APPLS: Evaluating Evaluation Metrics for Plain Language Summarization

Yue Guo[1*] **Tal August**[1] **Gondy Leroy**[2] **Trevor Cohen**[3] **Lucy Lu Wang**[3,4]
[1]University of Illinois Urbana-Champaign [2]University of Arizona
[3]University of Washington [4]Allen Institute for AI
{yueg, taugust}@illinois.edu; {cohenta, lucylw}@uw.edu

Abstract

While there has been significant development of models for Plain Language Summarization (PLS), evaluation remains a challenge. PLS lacks a dedicated assessment metric, and the suitability of text generation evaluation metrics is unclear due to the unique transformations involved (e.g., adding background explanations, removing jargon). To address these questions, our study introduces a granular meta-evaluation testbed, APPLS, designed to evaluate metrics for PLS. We identify four PLS criteria from previous work—informativeness, simplification, coherence, and faithfulness—and define a set of perturbations corresponding to these criteria that sensitive metrics should be able to detect. We apply these perturbations to the texts of two PLS datasets to create our testbed. Using APPLS, we assess performance of 14 metrics, including automated scores, lexical features, and LLM prompt-based evaluations. Our analysis reveals that while some current metrics show sensitivity to specific criteria, no single method captures all four criteria simultaneously. We therefore recommend a suite of automated metrics be used to capture PLS quality along all relevant criteria. This work contributes the first meta-evaluation testbed for PLS and a comprehensive evaluation of existing metrics.[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., 2023; 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) due to 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.

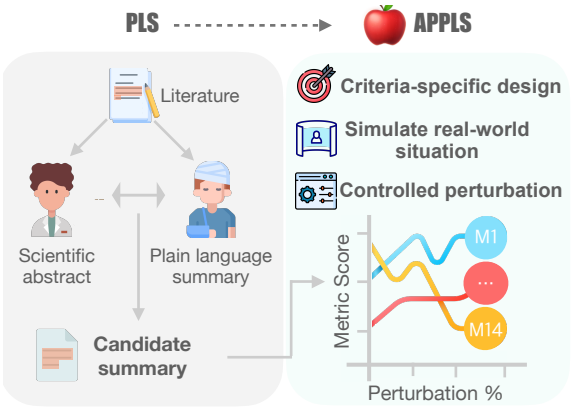

Figure 1: We present APPLS, the first granular testbed for analyzing evaluation metric performance for plain language summarization (PLS). We assess performance of 14 existing metrics, including automated scores, lexical features, and LLM prompt-based evaluations.

Our goal is to assess how well existing metrics capture the multiple criteria of PLS. We identify four criteria, informed by prior work (Pitcher et al., 2022; Ondov et al., 2022; Stoll et al., 2022; Jain et al., 2022), that a PLS metric should be sensitive to: *informativeness*, *simplification*, *coherence*, and *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.[2] By

[*]Work performed while at University of Washington.
[1]APPLS and our evaluation code can be found at https://github.com/LinguisticAnomalies/APPLS.
[2]We acknowledge that introducing any change in text likely affects multiple criteria, though we design our perturbations carefully to try and minimize these impacts.

incrementally introducing perturbations to the texts of two scientific PLS datasets, CELLS (Guo et al., 2024) and PLABA (Attal et al., 2023), we create our meta-evaluation testbed APPLS.

We analyze 14 metrics using APPLS, including the most widely used metrics in text simplification and summmarization literature, and recently-proposed prompt-based methods (Gao et al., 2023; Luo et al., 2023). We find that established metrics like ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), and QAEval (Deutsch et al., 2021) do not capture simplification and are inconsistent at capturing perturbations in informativeness, coherence, and faithfulness; SARI score (Xu et al., 2016), explicitly crafted for text simplification, is the only automated score that displays sensitivity towards simplification perturbations but not to other perturbations. LLM prompt-based evaluations are effective for assessing informativeness, faithfulness, and simplification, but not for coherence. Our analysis suggests that a single overall score cannot simultaneously respond to all four criteria.

Our main contributions are as follows:

- We present APPLS, the first granular testbed for analyzing evaluation metric performance for plain language summarization; the testbed is created by applying 11 perturbations along four dimensions to two scientific PLS datasets (§3, 4);
- We conduct a thorough analysis of 14 existing evaluation metrics (including automated metrics, lexical features, and LLM prompting methods), demonstrating mixed effectiveness in evaluating informativeness, coherence, faithfulness, and simplification (§5, 6);
- Based on our findings, we recommend an evaluation strategy for PLS that combines multiple automated metrics able to capture differences along all relevant dimensions.

2 Related Work

Limitations of Existing Metrics The primary approach for evaluating plain language summaries adopts evaluation metrics for summarization and simplification, coupled with human evaluation (Jain et al., 2021; Ondov et al., 2022). While ROUGE (Lin, 2004) and BLEU (Sulem et al., 2018) are frequently employed in PLS assessment, their efficacy is limited due to the reliance on high-quality reference summaries, which are often challenging to obtain for PLS or may not exist at all. Further, these metrics struggle to accurately identify hallucinations, especially crucial for PLS in the health domain to accurately inform health decisions (Wallace et al., 2021; Pagnoni et al., 2021; Wang et al., 2023). Though human evaluation offers thorough assessment (Hardy et al., 2019), the high costs and time needed impede scalability for larger datasets. While recent progress in prompt-based evaluation shows potential for assessing factuality (Luo et al., 2023) and summarization quality (Gao et al., 2023), their efficacy for PLS is yet to be validated. Our work aims to fill these gaps through a systematic examination of these metrics within the PLS context.

Dataset	Version	Word	Sentence
CELLS (n=6,311)	Abstract (*src.*)	$283_{\pm 132}$	$11_{\pm 6}$
	PLS (*tgt.*)	$178_{\pm 74}$	$7_{\pm 3}$
	Candidate summary	$134_{\pm 58}$	$5_{\pm 2}$
	GPT-simplified	$130_{\pm 34}$	$6_{\pm 2}$
PLABA (n=750)	Abstract (*src.*)	$240_{\pm 95}$	$10_{\pm 4}$
	Adaptation (*tgt.*)	$244_{\pm 95}$	$12_{\pm 5}$

Table 1: Diagnostic datasets statistics (mean±std).

Robust Analysis with Synthetic Data Synthetic data has been widely used in NLP to evaluate metrics, for tasks such as text generation (He et al., 2023; Sai et al., 2021), natural language inference (Chen and Eger, 2023; McCoy et al., 2019), question answering (Ribeiro et al., 2019), and reading comprehension (Sugawara et al., 2020). Yet, no prior work has focused on the PLS task or incorporated simplification into their benchmarks. Additionally, previous studies lack granular analyses to capture nuanced relationships between text changes and score changes. Our research endeavors to bridge these gaps by crafting perturbations that mirror real-world errors found in PLS.

3 PLS Evaluation Criteria

We identify four criteria that an effective PLS evaluation metric should be sensitive to based on both abstractive summarization (Sai et al., 2022; Koto et al., 2022) and plain language summarization paradigms (Pitcher et al., 2022; Ondov et al., 2022; Stoll et al., 2022; Jain et al., 2022). As in Gabriel et al. (2021), we define *sensitivity* as being correlated in the expected direction with the amount of change in that criteria.

- **Informativeness** measures the extent to which the plain language summary covers essential information from the source text (e.g., methods, main findings) and incorporates relevant back-

Table 2: Example perturbations for criteria in APPLS. Original text comes from the CELLS (Guo et al., 2024).

ground information (Smith et al., 2021; Beck et al., 1991).

- Simplification describes the degree to which information is conveyed in a form that non-expert audiences can readily understand. It is distinct from informativeness because it focuses on surface-level changes (e.g., shorter sentences) but not other changes relevant to content (e.g., background explanation).
- Coherence describes the logical arrangement of a plain language summary.
- Faithfulness denotes how well the summary aligns factually with the source text.

4 Constructing the APPLS Testbed

To assess metric sensitivity, we develop perturbations along each evaluation criteria dimension. We implement our perturbations in two large-scale PLS datasets, described in §4.1. We follow with a discussion of how perturbations are incorporated into these datasets and our approach for managing perturbation magnitude (§4.2) and validating perturbation quality (§4.3).

4.1 Diagnostic datasets

For our experiments, we use the CELLS (Guo et al., 2024) and PLABA (Attal et al., 2023) datasets. CELLS (Guo et al., 2024) is a parallel corpus of scientific abstracts (source texts) and their corresponding plain language summaries (target texts), which are written by the abstract authors or by other domain experts. CELLS aggregates papers from 12 biomedical journals, representing a diverse set of topics and summaries, and serves as the primary dataset in our testbed.

The PLABA (Attal et al., 2023) dataset includes expert-modified biomedical abstracts, simplified to improve understanding of health-related content. PLABA includes sentence-level alignments, which are useful for controlled perturbations. However, we did not select PLABA as the primary dataset due to its reliance on relatively contrived simplifications, which lack generalizability to other PLS datasets; these modifications include rule-based adjustments such as lexical simplification, shifting from passive to active voice, and segmenting long sentences. This results in high n-gram overlap between sources and target summaries, which is unrealistic and does not reflect PLS in the real world.

Therefore, PLABA serves as an auxiliary dataset to CELLS, helping to address its limitations discussed in Sections §4.2 and §4.3. Full results using PLABA as the diagnostic dataset are in App. H.

4.2 Applying perturbations to datasets

Illustrative examples of all perturbations are shown in Table 2. For the APPLS testbed, we propose and apply perturbations to a candidate summary, which is an extractive summary constructed in the oracle setting with additional lexical variation introduced through round-trip translation (Ormazabal et al., 2022) (Illustration in App. Figure 6).[3] We do not perturb the target text directly since the resulting candidate summary would be overly similar to the target, which would be unrealistic.

For CELLS, an extractive summary is created by selecting the set of source sentences yielding the highest ROUGE-L score when compared to the target summary, and this summary is then round-trip translated through German to derive the candidate summary. To identify the optimal extractive summary, we exhaustively evaluate all possible subsets of sentences from the source document while preserving their original order, ensuring the highest ROUGE-L score is achieved.[4] The PLABA dataset already contains sentence alignments, with sources and targets having similar lengths, so we produce the candidate summary through round-trip translation of the target alone. Details are in App. B.

We apply all perturbations to these candidate summaries as described below, where each perturbation introduces a change (e.g., add/swap sentences) at some magnitude (e.g., replace 50% of sentences). Due to the high costs associated with some of our perturbations (e.g., LLM-based simplification), we restrict our testbed to the test splits of our diagnostic datasets (stats in Table 10).[5] To mitigate the effects of randomness, we use two random seeds to produce all perturbations.

Informativeness

Delete sentences To simulate the loss of relevant information, we delete sentences until a single sentence remains. The magnitude of deletion is the ratio of remaining to original sentences.

Add sentences We insert up to the same number of sentences as in the candidate summary. To simulate out-of-domain hallucinations, we add sentences from ACL papers (Bird et al., 2008). For in-domain hallucinations, we add sentences from Cochrane abstracts.[6] The magnitude of addition is the ratio of added to original sentences.

Add definitions Background explanations are fundamental to PLS and involve adding external content like definitions or examples (Guo et al., 2024; Srikanth and Li, 2021). To simulate these, we add up to three definitions, the average number of nouns explained in CELLS (Guo et al., 2024), i.e., 100% perturbed adds three definitions.

Simplification

Replace sentences For CELLS, we first generate an LLM-simplified summary from the candidate summary. We align sentences between the candidate summary and LLM-simplified summary using the sentence alignment algorithm from Guo et al. (2024). We perturb the text by replacing random sentences from the candidate summary with their corresponding simplifications until full replacement is achieved. We use GPT-4 (Achiam et al., 2023) to generate simplifications due to its accessibility and demonstrated proficiency in text simplification (Lu et al., 2023). To ensure that our findings are not specific to the chosen model, we also generate simplifications and conduct experiments using GPT-3 (Brown et al., 2020), Llama2 (Touvron et al., 2023) and Claude.[7] For PLABA, we perturb text by replacing source sentences with round-trip translated versions of their aligned simplified targets (no LLM is used).

Coherence

Reorder sentences We shuffle sentences in the candidate summary and quantify perturbation percentage in terms of the absolute difference in sentence order between the original and shuffled candidate summaries, e.g., a document with reversed sentence order would be considered 100% perturbed. Details are in App. A.

Faithfulness

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

[3] To achieve the necessary level of control for detecting the sensitivity of automated scores to perturbation modifications, we opt to use an extractive summary instead of a language model to generate the candidate summaries.

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

[5] The CELLS dataset contains 63k pairs; only the test split with 6.3k pairs is used for APPLS construction.

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

[7] https://www.anthropic.com. Access date: 12/04/2023.

Verb swap We introduce two perturbations by substituting verbs with either synonyms or antonyms. An appropriate metric should be less sensitive to synonyms but more sensitive to antonyms.

Entity swap We replace entities using the KBIN method (Wright et al., 2022), which replaces entity spans with related concepts in the UMLS[8] while maximizing NLI contradiction and minimizing LM perplexity. This results in a fluent sentence that varies from the original one.

Negate sentences We negate sentences, and allow up to one negation per sentence.

The perturbation magnitude of number, verb, and entity swaps is determined by comparing the count of altered spans to the total number of eligible spans in the candidate summary. Full perturbation means all eligible spans are swapped.

4.3 Human validation of candidate summaries and LLM simplifications

We validate two design decisions of APPLS that involve other models modifying text—round-trip translation (RTT) for the extractive summary and GPT-4-based simplification perturbations—by conducting human evaluation. We sample 100 pairs each of (i) extractive summaries (pre-RTT) paired with candidate summaries (post-RTT) and (ii) GPT-simplified summaries paired with candidate summaries. Annotators were asked to assess content alignment (defined as having comparable entities and relations between entities) and rate informativeness, simplification, faithfulness, and coherence on 5-point Likert scales. Annotations were performed by two independent annotators, both with doctorates in the biological sciences, who were hired on UpWork and compensated at 21 USD/hr. Each annotator reviewed all sampled pairs for both evaluation tasks. Inter-rater agreement measured by Cohen's Kappa was 0.48, by Spearman rank correlation was 0.58, implying moderate agreement for both tasks (Artstein and Poesio, 2008). Details of the annotation tasks are given in App. C.

Human annotators affirmed that RTT text (candidate summary) retained its informativeness (98%), faithfulness (83%), coherence (100%), and simplicity (96%) compared to the extractive summary. For GPT-simplified sentences, evaluators rated its informativeness (95%), faithfulness (95%), coherence (98%), and simplicity (100%) compared to the candidate summary, with GPT-simplifications consistently rated as more simple than the candidate summary while preserving semantic content. In this context, we report the proportion of annotations equal to or better than neutral for each criterion. While the informativeness and faithfulness of GPT-simplified text are assessed to be very good at the passage level, the alignment algorithm used to produce sentence alignments for the simplification perturbation is imperfect and can introduce some errors. To mitigate the impact of such misalignment on the interpretation of results, we use the PLABA dataset for auxiliary diagnostics because it contains ground truth sentence-level alignments.

[8] https://www.nlm.nih.gov/research/umls/

5 Evaluation Metrics

Our analysis spans 8 established evaluation metrics, including the 5 most commonly reported in ACL'22 summarization/generation papers (empirical results in App. D). We also assess 5 lexical features associated with text simplification (§5.2) and LLM-based evaluations (§5.3).

5.1 Existing automated evaluation metrics

We compute the following 8 automated metrics:

- **Overlap-based metrics** measure n-gram overlap. We report ROUGE (computed as the average of ROUGE-1, ROUGE-2, and ROUGE-L) (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and SARI score (Xu et al., 2016).
- **Model-based metrics** use pretrained models to evaluate text quality. We adopt GPT-PPL, BERT-Score (Zhang et al., 2019), and LENS (Maddela et al., 2023).
- **QA-based metrics** capture content quality using a question-answering approach. We report QAEval (Deutsch et al., 2021) scores here.

Details for all metrics are available in App. E. All metrics assessed require target and generated texts; while SARI and LENS additionally make use of the source texts.

5.2 Lexical features

We also assess lexical features that have been shown to be associated with 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 per-

Figure 2: Average scores of existing metrics for perturbed texts in the CELLS dataset. Scores are averaged in 10 bins by perturbation percentage. Markers denote the de ned criteria associated with that perturbation. 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).

centage of text that is made up of the 1,000 most common English words[9].

- Speci city: Speci city quanti es the level of detail in the text. We use Speciteller (Ko et al., 2019) to compute the domain agnostic speci city of terms in the paragraph.
- Phrase Transitions Conjunctions (e.g., therefore) are important for ow 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.

5.3 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 prompt template from Gao et al. (2023) to have GPT-4 (gpt-4-0613) evaluate each candidate summary on the four PLS criteria and to provide an overall quality score. All scores range from 0 (worst) to 100 (best). We supply de nitions for each criterion in the prompt. We evaluate under three settings: (a) providing a single criterion in the prompt and requesting a score for that criterion; (b) providing all criteria in the prompt and requesting scores for each criterion as well as an overall score; and (c) the same setting as (b) but requiring explanations be generated alongside the provided scores. Model con gurations and prompts are in App. G.

6 Analysis Results

Automated metric responses to perturbations are in Figure 2, responses of lexical features are in Figure 3, and prompt-based evaluation results are shown in Figure 4. All trends are consistent across two random seeds.

To contextualize metric performance in PLS, we survey metric changes reported in ACL'22 papers on text generation and summarization (full results in App. D). 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 11.

Aside from SARI, current metrics exhibit shortcomings in evaluating simplicity. Metrics that are sensitive to simpli cation should consistently distinguish between more and less simpli ed text. SARI is the only automated metric among those we tested that is consistently sensitive to simpli ed text. As shown in Figure 2, metrics that exhibit sensitivity to simpli cation perturbations are GPT-PPL (decreasing as more perturbations are introduced; lower PPL is better) and SARI score. However, in follow-up evaluations with PLABA (shown in App.

[9] https://gist.github.com/deekayen/4148741

Figure 16), we see that GPT-PPL has undesirable sensitivity to text length, as found in prior work (Zhao et al., 2022).

ROUGE, BLEU, METEOR, BERTScore, and QAEval decrease in response to the simplification perturbation. While they show consistent response relative to the degree of perturbation, they are nonetheless not useful for assessing text simplicity. When we report metric changes swapping sources and targets (perturbing simplified texts to increase complexity), these metrics also decrease (App. Figure 13), suggesting that they are sensitive to n-gram changes and not text simplicity. LENS behaves erratically with increasing simplification perturbation percentage, indicating that it is not a good metric for text simplicity.

In addition to using GPT-4 (Achiam et al., 2023) to produced simplified text for the simplification perturbation, we also test three other LLMs: GPT-3 (Brown et al., 2020), Llama 2 (Touvron et al., 2023), and Claude[7]. In Figure 5, we show metric changes to the simplification perturbation generated by all four models. Similar score changes are observed for all models (except GPT-3 for SARI score, which is an outlier), demonstrating that the simplicity perturbation in our testbed is a reasonable and consistent measure of metric response to text simplification, and that SARI is generally able to distinguish between more and less simplified text.

Metrics effectively capture informativeness, coherence, and faithfulness, with room for improvement. For informativeness, ROUGE, BLEU, BERTScore, GPT-PPL, and QAEval are sensitive to information deletion and irrelevant additions, but decrease with the addition of background explanations through keyword definitions. For coherence, BERTScore and LENS excel in detecting perturbations, largely due to their ability to assess between-sentence relationships. BERTScore, GPT-PPL, and QAEval generally perform well for faithfulness-related perturbations, although GPT-PPL and BERTScore are somewhat sensitive to synonym verb swaps instead of antonym verb swaps, which is an undesirable trait. QAEval is best at being unresponsive to synonym verb swaps. Number swaps, however, remain undetected by all metrics. Results in Figure 2

Lexical features are useful measures of text simplicity. Figure 3 illustrates the response of lexical features to degrees of text simplification in

Figure 3: Relative change of each lexical feature with respect to the unperturbed state (0%). Different markers represent lexical feature categories.

CELLS, 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, sentence length decreases; common words and verbs increase; and nouns, adjectives, and term specificity decrease. Although prior work emphasizes the importance of conjunctions for comprehension (Kauchak et al., 2017), our study reveals a reduction rather than increase in conjunctions as texts become simpler. Overall, these trends demonstrate that lexical features are valuable indicators for text simplification. Results on PLABA are similar, with an inverse trend for paragraph length (App. Figure 12).

LLM prompt-based evaluations show promise in distinguishing between PLS criteria. Prompt-based scores demonstrate sensitivity to perturbations in informativeness, faithfulness, and simplification, while showing less sensitivity to changes in coherence (Figure 4). While providing a single criterion, all criteria, and all criteria with an explanation mostly yield similar results, trends for simplification and some types of faithfulness perturbations are more clear and consistent when all criteria are provided. This suggests that providing all criteria and requesting all scores simultaneously is most efficient and accurate.

Our results also indicate that additional explanations are not essential for PLS evaluation (results for settings b and c are similar). However, further studies are required to better understand the decision-making process of the LLM, assess the validity of its explanations, and explore how the quality of explanations impacts the score. Prompts and detailed results are provided in App. G.

Figure 4: Prompt-based evaluation scores for four criteria - informativeness, simpli cation, coherence, and faithfulness - along with an overall score. (a) providing a single criterion in the prompt and requesting a score for that criterion; (b) providing all criteria in the prompt and requesting scores for each criterion as well as an overall score; and (c) the same setting as (b) but with an additional requirement for explanations of the provided scores. Notably, prompt-based scores demonstrate sensitivity to perturbations in informativeness, faithfulness, and simpli cation, while showing less sensitivity to changes in coherence. The three prompt settings yield similar results, with the exception that providing all criteria (setting b and c) is more sensitive to entity swaps compared to providing a single criterion (setting a).

Figure 5: Variation in existing scores for simpli cation perturbations created by GPT-3, Llama2, and Claude on the CELLS dataset.

7 Discussion & Conclusion

Recent advances point to the possibility of automated PLS; however, the multifaceted nature of PLS makes evaluation challenging. We introduce the rst—to our knowledge—meta-evaluation testbed, APPLS, for evaluating PLS metrics. In APPLS, we apply controlled text perturba-

tions to existing PLS datasets based on several criteria (informativeness, simplification, coherence, and faithfulness). Using APPLS, we find that while some metrics reasonably capture informativeness, faithfulness, and coherence, SARI is uniquely sensitive to simplification perturbations, but exhibits insensitivity to other perturbations. Similar challenges are observed for QAEval, as no single metric was consistently sensitive to all perturbations across criteria. Therefore, an evaluation metric suite should be considered based on all desired criteria. From our results on APPLS, we identify the following metrics for each criterion as most promising from among those we tested: SARI for simplicity, GPT-PPL for informativeness, LENS for coherence, and QAEval for faithfulness. However, we warn that all of these automated metrics have limitations as identified in our results. Further research is necessary to identify more robust metrics for a comprehensive evaluation of PLS.

The quickly improving performance of language models on a variety of tasks has placed greater emphasis on extrinsic human evaluations (Clark et al., 2021) evaluating models in real-world use cases, often with end-users. However, extrinsic evaluations are time-intensive, difficult to implement correctly, and costly, making them only viable for the most promising models. Automated metrics offer a fast and low-cost method for identifying improvement trends even if they do not perfectly measure absolute improvement, or improvement at the instance level. Our selection of automated metrics grounded in criteria from the health communication literature offer a viable first step in evaluating systems. Initial automated evaluations can then be followed by extrinsic evaluations to ensure comprehensive analysis for real-world use.

Our APPLS testbed allows for extensible evaluation of PLS evaluation metrics. Although this study focuses on the health domain, APPLS can be adapted to other domains by changing the diagnostic dataset. Depending on the domain, evaluation criteria may need to be adjusted. For example, in legal contexts, faithfulness might be prioritized over informativeness, and additional criteria such as language specificity (e.g., avoidance of vague terminology) may be necessary. Using our perturbation pipeline, APPLS can transform any PLS dataset into a granular meta-evaluation testbed. New perturbations can be introduced, and new evaluation metrics can also be incorporated easily into analysis. Our testbed lays the groundwork for further advancements in automated PLS and PLS evaluation, aiming to foster more impactful, accessible, and equitable scientific communication.

Limitations

Our perturbations use synthetic data to simulate real-world textual phenomena seen in PLS. Although our approach is informed by prior work and provides valuable insights into metric behavior, further exploration of more sophisticated methods to simulate changes in these criteria is warranted. This is especially true for aligning sentences between scientific abstracts and plain language summaries, as sentence-level alignment for scientific summaries is still an open problem (Krishna et al., 2023).

We also acknowledge that text quality may deteriorate with synthetic perturbations in a way that affects multiple PLS criteria. However, by using synthetic data, we are benefiting from the ability to control our perturbations and extend our testbed creation framework to any dataset. It is infeasible to find naturally occurring text with the same controlled levels of each perturbation, with minimal changes to other aspects. Our aim is not to produce perfect outputs, but rather to establish a robust baseline for evaluating the performance of automated metrics for PLS evaluation. The results of our analysis complement qualitative examinations of model output conducted in other work, which further suggests that automated text generation evaluation metrics may be limited in their ability to assess generation performance of post-GPT-3 LLMs (Goyal et al., 2022).

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 the US National Library of Medicine [grant number R21LM013934] and by Azure cloud credits provided by the UW eScience Institute.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo

Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simpli cation 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. Scienti c Data, 10(1):8.

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 Transactions on Computer-Human Interaction, 30(5):1–38.

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.

Isabel L Beck, Margaret G McKeown, Gale M Sinatra, and Jane A Loxterman. 1991. Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. Reading research quarterly, pages 251–276.

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.

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

Yanran Chen and Steffen Eger. 2023. Menli: Robust evaluation metrics from natural language inference. Transactions of the Association for Computational Linguistics, 11:804–825.

Elizabeth Clark, Tal August, So a Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's `human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

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.

Ashwin Devaraj, Iain Marshall, Byron C Wallace, and Junyi Jessy Li. 2021. Paragraph-level simpli cation 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.

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. 2021. Go gure: A meta evaluation of factuality in summarization. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 478–487.

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 scienti c 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, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. ArXiv, abs/2209.12356.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. Journal of Biomedical Informatics, 149:104580.

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

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. Highres: Highlight-based reference-less evaluation of summarization. In Proceedings of the 57th

Annual Meeting of the Association for Computational Linguistics, pages 3381–3392.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12067–12097.

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.

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.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 6610–6617.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. Ffci: A framework for interpretable automatic evaluation of summarization. Journal of Artificial Intelligence Research, 73:1553–1607.

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.

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.

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. 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. 2023. Lens: A learnable evaluation metric for text simplification. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16383–16408.

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, Aitor Soroa, Gorka Labaka, and Eneko Agirre. 2022. Principled paraphrase generation with parallel corpora. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1621–1638.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4812–4829.

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.

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. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7219–7234.

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.

Reid Smith, Pamela Snow, Tanya Serry, and Lorraine Hammond. 2021. The role of background knowledge in reading comprehension: A critical review. Reading Psychology, 42(3):214–240.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5123–5137.

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. In 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pages 738–744. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

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.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey Kuehl, Erin Bransom, and Byron Wallace. 2023. Automated metrics for medical multi-document summarization disagree with human evaluations. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9871–9889, Toronto, Canada. Association for Computational Linguistics.

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. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2448–2460.

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 Zhang. 2022. Improving meta-learning for low-resource text classification and generation via memory imitation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 583–595.

Figure 6: Process for generating the candidate summary, on which we apply all perturbation operations.

A Applying Perturbations

Figure 6 shows how the candidate summary is extracted from the original scienti c abstract. An extractive summary is identi ed based on high ROUGE-L with the plain language target. The extractive summary is passed through round-trip translation via German to introduce lexical variation. The resulting candidate summary forms the basis for our perturbations.

For coherence, we count the total sentence displacement from their original positions, so for example, swapping the rst and last sentences would result in a higher perturbation percentage compared to swapping the rst and second sentences. We have published the perturbation code so that others can review it and deploy the testbed on different datasets.

B Round-trip translation for oracle extractive summary

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 nd that the English-German-English (en-de-en) translation sequence yields superior BLEU scores (Figure 7), and therefore, select the en-de-en sequence to produce the candidate summary for our testbed.

Figure 7: BLEU scores of round-trip translation for English-German-English (en-de-en) and English-Russian-English (en-ru-en) in CELLS oracle extractive summaries.

Figure 8: Comparison of BLEU scores between oracle extractive summary (extracted) and candidate summary (following roundtrip translation), using the scienti c abstract (src) as the reference for BLEU calculation.

To scrutinize the introduced variation through this extractive and round-trip translation pipeline, we evaluate the BLEU score. As depicted in Figure 8, the BLEU score for the candidate summary is lower than that of the oracle extractive summary. This suggests the successful introduction of text variations. Augmented by human evaluation results in Table 3, with 152 out of 198 raters indicating comparable simpli cation levels between the candidate summary and its extractive counterparts, we conclude that our extractive and round-trip translation approach successfully introduces lexical variation in our oracle extractive summaries without altering their simplicity level.

C Details of human evaluation

To validate the quality of candidate summaries and GPT-simpli ed summaries, we randomly select

Type	Unmatched	Criteria	Str. Agree	Agree	Neutral	Disagree	Str. Disagree
Round Trip Translation	1	Simpli cation	12	27	152	7	0
		Informativeness	188	4	3	4	0
		Faithfulness	155	6	4	20	14
		Coherence	30	11	156	2	0
GPT Simpli cation	0	Simpli cation	67	32	1	0	0
		Informativeness	37	37	21	5	0
		Faithfulness	38	43	14	4	1
		Coherence	10	47	41	2	0

Table 3: Counts of human evaluation ratings on each matched sentence for each criteria. For round trip translation, there are 200 ratings; for GPT simpli cation, there are 100 ratings. 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 simpli cation and coherence criteria). For GPT simpli cation, we see that the simpli cation perturbation leads to substantially more simple text, while also maintaining faithfulness and informativeness.

Figure 9: An example human evaluation task for assessing GPT-simpli ed summary quality.

100 summary pairs from each corpus for human evaluation. Each pair in the candidate summary annotation task consists of an oracle extractive sentence and its respective en-de-en round-trip translation sentence. Similarly, each pair in the GPT-simpli ed summary annotation task contains a chunk of the candidate summary along with its corresponding GPT-simpli ed 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 rst asked to assess whether the content of Text A matches the content of Text B, where a match is de ned 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 (simpli cation), maintains factual integrity (faithfulness), and exhibits a well-structured layout (coherence). All facets are assessed using a 1-5 Likert scale (1-strongly disagree,

Figure 10: Most common evaluation metrics reported in ACL'22 summarization and generation long papers.

Figure 11: Distributions of reported metric improvements over baseline (absolute value) reported in ACL'22 summarization and generation long papers.

5-strongly agree). Representative questions can be found in Figure 9. This research activity is exempt from institutional IRB review.

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 simpli cation tasks. We primarily concentrate on tasks related to generation, summarization, and simpli cation. 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 quali ed papers related to simpli cation tasks. Considering the signi cance of simpli cation in PLS, we expanded our search to all ACL 2022 papers, including long, short, system demonstration, and ndings papers. This led to the identi cation of 2 out of 22 papers with `simpl' in the title that reported SARI scores. As illustrated in Figure 10, the ve most frequently reported automated evaluation metrics are ROUGE, BLEU, GPT-PPL, METEOR, and BERTScore.

This investigation provides insight into the current adoption of evaluation metrics in natural language generation, summarization, and simpli cation 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 dif culty 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. We report the median of BERTScore values and deltas as reported in these publications, without considering the usage of different models or settings.

E Details on existing automated evaluation metrics

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

- ROUGE[10] (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.
- BLEU[10] (Papineni et al., 2002) computes n-gram precision of generated text against reference texts, including a brevity penalty.
- METEOR[10] (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[11] (Xu et al., 2016) is speci cally designed to evaluate text simpli cation tasks. The score weights deleted, added, and kept n-grams between the source, generated, and target texts.

[10]Implementation: Fabbri et al. (2021) BERTScore hash code: bert-base-uncased_L8_no-idf_version = 0.3.12(hug_trans=4.27.3).

[11]Implementation: Alva-Manchego et al. (2019)

Figure 12: Relative change of each lexical feature with respect to perturbations in the PLABA dataset. Different markers represent lexical feature categories.

Model-based metricsuse pretrained models to evaluate text quality.

- GPT-PPL,[12] usually computed with GPT-2, measures uency and coherence by calculating the average log probability assigned to each token by the GPT model, with lower scores indicating higher uency and coherence.
- BERTScore[10] (Zhang et al., 2019) quanti es the similarity between candidate summaries and targets using contextualized embeddings from the BERT model, computing the F1-score between embeddings to capture semantic similarity beyondn-gram matching.
- LENS (Maddela et al., 2023) employs an adaptive ranking loss to focus on targets closer to the system output in edit operations (e.g., splitting, paraphrasing, deletion).

QA-based metricscapture 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.

F Additional experiments for existing metrics

To illustrate that existing metrics are not sensitive to text simplicity but rather to length andn-gram overlap, we present metric scores computed when swapping source and target for simpli cation perturbations (Figure 13). When target text is used as reference, we start with the candidate summary and increase perturbation percentage by swapping in simpler text, going from more complex to more simple text. When source text is used as reference, we reverse the original source and target, starting with simple text and swapping in sentences from the candidate summary, thereby moving from more simple to more complex text. A metric sensitive to text simpli cation should move in opposite directions in these two settings as perturbation percentage increases. However, these metric scores uniformly decrease under both settings, regardless of the reference, demonstrating that these metrics are not responsive to simpli cation but more so to text length andn-gram overlap. We do not report performance of BERTScore and QAEval under this setting due to the higher cost of computing these model based metrics.

G LLM Prompt-Based Evaluation

We use GPT-4 for LLM evaluation. The generation process is con gured with a temperature parameter of 0, a maximum length of 150, and a penalty value of 0. For each input, the top-ranked text is selected as the GPT-simpli ed output. Example prompts used for evaluation are provided in Figure 14.

H Additional perturbation results for PLABA

We present full perturbation results on PLABA (Attal et al., 2023) in Figure 16. Trends for many perturbations are in the same direction as in CELLS. While many metrics now show a desirable reversed trend to simpli cation (increasing), we point out that this is inconsistent performance relative to CELLS and is due to the highn-gram overlap between the candidate summaries and targets in this case (we perturb by replacing source sentences with round-trip translated target sentences to form the candidate summary, which only introduces minor lexical variation). Adding text, especially de nitions, dramatically decreases many of these metrics due to the similar lengths of source and target texts in PLABA, again pointing to then-gram and length sensitivities of most of these metrics.

The impact of simpli cation perturbations on lexical features in the PLABA dataset is shown in Figure 12. Most trends are similar to CELLS, though paragraph length increases with higher perturbation percentage. In PLABA's target construction scheme, the target simpli ed texts length (244) are similar to the source abstracts (240).

[12] https://huggingface.co/transformers/v3.2.0/perplexity.html

Figure 13: Average scores of ROUGE, BLEU, METEOR, and SARI scores calculated using either the source text (complex) or target text (simple) as reference for simplification perturbations on the CELLS dataset. A metric sensitive to text simplicity should move in opposing directions under these two settings. However, ROUGE, BLEU, and METEOR decrease uniformly in both settings, suggesting that they are not sensitive to text simplicity.

Criteria:
-**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.

a. Single criterion provided
Imagine you are a human annotator now. You will evaluate the quality of a generated plain language summary for a scientific literature abstract. Please follow these steps:

1. Read the scientific abstract provided.
2. Read the generated plain language summary.
3. Compared to the scientific abstract, rate the generated summary on the following criteria: {one of the criteria}
4. Assign a score for the generated summary, rating on a scale from 0 (worst) to 100 (best).
5. You do not need to explain the reason. Only provide the score.

Scientific abstract:{abstract};
Generated plain language summary:{pls_gen};
Score:

b. All criteria provided
Imagine you are a human annotator now. You will evaluate the quality of a generated plain language summary for a scientific literature abstract. Please follow these steps:

1. Read the scientific abstract provided.
2. Read the generated plain language summary.
3. Compared to the scientific abstract, rate the generated summary on the following criteria: {all criteria}
4. Assign a score for the generated summary, rating on a scale from 0 (worst) to 100 (best).
5. You do not need to explain the reason. Only provide the score.

Scientific abstract:{abstract};
Generated plain language summary:{pls_gen};
Score:

c. All criteria provided, explanation needed
Imagine you are a human annotator now. You will evaluate the quality of a generated plain language summary for a scientific literature abstract. Please follow these steps:

1. Read the scientific abstract provided.
2. Read the generated plain language summary.
3. Compared to the scientific abstract, rate the generated summary on the following criteria: {all criteria}
4. Assign a score for the generated summary, rating on a scale from 0 (worst) to 100 (best).
5. *Explain the reason for the score.*

Scientific abstract:{abstract};
Generated plain language summary:{pls_gen};
Score:

Figure 14: Prompts used for GPT-4 based evaluation: (a): single criterion provided; (b) all criteria provided; and (c) all criteria provided and explanation needed.