIA Symposium, 1018–1027.

[92] Mary A M Rogers, Kelsey Lemmen, Rachel Kramer, Jason Mann, and Vineet Chopra. 2017. Internet-Delivered Health Interventions That Work: Systematic Review of Meta-Analyses and Evaluation of Website Availability. *Journal of Medical Internet Research* 19, 3 (2017), e90.

[93] 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, 8 (2017), 1003 – 1018.

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

[95] Cynthia Shanahan, Timothy Shanahan, and Cynthia Misischia. 2011. Analysis of Expert Readers in Three Disciplines. *Journal of Literacy Research* 43, 4 (2011), 393 – 429.

[96] Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S. Weld, and Doug Downey. 2022. VILA: Incorporating Visual Layout Structures for Scientific Text Classification. *Transactions of the Association for Computational Linguistics* 10 (2022), 376–392. https://doi.org/10.1162/tacl_a_00466

[97] 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 (2009), 266–71.

[98] Robyn Speer. 2022. *rspeer/wordfreq: v3.0.* https://doi.org/10.5281/zenodo.7199437

[99] 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. *Journal of the American Medical Association Surgery* 151 9 (2016), 831–7.

[100] 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. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *British Medical Journal* 349, 7987 (2014), g7015.

[101] 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, 8 (2020), 845 – 867.

[102] 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, 1 (2017), Article e9.

[103] 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, 632 (2016), v3.

[104] Hoang Van, David Kauchak, and Gondy Leroy. 2020. AutoMeTS: The Autocomplete for Medical Text Simplification. In *International Conference on Computational Linguistics*.

[105] Amir Pouran Ben Veyseh, Franck Dernoncourt, W. Chang, and Thien Huu Nguyen. 2021. MadDog: A Web-based System for Acronym Identification and Disambiguation. In *European Chapter of the Association for Computational Linguistics*.

[106] 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 (2014), 113–21.

[107] Esteban Walker and Amy S. Nowacki. 2011. Understanding Equivalence and Noninferiority Testing. *Journal of General Internal Medicine* 26, 2 (2011), 192–196.

[108] Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, and Jie fu. 2021. Pre-trained Language Models in Biomedical Domain: A Survey from Multiscale Perspective. *ArXiv* abs/2110.05006.

[109] Gerhard Weber and Peter Brusilovsky. 2015. ELM-ART – An Interactive and Intelligent Web-Based Electronic Textbook. *International Journal of Artificial Intelligence in Education* 26 (2015), 72–81.

[110] Ryen W. White and Ahmed Hassan Awadallah. 2014. Content Bias in Online Health Search. *ACM Transactions on the Web* 8, 4 (2014), 25:1–25:33.

[111] 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* 21 1 (2014), 49–55.

[112] 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*, Peggy Cellier and Kurt Driessens (Eds.). Springer International Publishing, Cham, 727–740.

[113] Yan Zhang. 2014. Searching for specific health-related information in MedlinePlus: Behavioral patterns and user experience. *Journal of the Association for Information Science and Technology* 65, 1 (2014), 53–68.

[114] Tian Zhao and Kyusong Lee. 2020. Talk to Papers: Bringing Neural Question Answering to Academic Search. In *ACL: System Demonstrations*. Association for Computational Linguistics, 30–36. https://doi.org/10.18653/v1/2020.acl-demos.5

[115] Alesia A. Zuccala. 2010. Open access and civic scientific information literacy. *Information Research* 15, 1 (2010), paper426.

A INTERVIEWS WITH HEALTHCARE CONSUMERS AND PROVIDERS

To validate the idea of helping readers understand medical research papers, we interviewed healthcare consumers and providers. We spoke with healthcare consumers with prior experience reading medical research (4 total, referred to as C1–4), and healthcare providers who had discussed findings from papers with their patients (2 total, H1–2). Interviewees 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.

The 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 treatment options. (C1, C4)
- **Comparing treatment options**: 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 [97] and illustrate the benefits of open-access medical literature [115] 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 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 §3.3 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 restricted their ability to explore the paper on their own. N3 wished for gists for scanning other sections of the paper that might not have an answering passage.

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 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.

C PAPER PLAIN IMPLEMENTATION

C.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.[4] 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:

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

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 attempt. The average number of attempts was 1.35, with a maximum of 4. Examples of re-generations are included in Table 6.

D STATISTICAL ANALYSIS

D.1 Modeling Mixed-Effects in Repeated Measures Studies

For the analysis in § 6.1.5, 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 [67]. LMMs are used as an analysis in medicine [32], the behavioral sciences [34], and human-computer interaction [47, 48].

For each of the quantitative measurements discussed in §6.1.3 (y), we fit a LMM with fixed effects β for the Paper Plain paper (x_1) and interface variant (x_2) factors.[5] We used the lme4 package in R [13] to fit the models. More precisely, we fit the following LMM:

$$E[y] = \beta_0 + \gamma_j + \beta_1 x_1 + \beta_2 x_2, \tag{1}$$

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 4. 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, section and terms (SD), and question and answer (QA) 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 3.

	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 4. Estimated fixed-effect coefficients for the LMM described in Appendix D for each measurement.

D.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 [59]. Using the Holm-Bonferroni [49] correction

[5] 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.

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.

D.3 Tests for Pairwise Differences in Fixed-Effects between Interfaces

To quantify pairwise differences in fixed-effects between 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 3.

D.4 Ordinal Regression for Likert-Scale Variables

As reading difficulty, confidence, and understanding were measured on a Likert-style 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 [31]. 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 3.

Question	Source	Extracted Answer	Simplified Answer
What condition does this paper study?	PICO	"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…"
How is the condition usually treated?	PICO	"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…"
What did the paper want to find out?	Cochrane	"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…"
What did the paper do?	Cochrane	"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…"
What were the new treatment(s), if any this paper looked into?	PICO	"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…"
What did the paper find?	Cochrane	"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…"
Are the findings different depending on a person's demographics?	PICO	"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…"
What are the limitations of the findings?	Cochrane	"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 5. Questions appearing in the key question index, in the order they appear in the index.

Original Passage	Generated Gist	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 6. Examples of generated plain language summaries, alongside our designation of whether they were usable in Paper Plain or whether they required regeneration. Errors in generation are indicated in **bold**.