APPLS: A Meta-evaluation Testbed for Plain Language Summarization

Yue Guo[1] **Tal August**[2] **Gondy Leroy**[3] **Trevor Cohen**[1] **Lucy Lu Wang**[1,2]
[1]University of Washington [2]Allen Institute for AI [3]University of Arizona
{yguo50, cohenta, lucylw}@uw.edu, tala@allenai.org

Abstract

While there has been significant development of models for Plain Language Summarization (PLS), evaluation remains a challenge. This is in part because PLS involves multiple, interrelated language transformations (e.g., adding background explanations, removing specialized terminology). No metrics are explicitly engineered for PLS, and the suitability of other text generation evaluation metrics remains unclear. To address these concerns, our study presents a granular meta-evaluation testbed, APPLS, designed to evaluate existing metrics for PLS. Drawing on insights from previous research, we define controlled perturbations for our testbed along four criteria that a metric of plain language should capture: informativeness, simplification, coherence, and faithfulness. Our analysis of metrics using this testbed reveals that current metrics fail to capture simplification, signaling a crucial gap. In response, we introduce POMME, a novel metric designed to assess text simplification in PLS. We demonstrate its correlation with simplification perturbations and validate across a variety of datasets. Our research contributes the first meta-evaluation testbed for PLS and a comprehensive evaluation of existing metrics, offering insights with relevance to other text generation tasks.[1]

1 Introduction

Plain language summaries of scientific information are important to make science more accessible (Kuehne and Olden, 2015; Stoll et al., 2022) and inform public decision-making (Holmes-Rovner et al., 2005; Pattisapu et al., 2020). Recently, generative models have made gains in translating scientific information into plain language approachable to lay audiences (August et al., 2022b; Goldsack et al., 2023; Devaraj et al., 2021). Despite these gains, the field has not reached consensus on effective automated evaluation metrics for plain language summarization (PLS) (Luo et al., 2022; Ondov et al., 2022). One reason is the multifaceted nature of the PLS task. Removal of unnecessary details (Pitcher et al., 2022), adding relevant background explanations (Guo et al., 2021), jargon interpretation (Pitcher et al., 2022), and text simplification (Devaraj et al., 2021) are all involved in PLS, posing challenges for comprehensive evaluation.

Our goal is to assess how well existing metrics capture the multiple criteria of PLS. We define four criteria, informed by prior work (Pitcher et al., 2022; Ondov et al., 2022; Stoll et al., 2022; Jain et al., 2022), that a measure of plain language should be sensitive to: 1) *informativeness*, 2) *simplification*, 3) *coherence*, and 4) *faithfulness*. We introduce a set of perturbations to probe metric sensitivity to these criteria, where each perturbation is designed to affect a single criterion with ideally minimal impact to others. Then, by incrementally introducing these perturbations to the texts of an existing scientific PLS dataset, CELLS (Guo et al., 2022), we produce APPLS, a novel, granular testbed to evaluate existing PLS metrics.

Using APPLS, we analyze 15 metrics, including the most widely used metrics in text simplification and summmarization and recently-proposed prompt-based evaluation (Gao et al., 2023; Luo et al., 2023). Established metrics like ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), and QAEval (Deutsch et al., 2021) demonstrate mixed sensitivities to informativeness, coherence, and faithfulness perturbations. All tested metrics, including those explicitly crafted for text simplification (Xu et al., 2016; Maddela et al., 2022), display a lack of sensitivity towards simplification perturbations.

In response to the lack of effective metrics for simplification, we introduce POMME, a new metric that evaluates text simplicity by leveraging language models (LMs) trained on in-domain (i.e., scientific) and out-of-domain (i.e., web) text. POMME

[1]The APPLS testbed and POMME will be made available at https://github.com/LinguisticAnomalies/APPLS

capitalizes on the fact that complex scientific text will be more similar to a scientific LM's domain-specific training data, while simpler text will align more closely with a general-domain LM. Given the inherent adaptability of LMs to various domains (Gururangan et al., 2020), POMME can be tailored to the specific domain of the data being evaluated.

Our main contributions are as follows:

- We present APPLS, the first granular testbed for analyzing evaluation metric performance for plain language summarization (§3, 4, 5);
- We assess the performance of existing evaluation metrics, demonstrating mixed effectiveness in evaluating informativeness, coherence, and faithfulness, and revealing their limitations in capturing simplification (§6, 8);
- We introduce a new metric, POMME, which employs language model perplexity to assess text simplicity, and validate its performance in our testbed and in three other datasets (§7, 8).

2 Related Work

Limitations of Existing Metrics While overlap-based metrics, such as ROUGE, BLEU, and METEOR, are known for their ease of use, their shortcomings in detecting crucial attributes like faithfulness (Wallace et al., 2021; Pagnoni et al., 2021), coherence (Barzilay and Lapata, 2008), and simplification (Silveira and Branco, 2012; Sulem et al., 2018) have been well-documented. Pre-trained model-based metrics like BERTScore, despite demonstrating strong correlations with human evaluations (Zhang et al., 2019), have been critiqued for their insensitivity towards factual inconsistencies (He et al., 2022). QA-based metrics, gaining momentum in evaluating faithfulness in summarization tasks (Durmus et al., 2020; Scialom et al., 2019; Deutsch et al., 2021), depend significantly on the employed question generation method (Gabriel et al., 2020) and the chosen answer verification approach (Deutsch and Roth, 2022), with their effectiveness in the PLS context yet unexplored. Recent prompt-based evaluations exhibit potential for evaluating factuality (Luo et al., 2023) and summarization quality (Gao et al., 2023), but have also not been tested for PLS. Our research aims to fill these gaps by conducting a systematic evaluation of these metrics under specific perturbations in a PLS context.

Robust Analysis with Synthetic Data Synthetic data has been widely used in various NLP tasks to evaluate metrics, including text generation (He et al., 2022; Sai et al., 2021), natural language inference (Chen and Eger, 2022; McCoy et al., 2019), question answering (Ribeiro et al., 2019), and reading comprehension (Sugawara et al., 2020). Yet, no prior work has specifically focused on the PLS task or incorporated simplification into their synthetic benchmarks. Additionally, previous studies have not conducted granular analyses to capture the nuanced relationship between text changes and score changes. Our research endeavors to bridge these gaps by crafting perturbations that mirror real-world errors and concentrating on the 'dose-response' relationship (Talbot and Aronson, 2011) between score changes and perturbations within the PLS context.

3 Desired Criteria for PLS Metric

We define four criteria that an effective evaluation metric for PLS should be sensitive to. We define sensitivity similar to prior work (Gabriel et al., 2020) as being correlated in the correct direction with the amount of perturbation. These criteria are informed by both abstractive summarization (Sai et al., 2022) and plain language summarization paradigms (Pitcher et al., 2022; Ondov et al., 2022; Stoll et al., 2022; Jain et al., 2022).

Informativeness measures the extent to which a plain language summary covers essential information such as methodology, primary findings, and conclusions in the original text. An informative summary conveys the central messages of the text, and does not remove crucial details (Devaraj et al., 2022; Pitcher et al., 2022) or hallucinate (Maynez et al., 2020; Goyal and Durrett, 2021). *We hypothesize that an ideal metric should decrease with the elimination of notable sentences and the insertion of irrelevant sentences, and increase with the incorporation of relevant definitions.*

Simplification describes the degree to which information is conveyed in a form that non-expert audiences can readily understand. This criterion prioritizes the use of simple vocabulary (Bingel et al., 2018; Laban et al., 2021), casual language (Pitcher et al., 2022), and concise sentences (De Belder and Moens, 2010; Scarton et al., 2018) that minimize excessive jargon and technical terminology unfamiliar to a lay audience.[2] Reducing

[2]Text simplification is its own distinct task; in comparison

Notations: ~~removals~~/ additions/ modifications

Original text	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them. The first step is an accurate assessment of the population prevalence of past infections… (Kline et al., 2021)		
Criterion	**Perturbation**	**Simulated real-world situation**	**Perturbed text**
Informativeness	Delete sentences	Salient information missing	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them. ~~The first step is an accurate assessment of the population prevalence of past infections~~…
	Add out-of-domain sentences	Out-of-domain hallucination	In this paper we address the problem of aggregating the outputs of classi ers solving different nlp tasks. Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them…
	Add in-domain sentences	In-domain hallucination	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them. This review synthesised the latest evidence on the reduction of antipsychotic doses for stable individuals with schizophrenia…
	Add definitions	Background explanation	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them. Coronaviruses are species in the genera of virus belonging to the subfamily Coronavirinae in the family Coronaviridae. Coronaviruses are enveloped viruses with a positive-sense RNA genome and with a nucleocapsid of helical symmetry. The genomic size of coronaviruses ranges from approximately 26 to 32 kilobases, extraordinarily large for an RNA virus. …
Simplification	Replace sentences	Paraphrasing with simple terms	SARS-CoV-2 is a virus that has infected over 59 million people globally and killed more than 1.39 million. Scientists are trying to learn more about the virus in order to design interventions to slow and stop its spread. One of the first steps is understanding how many people have been infected in the past, which requires accurate population prevalence studies…
Coherence	Reorder sentences	Poor writing flow	The first step is an accurate assessment of the population prevalence of past infections. Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them…
Faithfulness	Number swap	Human errors	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 64 million people and killed more than one of them…
	Entity swap	Human errors	Worldwide, canine adenovirus (CaV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them…
	Synonym verb swap	Human errors	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and stamped out more than one of them…
	Antonym verb swap	Human errors	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, infected more than 59 million people and saved more than one of them…
	Negate	Human errors	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, hasn't infected more than 59 million people and killed more than one of them.

Table 1: Example perturbations for criteria in APPLS. Original text comes from the CELLS (Guo et al., 2022).

complexity by substituting complex sentences with simpler ones has been empirically shown to reduce the difficulty of text (Van den Bercken et al., 2019). *We hypothesize that an ideal metric should exhibit sensitivity towards the substitution of complex sentences with simplified counterparts in the text.*

Coherence describes the logical arrangement of a plain language summary. A coherent summary presents information in a well-ordered fashion that facilitates ease of comprehension for the reader (DS, 2001; Sai et al., 2022). Barzilay and Elhadad (2002) underscored the significance of sentence sequencing in affecting user comprehension. We conjecture that the original sentence order reflects optimal coherence. *Consequently, we hypothesize that an ideal metric should demonstrate sensitivity to modifications in sentence sequencing.*

Faithfulness denotes how well the plain language summary aligns factually with the source text. A faithful summary should not substitute information or introduce errors, misconceptions, and inaccuracies (Devaraj et al., 2022; Prabhakaran et al., 2019). Faithfulness focuses on factual alignment, while informativeness measures the completeness and efficiency of the summary in conveying key points. *We hypothesize that an ideal metric should exhibit sensitivity towards changes in factual information such as entity swaps and sentence negation, while maintaining indifference to synonym swaps.*

4 Criteria-specific Perturbation Design

To assess existing metric sensitivity to our proposed criteria, we develop the following perturbations (illustrative examples in Table 1).

4.1 Informativeness

Delete sentences We simulate the omission of crucial information by ranking sentences based on similarity to others (assuming more similar is more likely to contain important information) (Zhong et al., 2020) and removing sentences starting from the most to least similar.

Add sentences We simulate the inclusion of unrelated information by adding sentences. We incorporate two forms of unrelated information: *out-of-*

to the PLS task, text simplification focuses on surface-level changes to simplify language and does not involve other relevant criteria important to PLS (e.g., informativeness).

domain, which integrates random sentences from an unrelated dataset, and *in-domain*, which includes sentences from a different summary within the same domain but on a different topic.

Add definitions Background explanation is fundamental to PLS and involves adding external content such as definitions or examples (Guo et al., 2022; Srikanth and Li, 2020). To simulate this phenomenon, we add definitions[3] of keywords identified by KeyBERT (Grootendorst, 2020).

4.2 Simplification

Replace sentences Taking advantage of the GPT-series models' ability to simplify text (Lu et al., 2023), we replace sentences in the original text with GPT-simplified versions. We use the *text-davinci-003* model to generate simplified summaries using the prompt "`explain the text in layman's terms to a primary school student.`" GPT configurations can be found in App. A.

4.3 Coherence

Reorder sentences We simulate changes in text coherence by randomly shuffling the order of sentences, as suggested by Sai et al. (2021).

4.4 Faithfulness

Number swap We randomly add a number from 1 to 5 to the original numerical value in the text.

Verb swap An appropriate metric should exhibit constancy for synonymous verbs but sensitivity for antonymous ones. To this end, we introduce two perturbations, where we identify verbs in text and substitute them with either synonyms or antonyms.

Entity swap We replace entities using the KBIN method (Wright et al., 2022), which links entity spans to concepts in the Unified Medical Language System (UMLS) and replaces them with different entities while maximizing NLI contradiction and minimizing LM perplexity. This results in a fluent sentence that contradicts the original one.

Negate sentences We negate sentences by identifying verbs and adding negation terms (e.g., not) preceding them. The goal of this perturbation is to create sentences similar to the original but communicating the exact opposite information.

5 Constructing the APPLS testbed

We implement our perturbations in an existing large-scale PLS dataset (§5.1). We describe how perturbations are incorporated into the dataset and our approach for managing perturbation magnitude (§5.2) and validating perturbation quality (§5.3). We employ this testbed in an analysis of existing (§6) and novel (§7) metrics for PLS (§8).

[3]Taken from http://wikidata.dbpedia.org/develop/datasets

	Src.	Tgt.	Oracle	GPT
Avg. # words	283	178	134	98
Avg. # sentences	11	7	5	4
Vocabulary size	76,275	46,522	46,658	27,937

Table 2: CELLS test set (n=6,311) characteristics for source (scientific abstract), target (plain language summary), oracle extractive hypothesis, and GPT-simplified oracle summaries.

5.1 Diagnostic dataset

For our experiments, we use the CELLS dataset (Guo et al., 2022): a parallel corpus of scientific abstracts (designated as *source*) and their corresponding plain language summaries (designated as *target*). The summaries are written by the abstract authors or by other domain experts. CELLS aggregates papers from 12 biomedical journals and is the most extensive and diverse abstract-level compilation for PLS presently available.

Many of the metrics we assess require three texts: source, target, and model-generated text (referred to as *hypothesis*). For our meta-evaluation testbed, we propose an *oracle extractive hypothesis*. This hypothesis is created by selecting a set of source sentences yielding the highest ROUGE-L score when compared to the target summary, and further introducing lexical variability into the text through round-trip translation (Ormazabal et al., 2022) (details in App. B). This produces a reasonable hypothesis that summarizes the source text while minimizing factual inaccuracies.[4]

We then apply perturbations to the oracle hypothesis, where each perturbation introduces some *change* (e.g., adding or swapping sentences) at some *magnitude* (e.g., replace 50% of sentences) to the oracle. Given the costs associated with some of our perturbations (e.g., GPT-based simplification), we restrict our perturbation experiments to the test set of CELLS (stats in Table 2).

[4]Why not use the extractive summary directly? Metrics like SARI expect simplified hypotheses and exhibit degenerate behavior when used to evaluate extractive summaries.

5.2 Applying perturbations to CELLS

For *informativeness* perturbations, we add sentences to the oracle hypothesis from ACL papers (Bird et al., 2008) to simulate out-of-domain hallucinations and Cochrane abstracts[5] for in-domain hallucinations. The perturbation percentage is the ratio of altered sentences to the initial count of sentences in the hypothesis (e.g., 100% perturbed for sentence addition adds the same number of sentences as in the original hypothesis). For sentence deletion, max perturbation is when a single sentence remains. For keyword definitions, we add up to three definitions, which reflects the average number of nouns explained in CELLS abstracts (Guo et al., 2022); 100% perturbation for this category entails the insertion of three definitions.

For *simplification* perturbations, we align sentences between the oracle hypothesis and the GPT-simplified summary using the sentence alignment algorithm originally employed in CELLS (Guo et al., 2022). We replace hypothesis sentences with corresponding GPT-simplified sentences. Full perturbation is when all hypothesis sentences are replaced with simplified counterparts.

For *coherence*, we quantify perturbation percentage based on the distance between the hypothesis and shuffled sentences in terms of absolute difference in sentence order. A document with reversed sentence order would be 100% perturbed.

For *faithfulness*, we determine the perturbation percentage of number, entity, and verb swaps by comparing the count of altered spans to the total number of eligible spans in the hypothesis. Full perturbation implies that all eligible spans are swapped. For sentence negation, we constrain the maximum number of negations to the sentence count in the hypothesis, allowing for only one negation per sentence. Therefore, full perturbation is achieved when each sentence contains a negation.

5.3 Human validation of oracle extractive hypotheses and GPT-simplified summaries

Because the creation of both oracle extractive hypotheses and GPT-simplified summaries involve text generation, we further validate their quality through human evaluation. For oracle hypotheses, we sample 100 paired sentences before and after round-trip translation of the extractive summaries. For GPT-simplified summaries, we sample 100 paired texts coupling a passage in the oracle hypothesis to a passage in the GPT-simplified summary. Annotators were asked to assess each pair of texts, first judging whether the content aligns (defined as containing the same relation triples), then rating informativeness, simplification, faithfulness, and coherence on a 5-point Likert scale. Annotations were performed by two independent annotators, both with doctorates in the biological sciences, who were hired on UpWork and paid a fair hourly wage. Each annotator reviewed all sampled pairs for both evaluation tasks. Inter-rater agreement, measured by Cohen's Kappa, is 0.46, implying moderate agreement (Artstein and Poesio, 2008). For task details, refer to App. C.

For round-trip translation, annotators validated that translated text was as informative, faithful, coherent, and simple as the original in the vast majority of cases. For GPT-simplified summaries, the evaluators rated the sentences as highly simplified, informative, faithful, and coherent when the original and simplified sentences were correctly aligned. However, a considerable fraction (46 out of 100) of the texts showed imperfect sentence-level alignment, suggesting potential information loss in partial simplification perturbations. Sentence-level alignment for scientific summaries is still an open problem (Krishna et al., 2023), and our evaluation highlights the need for improved alignment algorithms. For simplification perturbations, we additionally report on a smaller, manually sentence-aligned dataset (PLABA), due to the inconsistencies we found in automated alignment. We discuss these limitations further in §9.

6 Existing Metrics We Evaluate

Our analysis spans eight established evaluation metrics, including five metrics most commonly reported in ACL'22 summarization and generation papers (empirical results in App. D) and three additional metrics (§6.1). We also assess five lexical features associated with text simplification (§6.2) and LLM-based evaluations (§6.3).

6.1 Existing automated evaluation metrics

Overlap-based metrics measure n-gram overlaps, and are popular due to their ease of use.

- **ROUGE**[6] (Lin, 2004) measures n-gram overlap between generated and reference summaries, focusing on recall. We report the average of ROUGE-1, ROUGE-2, and ROUGE-L.

[5] https://community.cochrane.org

[6] Implementation: Fabbri et al. (2021)

- **BLEU**[6] (Papineni et al., 2002) computes n-gram precision of generated text against reference texts, including a brevity penalty.
- **METEOR**[6] (Banerjee and Lavie, 2005) employs a relaxed matching criterion based on the F-measure, and addresses the exact match restrictions and recall consideration of BLEU.
- **SARI**[7] (Xu et al., 2016) is specifically designed to evaluate text simplification tasks. The score weights deleted, added, and kept n-grams between the source and target texts.

Model-based metrics use pretrained models to evaluate text quality.

- **GPT-PPL**,[8] usually computed with GPT-2, measures fluency and coherence by calculating the average log probability assigned to each token by the GPT model, with lower scores indicating higher fluency and coherence.
- **BERTScore**[6] (Zhang et al., 2019) quantifies the similarity between hypothesis and targets using contextualized embeddings from the BERT model, computing the F1-score between embeddings to capture semantic similarity beyond n-gram matching.
- **LENS** (Maddela et al., 2022) employs an adaptive ranking loss to focus on targets closer to the system output in edit operations (e.g., splitting, paraphrasing, deletion).

QA-based metrics capture content quality using a question-answering approach.

- **QAEval** (Deutsch et al., 2021) generates question-answer pairs from the target text, then uses a learned QA model to answer these questions using the generated text. The score is computed as the proportion of questions answered correctly. We report QAEval LERC scores.

6.2 Lexical features

We also assess lexical features that have been shown to be predictive of text simplicity:

- **Length**: Shorter sentences are easier to understand (Kauchak et al., 2017). We report both sentence length and paragraph length.
- **Familiarity**: Simple text contains more common words (Leroy et al., 2018). We compute the percentage of text that is made up of the 1,000 most common English words.[9]
- **Specificity**: For biomedical text, MeSH term depth is predictive of sentence difficulty (Kauchak et al., 2014). MeSH terms are used to index articles in PubMed, and we determine the depth of an article's terms by splitting term identifiers on '.' (e.g., term depth for Zika Virus with identifier B04.820.578.344.350.995 is 6).
- **Phrase Transitions**: Conjunctions (e.g., therefore) are important for flow and can assist with comprehension (Kauchak et al., 2017). We report the number of conjunctions.
- **Function Words**: Simple text contains more verbs and fewer nouns (Mukherjee et al., 2017). We report the number of verbs, nouns, adjectives, adverbs, and numbers.

6.3 LLM prompt-based evaluations

Prompting LLMs for text generation evaluation has been explored in recent work (Gao et al., 2023; Luo et al., 2023). We adopt the template from Gao et al. (2023), prompting GPT-3 (*text-davinci-003*) to evaluate the hypotheses in our testbed across four dimensions—informativeness, simplification, coherence, and faithfulness—and provide an overall quality score. All scores are from 0 (worst) to 100 (best). We supply the definitions for each criterion in the prompt.

We test two experimental settings: one in which we only provide the source abstract (reference-free), and the other in which both the source abstract and the target plain language summary are provided (reference-provided). Model configurations and prompt details are available in App. E.

7 Novel Metric: POMME

We introduce a novel, lightweight metric (POMME) to assess text simplification by leveraging pretrained LMs. Prior work has relied on LMs like GPT-2 to assess readability and coherence through perplexity (Zhao et al., 2022; Kanthara et al., 2022), but these measures exhibit considerable sensitivity to text length (Wang et al., 2022), which is undesirable for PLS evaluation. Our own investigation corroborates this, showing divergent PPL responses to simplification in different datasets (Tables 4; 3).

To address these issues, our POMME metric employs the *difference* in perplexity scores from an in-domain and an out-of-domain LM, which acknowledges and takes advantage of the inherent domain shift from scientific text to plain language that PLS entails. The underlying hypothesis is that LMs pretrained on scientific text should assign lower perplexity scores to scientific texts than

[7] Implementation: Alva-Manchego et al. (2019)
[8] https://huggingface.co/transformers/v3.2.0/perplexity.html
[9] https://gist.github.com/deekayen/4148741

LMs pretrained on general English, with the inverse being true for plain language (Harris et al., 2012). Similar intuitions have driven the use of domain shifts between various LMs trained on diverse corpora to guide text generation (Liu et al., 2021), specifically in the scientific domain (August et al., 2022a).

To address differences in magnitude when comparing perplexity scores from models with distinct vocabulary sizes, we normalize POMME by computing and subtracting the perplexity Z-scores rather than using the raw values. POMME is computed as:

$$Z(x) = \frac{\log(x) - \mu_{\text{ref}}}{\sigma_{\text{ref}}}$$
$$\text{POMME} = Z\left(\text{PPL}_{\text{id}}\right) - Z\left(\text{PPL}_{\text{ood}}\right)$$

where μ is the mean and σ the standard deviation of the perplexity of texts in the reference dataset. We use BioMedLM (Bolton et al.) as our in-domain ("scientific") LM and T5 (Raffel et al., 2020) as our out-of-domain ("plain"). BioMedLM was trained exclusively on PubMed abstracts (matching the domain of CELLS) while T5 was trained on primarily general-domain data like web text and Wikipedia.

The essence of POMME is determining the relative position of a text's perplexity within the distribution of perplexity scores of a reference dataset with texts of at least two known levels of simplification. Using a reference dataset ensures that POMME scores are comparable across different datasets. A notable advantage of POMME is that it is model agnostic, allowing the use of any two models as the in- and out-of-domain LMs. Thus, POMME could be adapted to evaluate text simplification in other fields, such as legal text (Jain et al., 2021) or financial regulations (Salo et al., 2016). In the scope of this work, we limit the evaluation of POMME to the realm of biomedical text, given the availability of pretrained models and author-written paired plain-language summaries.

8 Analysis Results

To assist in interpreting metric performance in the APPLS testbed, we survey reported metric changes in ACL 2022 papers on text generation and summarization (full results in App. D). The median reported improvements are: ROUGE (+0.89), BLEU (+0.69), METEOR (+0.50), SARI (+1.71), BERTScore (+0.55), and PPL (-2.06). Table 7 lists these score differences along with observed score differences based on our perturbations. Line plots of existing metric scores corresponding to perturbations are shown in Figure 1. We summarize main findings of our analysis below.

Current metrics exhibit shortcomings in evaluating simplicity. We expect metrics that are sensitive to simplification to increase in response to our simplification perturbation. Metrics such as ROUGE, BLEU, METEOR, BERTScore, and QAEval decrease in response to the simplification perturbation. This is expected given their goal is to measure content overlap (GPT-simplified text has less n-gram/content overlap with the target). Disappointingly, the SARI metric also decreases, which deviates from its original design intent of measuring text simplification. LENS response to perturbations is similarly erratic. The only metric that exhibits sensitivity to simplification perturbations is GPT-PPL (decreasing as more perturbations are introduced; lower PPL is better); however, the metric does not consistently detect simplicity in other datasets and is difficult to compare between datasets. We comment on this further in relation to the POMME score below. LLM prompt-based evaluations also fail to respond to simplification perturbations; Table 6 shows score reductions in both the reference-free and reference-provided settings. Together, these findings underscore the need for a more effective simplicity assessment metric.

Metrics effectively capture informativeness, coherence, and faithfulness, but there is room for improvement. For informativeness-based perturbations, ROUGE, BLEU, BERTScore, PPL, and QAEval are sensitive to information deletion and irrelevant additions, but struggle to capture the effect of background explanations through keyword definitions. For coherence, BERTScore and LENS are proficient at detecting perturbations, most likely due to their ability to analyze structural and contextual sentence relationships. BERTScore, PPL, and QAEval generally perform well for faithfulness-related perturbations. Both PPL and BERTScore are somewhat sensitive to synonym verb swaps (an undesirable sensitivity for faithfulness). QAEval is best at being unresponsive to synonym verb swaps. All metrics fall short in effectively capturing number swaps.

Lexical features are useful measures of text simplicity. Figure 2 illustrates the response of lexical features to varying degrees of text

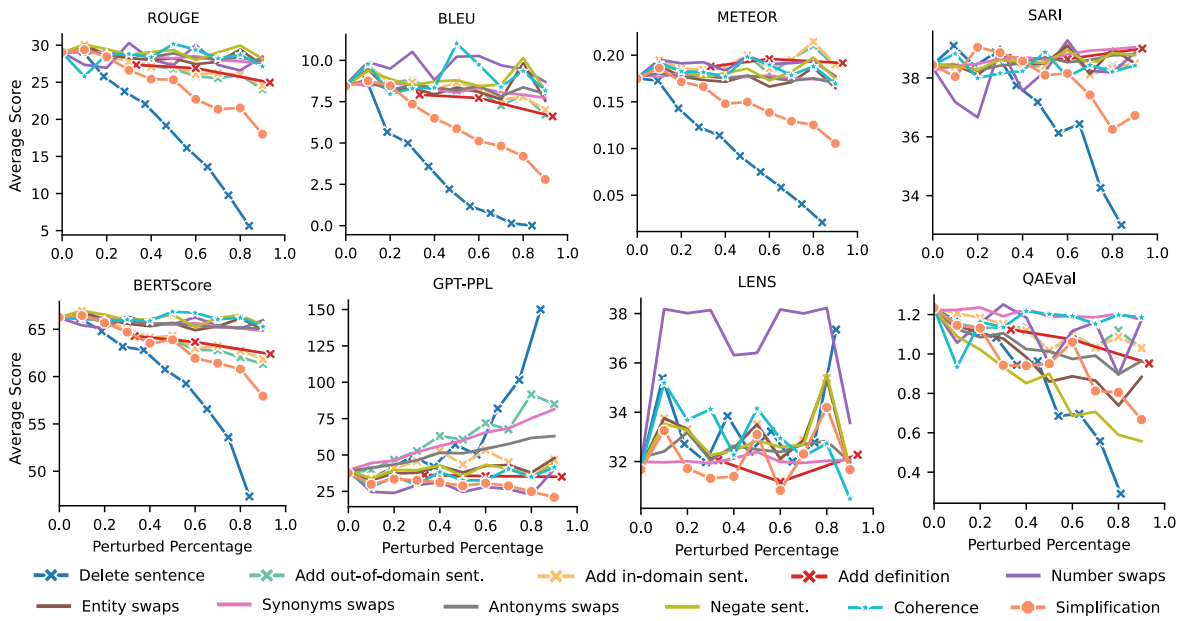

Figure 1: Average scores of existing metrics for perturbed texts, plotted as line charts. Scores are averaged in 10 bins by perturbation percentage. Markers denote perturbations arising from our four defined criteria. GPT-PPL is the only metric exhibiting sensitivity to the simplification perturbation (i.e., PPL decreases when simplification perturbation % increases, signifying simpler text). Different metrics are sensitive to the other three criteria. QAEval, for example, demonstrates sensitivity to the antonym verb swap, entity swap, and sentence negation faithfulness perturbations. For reference, the median reported improvements in ACL'22 summarization and generation papers are ROUGE (+0.89), BLEU (+0.69), METEOR (+0.50), SARI (+1.71), BERTScore (+0.55), and PPL (-2.06) (details in App. D). For estimated score changes corresponding to full perturbation (100%), refer to App. Table 7.

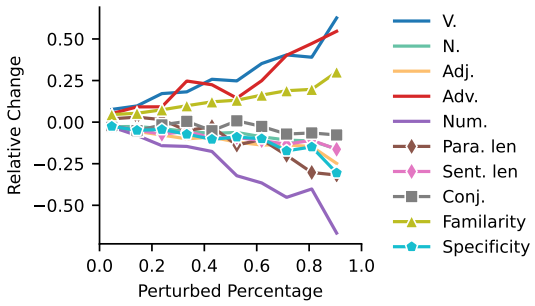

Figure 2: Relative change in lexical features with respect to the unperturbed state (0%). This change is the ratio of the deviation from the lexical feature count at the current perturbation percentage to the lexical feature count in the unperturbed state. Different markers represent lexical feature categories.

simplification, confirming trends observed in previous studies (Kauchak et al., 2014; Leroy et al., 2018; Kauchak et al., 2017; Mukherjee et al., 2017). As simplification increases, paragraph and sentence lengths decrease, while the presence of common words and verbs rises, and the occurrence of nouns, adjectives and term specificity declines. Although prior research emphasizes the importance of conjunctions for comprehension (Kauchak et al., 2017), our perturbation reveals a reduction rather than an increase in conjunctions as texts become simpler. Overall, these trends demonstrate that lexical features are valuable indicators for assessing text simplification.

LLM prompt-based evaluations do not distinguish between PLS criteria. Prompt-based evaluations are not sensitive to simplification perturbations, and in most cases, do not distinguish between the four criteria when scoring summaries (Figure 8). Despite findings from Luo et al. (2023) showing agreement between ChatGPT scores and human ratings, our results suggest that the capacity of LLMs for generative text evaluation warrants further examination. We also note that the reference-free and reference-provided settings yield very different scores along all four criteria, indicating that scores produced with this method are difficult to compare across settings and datasets. Detailed results are provided in App. E.

POMME is sensitive to simplification perturbations. In Table 7, we observe that POMME is significantly correlated with the simplification per-

	BioMedLM-PPL			T5-PPL			POMME		
Datasets	**Source**	**Target**	**Δ (↑)**	**Source**	**Target**	**Δ (↓)**	**Source**	**Target**	**Δ (↑)**
CELLS	-0.36	0.36	**0.72**	0.52	-0.52	**-1.04**	-0.88	0.88	**1.76**
PLABA	-0.79	-0.14	**0.65**	0.29	0.31	0.02	-1.08	-0.45	**0.63**
MSD	3.30	3.30	0.0	-1.89	-1.94	**-0.05**	5.19	5.24	**0.05**
WikiSimple	1.28	2.47	**1.19**	-1.12	-3.23	**-2.11**	2.40	5.70	**3.30**

Table 3: BioMedLM-PPL, T5-PPL and POMME values for existing simplification datasets, comparing source (complex version) and target (simple version). A higher POMME value indicates a higher degree of text simplification. CELLS dataset functions as the reference in POMME computations. The difference, denoted by Δ, is calculated by subtracting the source score from the target score. **Bolded** values indicates statistical significance in the correct direction with Bonferroni-Holm correction for multiple hypothesis testing (Holm, 1979) (i.e., target is simpler than source).

	CELLS / APPLS		PLABA	
Perturb%	**PPL**	**POMME**	**PPL**	**POMME**
20%	**-44.81**	-5.32	3.06	-0.38
40%	**-17.52**	-0.88	5.01	**0.69**
60%	**-14.14**	-0.29	4.59	**0.83**
80%	**-11.55**	0.27	3.38	**0.76**
100%	**-15.45**	**0.42**	1.80	**0.68**

Table 4: The difference in scores between a particular level of perturbation and the unperturbed baseline. **Bolded** values indicates statistical significance in the correct direction with Bonferroni-Holm correction for multiple hypothesis testing (Holm, 1979).

turbation, with higher POMME scores corresponding to greater simplification. We further validate the sensitivity of POMME in three additional text simplification datasets: MSD (Cao et al., 2020), WikiSimple (Woodsend and Lapata, 2011), and PLABA (Attal et al., 2023). Examination of the target texts in Table 3 shows consistently higher POMME values compared to the source texts (Δ is positive), indicating the target texts are simpler. We also present the PPL scores as computed by the in- and out-of-domain LMs. The inconsistency of single model PPLs is evident. For instance, the BioMedLM-PPLΔ for MSD is 0.0 and the T5-PPLΔ for PLABA is 0.02, which suggests that the source texts (scientific abstracts) are simpler or as simple as the targets (plain language summaries).

Using a fixed reference dataset to compute the mean and standard deviation of source and target perplexity enables cross-dataset comparisons of POMME. For example, both MSD and WikiSimple appear considerably simpler than CELLS based on POMME, which is consistent with the content of these datasets—MSD has sentence-level text that is simpler than CELLS' paragraphs, and WikiSimple, which is derived from English and Simple Wikipedias, contains mostly plain language.

Of these validation datasets, PLABA is also in the biomedical domain and contains manually-aligned sentences between the scientific abstracts and human-written plain text. To validate our simplification perturbation in the presence of gold-aligned text, we apply simplification perturbations to PLABA by substituting abstract sentences with their plain text counterparts. As shown in Table 4, GPT2-PPL is insensitive to these perturbations, potentially due to the comparable text lengths of the scientific abstracts and plain text in PLABA. In contrast, POMME demonstrates a consistent response to perturbations, yielding higher scores for more extensively perturbed text.

9 Discussion & Conclusion

Recent advances in NLP point to the possibility of automated plain language summarization (PLS); however, the multifaceted nature of PLS has complicated efforts to define useful evaluation criteria. We introduce the first—to our knowledge—meta-evaluation testbed, APPLS, for evaluating PLS metrics. APPLS applies controlled text perturbations to an existing PLS dataset. Each perturbation is associated with a criteria for PLS: informativeness, coherence, faithfulness, and simplification.

Using APPLS, we find that while some metrics effectively capture informativeness, faithfulness, and coherence, they face challenges in assessing simplification. Most metrics decreased, rather than increased, with further simplification perturbations. The one metric sensitive to simplification, GPT-2 perplexity, exhibited inconsistent sensitivity in other text simplification datasets.

In response, we propose a novel metric, POMME, to address the shortcomings of current metrics in capturing simplification for PLS. POMME uses

normalized perplexity differences between an in-domain and out-of-domain language model. POMME maintains the desirable qualities of language model perplexity we observed in our analysis while being robust and comparable across datasets.

A major benefit to our testbed and metric are their extensiblity. APPLS requires only paired source and target documents (e.g., scientific abstracts and their plain language summaries). Using the perturbation pipeline, APPLS can turn any PLS dataset into a granular meta-evaluation testbed. Similarly, POMME only requires two language models, representing in- and out-of-domain models. Our testbed and metric lay the groundwork for further advancements in automated PLS, with the hope of facilitating more impactful, accessible, and equitable scientific communication.

Limitations

Our perturbations use synthetic data to simulate real-world textual phenomenon seen in PLS. Although our approach is informed by theory and provides valuable insights into metric behavior, further exploration of more sophisticated methods to better simulate changes in these criteria is warranted. This is especially true for aligning sentences between scientific abstracts and plain language summaries, as we observed in our human evaluation that our sentence alignment algorithm led to many partial or incorrect matches. In future iterations, we plan to explore improved sentence alignment algorithms or the utilization of larger manually aligned datasets, such as PLABA (Attal et al., 2023).

We have also focused our analysis on commonly used metrics reported in prior work on simplification, summarization, and generation. Investigating the performance of metrics not included in this work, as well as the generalizability of our methods to meta-evaluation for other generative NLP tasks, is a future goal.

Acknowledgements

This research was supported in part by US National Library of Medicine [grant number R21LM013934].

References

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022a. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Regina Barzilay and Noemie Elhadad. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan R Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, Yee Fan Tan, et al. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. Pubmedgpt 2.7b.

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. *arXiv preprint arXiv:2005.00701*.

Yanran Chen and Steffen Eger. 2022. Menli: Robust evaluation metrics from natural language inference. *arXiv preprint arXiv:2208.07316*.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Daniel Deutsch and Dan Roth. 2022. Benchmarking answer verification methods for question answering-based summarization evaluation metrics. *arXiv preprint arXiv:2204.10206*.

Ashwin Devaraj, Iain Marshall, Byron C Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984.

Ashwin Devaraj, William Sheffield, Byron C. Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2022:7331–7345.

McNamara DS. 2001. Reading both high-coherence and low-coherence texts: effects of text sequence and prior knowledge. *Can J Exp Psychol*, 55(1):51–62.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2020. Go figure: A meta evaluation of factuality in summarization. *arXiv preprint arXiv:2010.12834*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2023. Domain-driven and discourse-guided scientific summarisation. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, pages 361–376. Springer.

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

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2022. Cells: A parallel corpus for biomedical lay language generation. *arXiv preprint arXiv:2211.03818*.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Zellig Harris, Michael Gottfried, Thomas Ryckman, Anne Daladier, and Paul Mattick. 2012. *The form of information in science: analysis of an immunology sublanguage*, volume 104. Springer Science & Business Media.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2022. On the blind spots of model-based evaluation metrics for text generation. *arXiv preprint arXiv:2212.10020*.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Margaret Holmes-Rovner, Sue Stableford, Angela Fagerlin, John T Wei, Rodney L Dunn, Janet Ohene-Frempong, Karen Kelly-Blake, and David R Rovner. 2005. Evidence-based patient choice: a prostate cancer decision aid in plain language. *BMC Medical Informatics and Decision Making*, 5(1):1–11.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388.

Raghav Jain, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. A survey on medical document summarization. *arXiv preprint arXiv:2212.01669*.

Shankar Kanthara, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*.

David Kauchak, Gondy Leroy, and Alan Hogue. 2017. Measuring text difficulty using parse-tree frequency. *Journal of the Association for Information Science and Technology*, 68(9):2088–2100.

David Kauchak, Obay Mouradi, Christopher Pentoney, and Gondy Leroy. 2014. Text simplification tools: Using machine learning to discover features that identify difficult text. In *2014 47th Hawaii international conference on system sciences*, pages 2616–2625. IEEE.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. In *European Chapter of the Association for Computational Linguistics*.

Lauren M Kuehne and Julian D Olden. 2015. Lay summaries needed to enhance science communication. *Proceedings of the National Academy of Sciences*, 112(12):3585–3586.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. *arXiv preprint arXiv:2107.03444*.

Gregoire Leroy, Emma L Carroll, Mike W Bruford, J Andrew DeWoody, Allan Strand, Lisette Waits, and Jinliang Wang. 2018. Next-generation metrics for monitoring genetic erosion within populations of conservation concern. *Evolutionary Applications*, 11(7):1066–1083.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Junru Lu, Jiazheng Li, Byron C. Wallace, Yulan He, and Gabriele Pergola. 2023. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. In *Findings*.

Junyu Luo, Junxian Lin, Chi Lin, Cao Xiao, Xinning Gui, and Fenglong Ma. 2022. Benchmarking automated clinical language simplification: Dataset, algorithm, and evaluation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3550–3562.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2022. Lens: A learnable evaluation metric for text simplification. *arXiv preprint arXiv:2212.09739*.

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

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Partha Mukherjee, Gondy Leroy, David Kauchak, Brianda Armenta Navarrete, Damian Y Diaz, and Sonia Colina. 2017. The role of surface, semantic and grammatical features on simplification of spanish medical texts: A user study. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1322. American Medical Informatics Association.

Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2022. Principled paraphrase generation with parallel corpora. *arXiv preprint arXiv:2205.12213*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Nikhil Pattisapu, Nishant Prabhu, Smriti Bhati, and Vasudeva Varma. 2020. Leveraging social media for medical text simplification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 851–860.

Nicole Pitcher, Denise Mitchell, and Carolyn Hughes. 2022. Template and guidance for writing a cochrane plain language summary.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Conference on Empirical Methods in Natural Language Processing*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184.

Ananya B Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M Khapra. 2021. Perturbation checklists for evaluating nlg evaluation metrics. *arXiv preprint arXiv:2109.05771*.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Marika Salo, Helena Haapio, and Stefania Passera. 2016. Putting financial regulation to work: Using simplification and visualization for consumer-friendly information. In *Networks. Proceedings of the 19th International Legal Informatics Symposium IRIS*, pages 399–406.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Sara Botelho Silveira and António Branco. 2012. Enhancing multi-document summaries with sentence simplification. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer ....

Neha Srikanth and Junyi Jessy Li. 2020. Elaborative simplification: Content addition and explanation generation in text simplification. *arXiv preprint arXiv:2010.10035*.

Marlene Stoll, Martin Kerwer, Klaus Lieb, and Anita Chasiotis. 2022. Plain language summaries: A systematic review of theory, guidelines and empirical research. *Plos one*, 17(6):e0268789.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8918–8927.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*.

John Talbot and Jeffrey K Aronson. 2011. *Stephens' detection and evaluation of adverse drug reactions: principles and practice*. John Wiley & Sons.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605.

Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2022. Perplexity from plm is unreliable for evaluating text quality. *arXiv preprint arXiv:2210.05892*.

Kristian Woodsend and Mirella Lapata. 2011. Wikisimple: Automatic simplification of wikipedia articles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 927–932.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. *arXiv preprint arXiv:2203.12990*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yingxiu Zhao, Zhiliang Tian, Huaxiu Yao, Yinhe Zheng, Dongkyu Lee, Yiping Song, Jian Sun, and Nevin L Zhang. 2022. Improving meta-learning for low-resource text classification and generation via memory imitation. *arXiv preprint arXiv:2203.11670*.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9709–9716.

A GPT Configurations for Simplification Perturbation

We use GPT-3 (*text-davinci-003*) for text simplification. The generation process is configured with a temperature parameter of 0.7, a maximum length of 1000, and a penalty value of 0. For each input, the top-ranked text is selected as the GPT-simplified output.

B Round-trip translation for oracle extractive hypothesis

We use round-trip translation to introduce lexical variation into our oracle extractive summaries. This is important when computing metrics such as SARI, which exhibit degenerate behavior when the hypothesis is an extractive subset of the source. We examine two languages for round-trip translation: German and Russian. By employing the BLEU score as a performance metric for the round-trip generated text relative to the original source, we find that the English-German-English (en-de-en) translation sequence yields superior BLEU scores (Figure 3), and therefore, select the en-de-en sequence to produce the oracle extractive hypothesis for our testbed.

C Details of human evaluation

To validate the quality of oracle extractive hypotheses and GPT-simplified summaries, we randomly select 100 summary pairs from each corpus for human evaluation. Each pair in the oracle extractive hypotheses consists of an oracle extractive sentence and its respective en-de-en round-trip-translation sentence. Similarly, each pair in the GPT-simplified summaries contains a hypothesis chunk along with its corresponding GPT-simplified summary chunk.

Each pair is reviewed by two independent annotators. Annotators were hired through UpWork and have Bachelors and Doctorate degrees in the biological sciences. In the evaluation, the text pairs are labeled as Text A and Text B, without any indication that either text is generated. The annotators are first asked to assess whether the content of Text A matches the content of Text B, where a match is defined as containing the same relation tuples. If the texts match, the annotators further evaluate Text B in relation to Text A, assessing whether Text B encapsulates key points (informativeness), is more comprehensible (simplification), maintains factual integrity (faithfulness), and exhibits a well-structured layout (coherence). All facets are assessed using a 1-5 Likert scale (1-strongly disagree, 5-strongly agree). Representative questions can be found in Figure 4. This research activity is exempt from institutional IRB review.

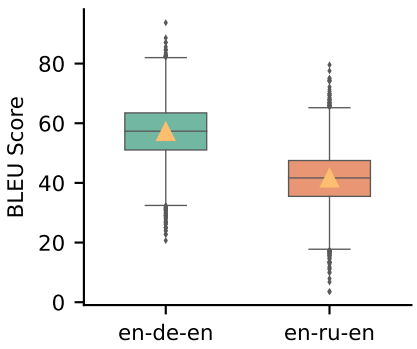

Figure 3: BLEU scores of round-trip translation for English-German-English (en-de-en) and English-Russian-English (en-ru-en) in CELLS oracle extractive hypotheses.

For round-trip translation, annotators reported that translated sentences maintained or increased simplicity (95.5%), informativeness (97.5%), faithfulness (82.5%), and coherence (98.5%) in the large majority of cases. The inter-rater agreement, measured by Cohen's Kappa, is 0.46, implying moderate agreement (Artstein and Poesio, 2008). For GPT-simplified summaries, one annotator completed all annotations. Of the 100 examples, 46 were labeled unmatched, indicating partial or no match in content. For the 54 examples that matched well, the GPT-simplified text was generally found to be much simpler (74% strong agree, 17% agree), while maintaining informativeness (100%), faithfulness (98%), and coherence (100%). These results emphasize the need for further work on sentence-level alignment. As such, we note that partial simplification perturbations in our dataset potentially introduce loss of information. To offset this issue, we also evaluate simplification perturbations on a separate but smaller dataset with gold sentence alignments, PLABA. Complete evaluation results can be found in Table 5.

D Empirical Study of Evaluation Metrics Reported in ACL 2022 Publications

Our study undertakes a comprehensive analysis of scores reported in the long papers of ACL 2022 to identify the most prevalently reported metrics in summarization and simplification tasks. We pri-

Type	Unmatched	Criteria	Str. Agree	Agree	Neutral	Disagree	Str. Disagree
Round Trip Translation	1	Simplification	12	27	152	7	0
		Informativeness	188	4	3	4	0
		Faithfulness	155	6	4	20	14
		Coherence	30	11	156	2	0
GPT Simplification	46	Simplification	40	9	2	2	1
		Informativeness	49	3	2	0	0
		Faithfulness	49	3	1	1	0
		Coherence	50	2	2	0	0

Table 5: Counts of human evaluation ratings on each matched sentence for each criteria. For round trip translation, there are a total of 200 ratings (2 annotators rating 100 sentences each). For GPT simplification, 1 annotator rated 100 sentences. Of these, 46 of 100 sentences had imperfect alignment (unmatched), so we report on the 54 ratings for aligned sentences. Overall, we see that round trip translation maintains strong faithfulness to the original, does not remove important information, and remains equally simple and coherent (shown by a majority of neutral ratings for the simplification and coherence criteria). For GPT simplification, in cases where sentences aligned, we see that the simplification perturbation leads to substantially more simple text, while also maintaining faithfulness and informativeness.

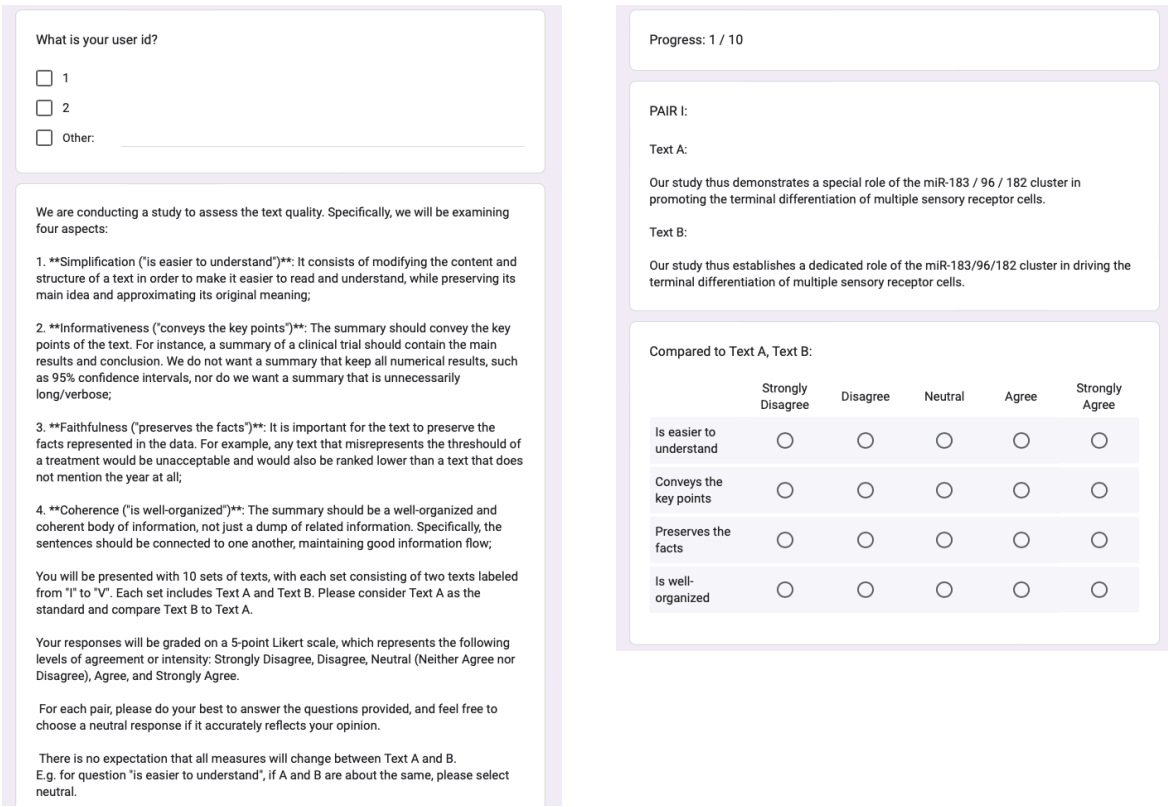

Figure 4: An example human evaluation task for assessing GPT-simplified summary quality.

marily concentrate on tasks related to generation, summarization, and simplification. Our inclusion criteria are: 1) long papers with 'generat,' 'summar,' or 'simpl' in the title; and 2) papers that report scores for both the current model and at least one baseline model in the main text. We exclude scores from ablation studies.

Of the 601 long papers accepted to ACL 2022, 109 satisfy our inclusion criteria, which we categorize into 31 summarization and 78 generation papers, with no qualified papers related to simplification tasks. Considering the significance of simplification in PLS, we expanded our search to all ACL 2022 papers, including long, short, system demonstration, and findings papers. This led to the identification of 2 out of 22 papers with 'simpl' in the title that reported SARI scores. As illustrated in Figure 5, the five most frequently reported automated evaluation metrics are ROUGE, BLEU, GPT-PPL, METEOR, and BERTScore.

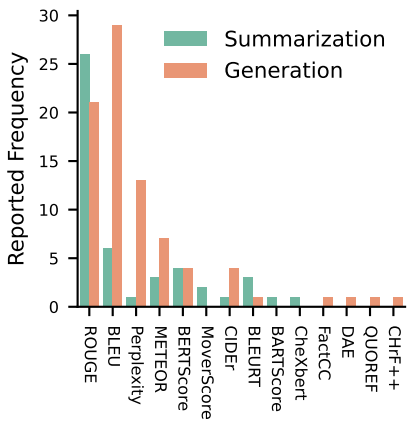

Figure 5: Most common evaluation metrics reported in ACL'22 summarization and generation long papers.

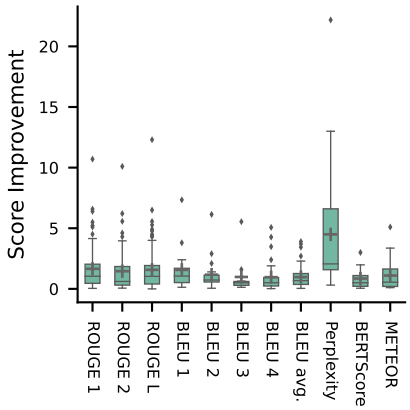

Figure 6: Distributions of reported metric improvements over baseline (absolute value) reported in ACL'22 summarization and generation long papers.

This investigation provides insight into the current adoption of evaluation metrics in natural language generation, summarization, and simplification tasks. We observe that a majority of papers employ the same metrics across these tasks, and the reported improvements are often relatively small compared to the overall ranges for each measure. We also underscore the difficulty of interpreting changes in some of these metrics, especially model-based metrics, which lack grounding to lexical differences in text such as n-gram overlap.

By presenting the reported score differences from ACL papers, we hope to contextualize the metric changes observed through testing in our meta-evaluation testbed. Median reported improvements for the most commonly reported metrics and SARI are: ROUGE (+0.89), BLEU (+0.69), PPL (-2.06), METEOR (+0.50), BERTScore (+0.55), and SARI (+1.71), as shown in Figure 6.

	Ref. Free	Ref. Provided
Informativeness (↓)		
Delete sentence	**-61.09**	**-22.33**
Add out-of-domain sent	**-19.91**	**-34.88**
Add in-domain sent	**-7.32**	-6.99
Add definition (↑)	-0.6	12.64
Simplification (↑)	-15.98	-11.46
Coherence (↓)	-0.17	-1.06
Faithfulness (↓)		
Number swaps	0.45	-0.44
Entity swaps	**-5.48**	-7.24
Synonyms verb swaps	-4.02	-10.54
Antonyms verb swaps	**-6.05**	**-8.27**
Negate sentence	**-18.59**	**-21.94**

Table 6: Overall score derived from the prompt-based evaluation for two settings: reference-free and reference-provided. **Bolded** values indicate statistical significance in the correct direction with Bonferroni-Holm correction for multiple hypothesis testing (Holm, 1979).

E LLM Prompt-Based Evaluation

We use GPT-3 for LLM evaluation. The generation process is configured with a temperature parameter of 0, a maximum length of 100, and a penalty value of 0. For each input, the top-ranked text is selected as the GPT-simplified output. Example prompts used for evaluation are provided in Figure 7.

Figure 8 shows the results for GPT-3 LLM evaluation, for both the reference-free and reference-provided settings. Though the evaluation is sensitive to some perturbations (deletion, addition, negation), it is insensitive to other perturbations (coherence, swaps) and sensitive to simplification in the inverse direction as would be expected (simplification score drops when more source text is replaced by simplified text). Additionally, the LLM evaluation is generally unable to distinguish between the four criteria, as most perturbations lead to the same score trends for simplification, coherence, faithfulness, and to a lesser degree informativeness. These patterns are similar to those observed in the overall score, indicating that the LLM evaluation as performed is not useful for providing facet-based judgments.

We also observe that in the reference-provided setting, scores for some perturbations are much higher (e.g., deletion) while others are much lower (e.g., add out-of-domain) than in the reference-free setting. The lack of a reference point or a way to normalize these scores makes it impossible to compare them across settings or datasets.

a. Reference Free Prompt:
Imagine you are a human annotator now. You will evaluate the quality of generated plain lanugage summary written for a scientific literature abstract. **Please follow these steps:**
1. Carefully read the scientific literature abstract, and be aware of the information it contains.
2. Read the proposed generated plain langauge summary.
3. Compared to the scientific abstract, rate the summary on four dimensions: informativeness, simplification, coherence, and faithfulness. Assign a score for each aspect and provide an overall score. You should rate on a scale from 0 (worst) to 100 (best).
4. You do not need to explain the reason. Only provide the scores.

Definitions are as follows:
-**Informativeness**: measures the extent to which a plain language summary encapsulates essential elements such as methodologies, primary findings, and conclusions from the original scientific text. An informative summary efficiently conveys the central message of the source material, avoiding the exclusion of crucial details or the introduction of hallucinations (i.e., information present in the summary but absent in the scientific text), both of which could impair reader comprehension.
-**Simplification**: encompasses the rendering of information into a form that non-expert audiences can readily interpret and understand. This criterion prioritizes the use of simple vocabulary, casual language, and concise sentences that minimize excessive jargon and technical terminology unfamiliar to a lay audience.
-**Coherence**: pertains to the logical arrangement of a plain language summary. A coherent summary guarantees an unambiguous and steady progression of ideas, offering information in a well-ordered fashion that facilitates ease of comprehension for the reader. We conjecture that the original sentence order reflects optimal coherence.
-**Faithfulness**: denotes the extent to which the plain language summary aligns factually with the source scientific text, in terms of its findings, methods, and claims. A faithful summary should not substitute information or introduce errors, misconceptions, and inaccuracies, which can misguide the reader or misrepresent the original author's intent. Faithfulness emphasizes the factual alignment of the summary with the source text, while informativeness gauges the completeness and efficiency of the summary in conveying key elements.

The scientific abstract and the generated plain language summary are given below:
Scientific abstract: {}
Generated plain language sumamry:{}

b. Reference Provided Promt:
Imagine you are a human annotator now. You will evaluate the quality of generated summary written for a scientific literature abstract. **Please follow these steps:**
1. Carefully read the scientific abstract and plain language summary written by human, and be aware of the information it contains.
2. Read the proposed genereated summary.
3. Compared to the scientific abstract and human-written plain language summry, rate the generated summary on four dimensions: informativeness, simplification, coherence, and faithfulness. Assign a score for each aspect and provide an overall score. You should rate on a scale from 0 (worst) to 100 (best).
4. You do not need to explain the reason. Only provide the scores.

Definitions are as follows:
-**Informativeness**: measures the extent to which a plain language summary encapsulates essential elements such as methodologies, primary findings, and conclusions from the original scientific text. An informative summary efficiently conveys the central message of the source material, avoiding the exclusion of crucial details or the introduction of hallucinations (i.e., information present in the summary but absent in the scientific text), both of which could impair reader comprehension.
-**Simplification**: encompasses the rendering of information into a form that non-expert audiences can readily interpret and understand. This criterion prioritizes the use of simple vocabulary, casual language, and concise sentences that minimize excessive jargon and technical terminology unfamiliar to a lay audience.
-**Coherence**: pertains to the logical arrangement of a plain language summary. A coherent summary guarantees an unambiguous and steady progression of ideas, offering information in a well-ordered fashion that facilitates ease of comprehension for the reader. We conjecture that the original sentence order reflects optimal coherence.
-**Faithfulness**: denotes the extent to which the plain language summary aligns factually with the source scientific text, in terms of its findings, methods, and claims. A faithful summary should not substitute information or introduce errors, misconceptions, and inaccuracies, which can misguide the reader or misrepresent the original author's intent. Faithfulness emphasizes the factual alignment of the summary with the source text, while informativeness gauges the completeness and efficiency of the summary in conveying key elements.

The scientific abstract, plain language summary, and generated summary are given below:
Scientific abstract: {}
Plain language summary: {}
Generated summary: {}

Figure 7: Prompts used for LLM evaluation. (a): Reference-free; (b) Reference-provided.

F Existing metrics performance

We present the estimated metric changes resulting from 100% perturbations in the APPLS testbed. Table 7 shows the slope coefficients associated with 8 automated evaluation metrics and POMME. Table 8 presents the slope coefficients for the simplification-associated lexical features.

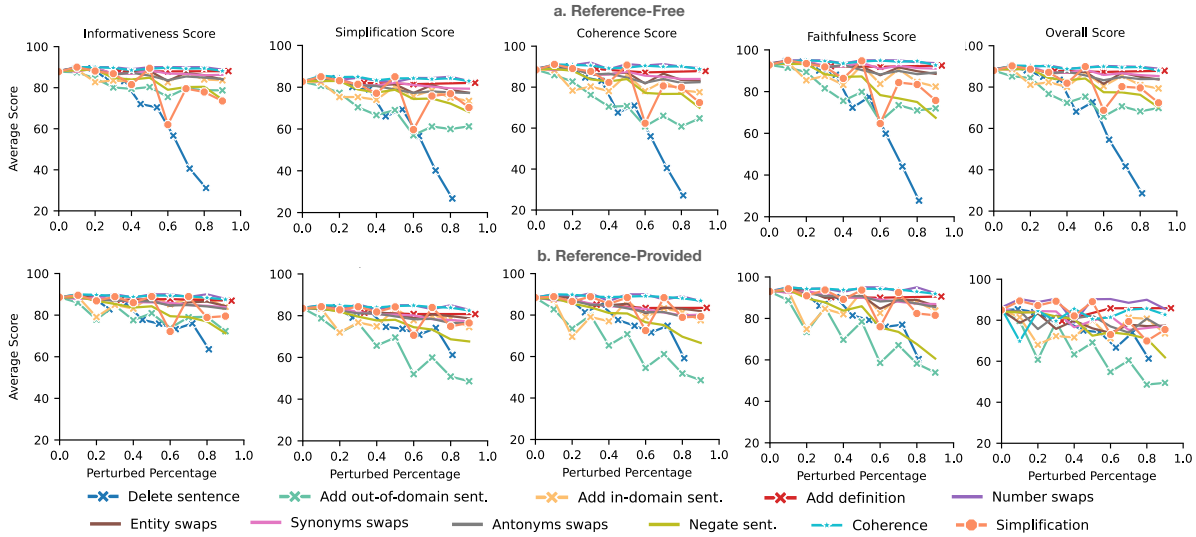

Figure 8: Prompt-based evaluation scores for four criteria - informativeness, simplification, coherence, and faithfulness - along with an overall score. (a): Reference free; (b) Reference provided. Notably, prompt-based scores exhibit a reverse correlation with simplification perturbation (i.e., scores diminish as text simplifies) and demonstrate insensitivity towards coherence and faithfulness perturbations, except in instances of sentence negation.

	Overlap-based				**Model-based**			**QA-based**	**New**
	ROUGE	BLEU	METEOR	SARI	BERTSc.	PPL(↓)	LENS	QAEval	POMME
Informativeness (↓)									
Delete sentence	**-25.19**	**-11.16**	**-0.18**	**-5.24**	**-16.78**	**82.79**	1.41	**-0.95**	5.89
Add out-of-domain sent	**-5.23**	**-2.07**	0.01	-0.06	**-5.26**	**49.41**	0	**-0.19**	-0.1
Add in-domain sent	**-4.73**	**-1.72**	0.01	-0.01	**-4.79**	**9.50**	0	**-0.22**	**-1.40**
Add definition (↑)	-2.77	-1.49	0.01	0.27	-2.51	-1.96	-0.71	-0.22	-1.45
Simplification (↑)	-11.31	-5.84	-0.07	-1.83	-8.32	**-15.45**	0.10	-0.55	**1.02**
Coherence (↓)	**-1.28**	0.01	0	0.05	**-0.64**	**5.37**	**-1.75**	-0.03	1.0
Faithfulness (↓)									
Number swaps	-0.70	0.49	0.01	0.46	-0.39	-2.96	3.04	-0.09	**-1.21**
Entity swaps	**-1.79**	-0.93	**-0.01**	0.09	**-1.33**	**8.96**	-0.01	**-0.41**	0.72
Synonyms verb swaps	-1.80	-0.90	-0.01	0.87	-1.45	43.63	0	-0.05	1.13
Antonyms verb swaps	**-1.22**	-0.62	-0.01	0.49	**-1.31**	**25.74**	0.03	**-0.28**	0.99
Negate sentence	-0.86	-0.30	0	0.33	**-0.84**	**4.84**	0	**-0.67**	0.22
ACL'22 improvement	0.89	0.69	0.50	1.71	0.55	-2.06	-	-	-

Table 7: Slope coefficients of linear regression between perturbation percentage and automated evaluation metrics. The slopes represent the estimated change in evaluation score for full perturbation. **Bolded** values indicates statistical significance in the correct direction with Bonferroni-Holm correction for multiple hypothesis testing (Holm, 1979). We also provide the median reported score improvements from ACL'22 papers on text generation and summarization to help contextualize these deltas.

	Length		**Familiarity**	**Specificity**	**Trans.**	**Function Words**				
	Para.	Sent.	Common Words	MeSH Terms	Conj.	V.	N.	Adj.	Adv.	Num.
Simplification	-41.63	-5.22	0.12	-0.90	-0.01	0.06	-0.05	-0.04	0	0

Table 8: Slope coefficients of linear regression between perturbation percentage and lexical feature-based metrics. The slopes represent the estimated change in evaluation score for a 100% perturbation. All values are statistically significant with Bonferroni-Holm correction for multiple hypothesis testing (Holm, 1979).