A Meta-evaluation of Automatic Metrics for Elaborative Simplification

Abdullah Alshatti, Steven Schockaert, Fernando Alva-Manchego
School of Computer Science and Informatics, Cardiff University, UK
{AlshattiAM, SchockaertS1, AlvaManchegoF}@cardiff.ac.uk

Abstract

Elaborative simplification aims to improve the readability of texts by adding content that helps the readers. However, evaluating these elaborations remains challenging due to their subjective nature and the lack of suitable annotated datasets. To support the evaluation of elaborative simplification models, we introduce a new dataset with human ratings of elaborations generated by Large Language Models (LLMs), focusing on two quality criteria: cohesion and informativeness. Using these human judgments as a reference, we conduct a meta-evaluation of existing automatic evaluation approaches, with a focus on LLM-as-a-judge strategies. Our experiments suggest that evaluations made by smaller LLMs correlate poorly with human judgments, while larger models with structured prompting exhibit higher agreement. Informativeness evaluation proved to be challenging due to its subjectivity, as evidenced by the low inter-annotator agreement compared to cohesion.

Keywords: Text Simplification, Elaboration, Text Evaluation

1. Introduction

Elaborative simplification improves readability by adding explanations or information aligned with the original text's meaning and context (Srikanth and Li, 2021). Table 1 illustrates such an elaboration. The quality of elaborations is crucial, as inaccuracies or poorly worded extra information may render the text nonsensical and harder to understand (Long and Ross, 1993; Shardlow, 2014).

However, a critical gap exists in evaluating elaborative simplification, as traditional reference-based metrics have proven inadequate due to their reliance on lexical overlap with fixed references (Srikanth and Li, 2021). Even embedding-based metrics such as BERTscore (Zhang et al., 2020) are often poorly correlated with human judgments of quality (Moramarco et al., 2022; Li et al., 2024; Kryscinski et al., 2020; Fabbri et al., 2021). Overall, while traditional automated metrics offer scalability and objectivity, often fall short in generative tasks.

A potential solution is to rely on LLM-as-a-judge approaches, where the quality of elaborations is assessed by prompting an LLM, eschewing the need for reference answers. However, there are currently no annotated datasets that can be used for evaluating the reliability of LLM judges in elaborative simplification, nor is there a standardized framework that can be used for designing suitable rubrics to assist them.

In this paper, we address this gap by introducing ElabEval, a manually annotated dataset of elaboration quality.[1] Specifically, we collect human quality ratings for LLM-generated elaborations for anchor sentences from the ElabQUD dataset (Wu et al., 2023). Our annotations cover two aspects of quality: (i) *cohesion*, measuring the extent to which an elaboration is sensible within the given context, and (ii) *informativeness*, measuring how useful the provided information is likely to be to the reader.

CONTEXT:	Anderson became interested in people like Landa when she noticed something strange about a call center near her house.
ELABORATION:	Workers at call centers help people over the phone.

Table 1: Example of an elaboration from ElabQUD (Wu et al., 2023), answering the implicit question under discussion: *What do call centers do?*

Using this dataset, we conduct a meta-evaluation of standard reference-based metrics and reference-free LLM-as-a-judge approaches. We also leverage the LLM-as-a-Judge approach to provide a scalable alternative to human assessment and compare its reliability against both traditional automatic metrics and human judgments, employing a variety of LLMs to evaluate elaborations, ensuring a more comprehensive assessment across model sizes and architectures. Our analysis confirms that standard metrics, which rely on comparing the generated elaborations with reference elaborations from the ElabQUD dataset, perform poorly. When it comes to (reference-free) LLM judges, we found smaller models to perform surprisingly poorly, barely outperforming random guessing. However, we also found that frontier models such as gpt5 can provide reliable assessments when sufficiently detailed instructions are provided.

Overall, our results show that the proposed

[1] Resources available on: `https://github.com/Abdullah-alshatti/ElabEval`

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 193–209
11 May 2026.

dataset provides a realistic reference for meta-evaluating automatic evaluation methods for elaborative simplification, while also revealing the current limitations of LLM-as-a-judge approaches.

2. Related Work

Datasets. Srikanth and Li (2021) introduced the term "elaborative simplification" to describe content addition in text simplification to improve readability. Through crowdsourcing, they collected a dataset of 1.3K naturally occurring elaborations in the Newsela corpus (Xu et al., 2015) focusing on the contextual aspect for these elaborations. Based on this annotated dataset, Wu et al. (2023) used human annotation and encoder-decoder models to generate an implicit Question Under Discussion (QUD) to help guide LLMs in producing contextually relevant elaborations. Laban et al. (2023) created the SWIPE dataset, which reconstructs the document-level editing process from English Wikipedia (EW) articles to paired Simple Wikipedia (SEW) articles by leveraging the entire revision history during the pairing process in order to better identify simplification edits. In addition to other types of simplifications, this dataset contained instances of elaboration as well.

Since the quality of Wikipedia-based datasets in simplification is questioned (Trokhymovych et al., 2024), we opted for ElabQUD, which is the Newsela corpus, a professionally curated resource of English-language texts. This choice ensured higher data quality while also introducing QUD-based elaboration structures that enriched the diversity of our dataset.

Methods. Existing approaches to generate elaborative simplifications mainly use Transformer (Vaswani et al., 2017) based models , relying on prompting and fine-tuning of pre-trained language models. For example, Srikanth and Li (2021) fine-tuned GPT-2 (Radford et al., 2019), using the simplest texts in Newsela and their annotated elaborations, then provided the model with the text preceding the elaboration in a simplified text as input, and the model would generate the elaboration as output. Wu et al. (2023) used GPT-3 (Brown et al., 2020) for zero-shot elaboration generation, experimenting with including an automatically generated and manually written QUD in the prompt, finding that the latter type produced the best elaborations. For German, Hewett et al. (2024) used Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024) as both out-of-the-box model and fine-tuned on B1 and A2 German texts, following the CEFR language proficiency framework. They also used prompt variations to generate elaborations with generic, background, and contextual information.

These approaches are also used in definition generation, a related task. Yarbro and Olney (2021) used a dataset containing words definitions and a list of contexts associated them to fine-tune a GPT-2 based model to generate definitions for English words with only the word and a context as inputs. Asthana et al. (2024) used four LLMs: GPT-4 (OpenAI, 2024), PaLM-2 (Anil and et al, 2023), Falcon-40b (Almazrouei et al., 2023), and BLOOM-176b (Workshop et al., 2023). They provided the models with the term, definition, and difficult concept and used two types of prompts that reflect two simplification strategies, to rewrite the definition by adding an explanation for the difficult concept and to rewrite the definition and simplify the difficult concept word.

We adopted an approach similar to these works to generate elaboration instances for our dataset. We provide the LLM models with the contexts that need to be elaborated on, using two types of prompting approaches: QUD-based prompting and a descriptive prompt without specifying the elaboration type that needs to be generated.

Evaluation. Previous work on elaborative simplification (Wu et al., 2023; Laban et al., 2023) reported automatic metrics scores like BERTScore (Zhang et al., 2020), BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016) for completeness but relied on humans to evaluate the elaborations through comparing various elaborations based on how coherent and elaboration-like they are. These works did not provide clear definitions of what that means, making the evaluations vague, unspecified, and reliant on the interpretation of the evaluators.

Existing text simplification metrics face significant limitations when evaluating elaborative simplification compared to standard simplification; current metrics either aggregate meaning preservation and simplification into single overall scores or target only meaning preservation, confirming that no single automatic metric captures all necessary evaluation criteria (Cripwell et al., 2024; Alfear et al., 2024; Guo et al., 2024; Alva-Manchego et al., 2021). Such approaches penalize novel content generation, resulting in poor correlation with human judgments (Li et al., 2024; Barayan et al., 2025).

Studies in Natural Language Generation tasks have found the LLM-as-a-judge approach can achieve high correlation with human judgments (Wang et al., 2025), outperforming traditional automatic metrics in this respect (Nguyen et al., 2024; Chen et al., 2023). We also found the LLM-as-a-judge approach to be suitable for evaluating elaborations, as they have an extensive output space and cannot be fully captured by a fixed number of references. However, for this to work, appropriate criteria for evaluating the quality of the elaborations need to be established first. We introduce an explicit evaluation framework for elaborative simplification that evaluates the elaborations via a two-stage rubric: annotators first make a binary cohesion judgment (filtering out irrelevant, inconsistent, or merely repetitive elaborations) and then rate only cohesive outputs for informativeness on a three-level scale.

3. The ElabEval Dataset

This section describes the curation of ElabEval, our annotated English dataset for the *meta-evaluation* of elaboration quality, i.e. for assessing the reliability of elaboration evaluation metrics.

3.1. Curation of Elaborations

Source Texts. We selected 100 news articles from ElabQUD (Wu et al., 2023) to use as contexts for the elaborations in our dataset. In ElabQUD, for each context, both an implicit Question Under Discussion (QUD) and an elaboration are provided. As context for each elaboration, we considered up to 5 sentences prior to the elaboration, following the same setup as Srikanth and Li (2021). The source text underwent a preprocessing step that involved removing the special characters to provide the context as a single clean input to the LLMs.[2]

LLM-Generated Elaborations. To obtain a dataset covering a wide range of elaboration quality, we first observed that smaller LLMs produce highly variable outputs. Based on this, we selected two such models to generate elaborations for our dataset: DeepSeek-R1-Distill-Qwen-7B[3] (Guo et al., 2025) and FLAN-T5_Instruct-Mistral7B[4] (Jiang et al., 2023). These models were chosen because they are open-weight, supporting reproducibility, and they consistently produced a mix of high- and low-quality elaborations. Both models were accessed via the Hugging Face Transformers library. They were used with all parameters set to default values, and the maximum generated elaboration length was capped at 100 tokens to ensure comparability across the models. For each context, we generated four elaborations (one for each prompt type and model). For each source context, the LLMs were instructed to generate an elaboration using the following two prompts:

- **Descriptive Prompt:** We provide specific instructions for the elaboration generation task with a target length of one sentence. The prompt was: *"You are tasked with generating a brief elaboration of one sentence from a given context. The context will be in the form of a paragraph consisting of multiple sentences. You're required to generate an appropriate new sentence that adds to the reader's understanding of the context in a clear and coherent way. Ensure the new sentence is concise and directly relevant to the information presented. Context: {sentence} Answer: "*
- **QUD:** We mirrored the methodology of (Wu et al., 2023), creators of ElabQUD. Specifically, the model was provided with inputs structured in the following format: *"Context: < context >, Question: < question >, Answer:"*

[2]The QUD framework views each sentence as the answer to an implicit or explicit question from prior context.

[3]https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

[4]https://huggingface.co/SanketAI/FLAN-T5_instruct-mistral7b

3.2. Collecting Human Judgments

Concerns have been raised regarding the declining quality of outputs generated by widely used crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk) (Chmielewski and Kucker, 2020). In addition, several specialized services, including FigureEight, have become unavailable for academic research following their acquisition by commercial entities (Gilardi et al., 2023). These concerns led us toward the recruitment of annotators with verified competencies. Although this strategy enhanced the reliability of the resulting data, it simultaneously imposed constraints on the attainable scale of the dataset.

Annotators. We recruited three native English-speaking evaluators: two postgraduate students with backgrounds in linguistics and a non-academic staff member. They were selected from a pool of 11 candidates based on a qualification task that assessed their understanding of the annotation guidelines (available in Appendix A).

Annotation Criteria. We focused on two primary criteria to assess elaborations: cohesion and informativeness. *Cohesion* is meant to be more objective, focusing on whether a given elaboration "makes sense", and is thus evaluated as a binary property. More precisely, for an elaboration to be cohesive, it should:

1. maintain relevance to the given context;
2. be free from logical inconsistencies;
3. not contain any misleading examples; and
4. not merely repeat parts of the original text.

Informativeness assesses the utility of the provided information, relative to the given context. Given

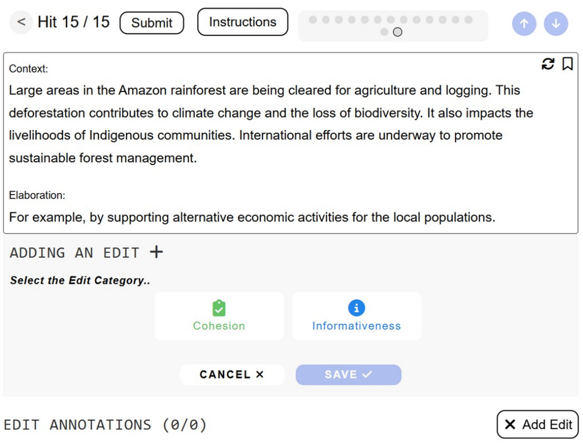

Figure 1: Screenshot of the annotation tool interface

the subjective nature of this criterion, it was evaluated using a three-point Likert scale. In particular, an elaboration is said to be *uninformative* if it offers no useful information for understanding the context, *somewhat informative* if it provides basic information that might benefit some readers, and *informative* if it would benefit most readers (e.g. by providing in-depth perspectives).

Annotation Tool. To collect the human ratings, we used the "thresh" annotation tool (Heineman et al., 2023), which is a customizable open-source platform for textual annotation.[5] Figure 1 shows a screenshot of our annotation interface created using this tool.

Annotation Process. Each annotator evaluated 500 elaborations: 100 gold-standard elaborations from ElabQUD and 400 generated by LLMs. For each instance, annotators first provided a cohesion judgement. Since incohesive elaborations are not informative, only those judged to be cohesive by an annotator were subsequently assessed for informativeness by the same annotator. As a result, informativeness ratings are only available for 226 out of the overall 500 instances: 165 have three evaluations for informativeness, 61 have two evaluations, and 5 instances have just one. In our analysis of informativeness, we only focus on those instances with three informativeness annotations. We convert the three ratings to a single value by summing them. By interpreting the individual ratings numerically, on a scale from 1 to 3, we thus obtain an overall informativeness score between 3 and 9. For cohesion, the final label was obtained via majority voting based on the scores given by the annotators.

[5] https://github.com/davidheineman/thresh

System		Level 1	Level 2	Level 3
Golden Elaborations		18	52	25
T5-Mistral	(DP)	8	2	0
	(QUD)	12	13	6
DeepSeek-R1-Qwen	(DP)	9	15	13
	(QUD)	18	28	7

Table 2: Distribution of informativeness labels (Level 1: Uninformative, Level 2: Somewhat Informative, Level 3: Informative) for the golden elaborations, the systems and prompting methods used: (DP) Descriptive Prompt and (QUD) Questions Under Discussion.

System		Cohesive	Incohesive
Golden Elaborations		95	5
T5-Mistral	(DP)	10	90
	(QUD)	31	69
DeepSeek-R1-Qwen	(DP)	37	63
	(QUD)	53	47

Table 3: Distribution of cohesion labels for the golden elaborations, and for the elaborations generated by two LLMs and two prompting strategies: (DP) Descriptive Prompt and (QUD) Questions Under Discussion.

3.3. Annotation Analysis

We measured inter-annotator agreement using Krippendorff's α (Krippendorff, 1980) and Fleiss' κ (Fleiss, 1971). Cohesion achieved moderate agreement ($\alpha = 0.61, \kappa = 0.61$), with unanimous judgments in 356 out of 500 instances (71.2%). In contrast, agreement for informativeness was fair ($\alpha = 0.24, \kappa = 0.08$), reflecting the more subjective nature of this criterion. Table 2 shows the distribution of informativeness labels in our dataset. Among the 226 instances judged as cohesive using annotator majority vote, agreement on informativeness was observed in only 38 cases (16.8%), with 26 labeled as *somewhat informative*, 5 as *uninformative*, and 7 as *informative*. This suggests that human annotators rarely agree on the informativeness of an elaboration, likely because this aspect is subjective and highly dependent on individual prior knowledge. Consequently, this resulted in significant data sparsity, which constrained our ability to conduct a fine-grained evaluation of informativeness. Nonetheless, these agreement levels provide a realistic reference point for our meta-evaluation.

The distribution of cohesion values across the final annotated dataset is presented in Table 3. The table indicates that elaborations generated by T5-Mistral were of lower quality relative to those produced by DeepSeek-R1-Qwen. Furthermore, the results demonstrate that elaborations generated using the QUD approach yielded higher quality instances compared to the Descriptive Prompt approach. Surprisingly, five of the 'gold' instances from the ElabQUD corpus were in fact identified as incohesive by our annotators. Upon manual review, four of these five incohesive instances were confirmed to be indeed incohesive based on the context, suggesting that even reference corpora like ElabQUD contain instances that do not meet our cohesion criteria.[6] We further analyzed the 356 elaborations with unanimous cohesion judgments (166 cohesive and 190 incohesive). Manual inspection of the incohesive cases revealed several recurring issues: 103 elaborations contained text repetitions, 59 were logically inconsistent with the context, 27 included hallucinations, and 1 contained a misleading example.

4. Meta-evaluation of Automatic Metrics

In this section, we analyze how well existing strategies for evaluating elaboration quality correlate with human judgments, focusing on traditional reference-based metrics (Sec. 4.1) and (reference-free) LLM-based judgments (Sec. 4.2). For cohesion, we used a balanced subset of 462 instances with the same number of cohesive and incohesive instances. For informativeness, we used a subset of 146 instances that all annotators rated as cohesive, thus having three evaluations, where we also excluded instances that had extreme rating disagreements (i.e. being labeled as *uninformative* by one annotator and *informative* by another).

4.1. Reference-based Metrics

To the best of our knowledge, there are no automated metrics specifically tailored for evaluating elaborative simplifications. For our study, we selected three standard metrics similar to those utilized in earlier works (Srikanth and Li, 2021; Wu et al., 2023):

- BLEU (Papineni et al., 2002) measures the precision of n-grams in a candidate elaboration compared to a reference. We used the implementation available in the Evaluate library.[7]
- METEOR (Banerjee and Lavie, 2005) goes beyond exact word matches by incorporating stemming and synonymy, for a better measure of semantic equivalence. We used the implementation available in the Evaluate library.

	Cohesion AUC	**Informativeness** Spearman ρ
BERTscore F1	0.55 ± 0.12	0.26 ± 0.30
BLEU	0.55 ± 0.11	0.27 ± 0.32
Meteor	0.52 ± 0.12	0.19 ± 0.32

Table 4: Meta-evaluation of reference-based metrics.

- BERTScore (Zhang et al., 2020) leverages contextual embeddings from pre-trained language models to capture semantic similarity that goes beyond mere lexical overlap. In our study, we calculated the BERTScore F1 metric using *distilbert-base-uncased* as the contextual embedding model, relying on the implementation available in the Transformers library, chosen for better computational efficiency.[8]

These automatic metrics allow us to rank the elaborations from best to worst, and we assess how well these rankings agree with the human ratings. We used the gold elaborations from ElabQUD as reference texts and compared the automatic-metric scores with the collected human ratings. For cohesion, which is a binary feature in our dataset, we use the Area under the ROC Curve (AUC). For informativeness, we assess rank correlation between metric scores and the ordering induced by aggregated human ratings using Spearman's ρ. The results in Table 4 confirm that reference-based metrics perform poorly for elaboration evaluation. For cohesion, the AUC scores are close to the expected performance of random guessing (0.5). For informativeness, we see a weak (but statistically significant) positive correlation.[9]

4.2. LLM Judges

We experimented with 6 open-weight models of various sizes (falcon-3-7b, mistral-7b, gemma-3-4b-it, llama-3.1-8b, qwen3-next-80b, deepseek-chat-v3.1), and three closed-weight models (gpt5, gpt4o, gpt4o-mini). The LLMs were prompted with three configurations:

- **Full Guidelines**: prompts mirroring the comprehensive instructions given to human annotators;
- **Concise Prompt + CoT**: a condensed version of the guidelines (omitting examples) with a Chain-of-Thought (CoT) approach;
- **Full Guidelines + CoT**: the complete human guidelines combined with CoT.

[6] See Appendix D for a discussion of these instances.

[7] https://huggingface.co/docs/evaluate/index

[8] https://huggingface.co/docs/transformers/en/index

[9] p-value are (0.0014) for BERTscore, (0.0011) for BLEU, and (0.0197) for Meteor.

		Cohesion Acc	Informativeness Spearman ρ
Full Guidelines	gpt4o	**0.84** ± 0.06	0.53 ± 0.25
	qwen3-next-80b	0.82 ± 0.06	0.40 ± 0.29
	gpt5	0.78 ± 0.07	0.49 ± 0.25
	deepseek-chat-v3.1	0.75 ± 0.07	0.47 ± 0.28
	gpt4o-mini	0.64 ± 0.08	**0.54** ± 0.23
	falcon-3-7b	0.58 ± 0.09	0.48 ± 0.24
	gemma-3-4b-it	0.53 ± 0.09	0.27 ± 0.29
	llama-3.1-8b	0.50 ± 0.08	0.41 ± 0.28
	mistral-7b	0.48 ± 0.09	0.16 ± 0.35
Concise Prompt + CoT	gpt4o	**0.84** ± 0.06	**0.55** ± 0.24
	gpt5	0.81 ± 0.07	0.43 ± 0.27
	qwen3-next-80b	0.75 ± 0.08	0.46 ± 0.28
	deepseek-chat-v3.1	0.71 ± 0.08	0.48 ± 0.23
	gpt4o-mini	0.64 ± 0.08	0.52 ± 0.23
	falcon-3-7b	0.61 ± 0.09	0.48 ± 0.25
	gemma-3-4b-it	0.57 ± 0.08	0.46 ± 0.24
	llama-3.1-8b	0.54 ± 0.08	0.45 ± 0.27
	mistral-7b	0.50 ± 0.09	0.37 ± 0.27
Full Guidelines + CoT	gpt5	**0.86** ± 0.06	0.44 ± 0.26
	gpt4o	0.84 ± 0.06	**0.59** ± 0.22
	qwen3-next-80b	0.75 ± 0.07	0.46 ± 0.29
	deepseek-chat-v3.1	0.70 ± 0.08	0.41 ± 0.27
	gpt4o-mini	0.66 ± 0.08	0.50 ± 0.22
	falcon-3-7b	0.59 ± 0.08	0.48 ± 0.24
	llama-3.1-8b	0.56 ± 0.08	0.45 ± 0.26
	gemma-3-4b-it	0.53 ± 0.08	0.43 ± 0.26
	mistral-7b	0.47 ± 0.09	0.37 ± 0.28

Table 5: LLM evaluation results for cohesion and informativeness, with 95% confidence intervals.

The LLM-based evaluation does not rely on reference answers, and we separately prompt the models to assess cohesion and informativeness. For cohesion, the models generate a binary judgment. For informativeness, we tested two approaches: asking the models to use a 7-point scale (mimicking the combined human ratings) and asking for an informativeness degree between 0 and 100. For this analysis, we used the latter, as it performed better based on the average values of 3 runs.[10]

Table 5 shows that all three prompting strategies perform similarly. The Full Guidelines + CoT configuration produced the overall best results for evaluating cohesion, while Concise Prompt + CoT yielded the overall best results for informativeness. Comparing the different models, we can see a strong correlation between model size and the results for cohesion, with gpt5, gpt4o and qwen3-next-80b performing particularly well. In contrast, the smaller models are not capable of assessing cohesion to a meaningful extent, with performances around random guessing. For informativeness, the results show a moderate agreement with the human ratings, and the impact of model size is less clear (e.g. gpt4o-mini performing similarly to gpt4o).

[10] Appendix B compares the two evaluation approaches.

Error Type	Count
Text Repetitions	103
Logical Inconsistency	59
Hallucinations	27
Misleading examples	1

Table 6: Overview of errors in clear-cut "Incohesive" instances

Model	INC	REP	HAL
gpt4o-mini	5	5	0
gpt4o	2	1	0
gpt5	4	4	1

Table 7: Prevalence of different error types among the clear-cut cases of cohesion prediction: Logical Inconsistency (INC), Text Repetitions (REP) and Hallucinations (HAL).

4.3. Analysis

We further analyse the predictions of the GPT models, which performed best on our experiments.

4.3.1. Cohesion

To more effectively examine the false positives in the GPT models' cohesion predictions, we used a dedicated subset of 356 instances in which all annotators unanimously agreed on the cohesion label. Table 6 summarizes the error categories observed among the incohesive instances within this subset. Table 7 shows the frequency of each fault type among the elaborations that the GPT models failed to correctly identify as incohesive per error. While the overall number of false positives is small in these clear-cut cases, it is surprising that even gpt5 fails to detect some repetitions and logical inconsistencies, despite being explicitly prompted to regard such cases as incohesive.

We observed instances where even the largest models (gpt4o and gpt5) misclassified texts as cohesive despite the presence of text repetitions; Table 8 presents an example of this phenomenon. Furthermore, 10 cohesive instances that achieved the highest level of informativeness by the human annotators were unanimously identified as incohesive by all GPT models; an example of such a case is presented in Table 9. This indicates that despite being provided with the same guidelines as the human annotators, the LLMs did not fully understand the nature of the cohesion criterion.

CONTEXT: The reason why is complicated. They said its genes were too much like the genes of other giraffes. All plants and animals have genes. They play a big part in what animals and plants look and act like. Genes are passed down from parents.
ELABORATION: Genes are passed down from parents to children.

Table 8: Example of an incohesive instance containing text repetition misclassified as cohesive.

CONTEXT: Then about 6 million years ago another big change occurred. Big cats split into several different species. They became lions, tigers, jaguars and leopards. But there's a problem: What scientists find by looking at big cat DNA doesn't agree with what the fossils tell them. Scientists are hoping to figure out where big cats first appeared.
ELABORATION: This period was known as the Mesozoic Era. This was a time of great diversity and evolution.

Table 9: Example of a cohesive instance misclassified as incohesive.

4.3.2. Informativeness

Figure 2 plots gpt4o informativeness predictions against the combined human ratings. The model defaults to predict that elaborations are uninformative, with a clear cluster of predictions around a z-score of -1, and a smaller cluster around 0.5. Such values are also predicted for elaborations that humans judged as highly informative. In contrast, elaborations rated as the least informative by humans never receive the highest scores by the model. This shows that the model is more prone to underscoring informative elaborations, than to overscoring uninformative ones.

One challenge in this evaluation was the incompatibility between model predictions (0-100 scale) and the aggregated human judgments (discrete 3-9 scale). To enable a meaningful statistical comparison (e.g. for the scatter plot in Figure 2), we applied z-score standardization to both predicted and ground truth values.

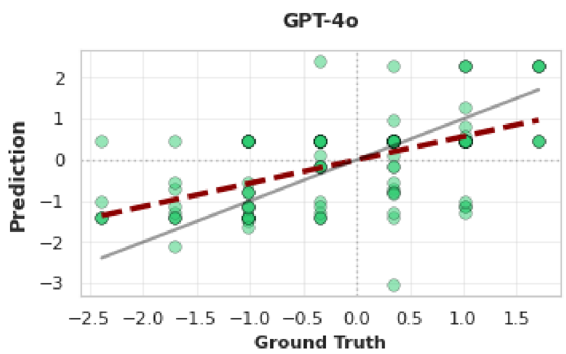

Figure 2: gpt4o Informativeness evaluation scatter plot.

CONTEXT: To reward his children, his father would bring books home. Usually, they were used books that cost just 5 cents. Billington is 85 years old. Even so, he does not plan to stop working anytime soon. He hopes to keep making it easier for people to use the library's huge collection.
ELABORATION: He feels a sense of accomplishment and satisfaction.

Table 10: Example of an instance underscored by all GPT models in informativeness.

Finally, Table 10 shows an example that was underscored by the GPT models. These instances tend to be short elaborations (around 10-15 words) that provide factual background information rather than direct explanations.

5. Conclusion

In this paper, we introduced a human-annotated dataset for the meta-evaluation of elaborative simplification evaluation methods. Our results show that standard reference-based metrics correlate weakly with human judgments, highlighting their clear limitations. LLM-as-a-judge approaches achieve higher correlation, provided that sufficiently large models are used. While their performance is impressive, especially in the light of the variability observed in human annotations, our qualitative analysis revealed that they remain an imperfect proxy for human judgment. Overall, the dataset provides a realistic reference for studying the strengths and limitations of automatic evaluation methods, especially as far as cohesion is concerned. Evaluating informativeness proved more challenging, as reflected by relatively low inter-annotator agreement. In future work, we will aim to address this by referring to a more clearly defined target audience in the definition of informativeness, and by splitting this criterion into more easily defined sub-criteria (e.g. How crucial is the information provided in this elaboration for understanding the text?, What proportion of the target audience would already be familiar with this information?).

Limitations

Due to budgetary constraints, we limited our scope by recruiting only three annotators and focusing our evaluation on elaborations generated from just two local LLMs. This restriction was required as the available funds were only sufficient to compensate the participants for evaluating 500 instances each. Furthermore, given that our study focuses on elaborations generated by locally hosted LLMs, model selection was constrained to those compatible with the Nvidia RTX 4090 graphics card installed on the workstation used for model execution.

Ethics Statement

This study received a favourable ethical opinion from the School Research Ethics Committee. All participants filled consent forms containing instructions on how their data would be used. Annotators were compensated for their work with vouchers worth £125 each to annotate 500 instances, at a rate equivalent to the legal minimum wage. AI assistants were occasionally used in producing this work. AI tools were utilized for researching and identifying relevant research, code completion and optimization, and refining the writing through spell-checking and paraphrasing.

Lay Summary

To make complex writing easier to understand, a technique called "elaborative simplification" can be used. Instead of just swapping big words for small ones, this method adds extra context or explanations to help the reader follow along. However, it is difficult for researchers to measure if these added explanations are actually "good" or helpful.

To help with this, we created a new dataset that contains human ratings for explanations generated by AI. We asked people to judge these explanations based on two factors: cohesion (how well the explanations fit the context) and informativeness (how useful the added information actually is). We then tested whether other AI models could automatically grade these explanations as accurately as a human would.

Our findings show that while smaller AI models struggle to match human judgments, larger AI models can be quite effective at judging quality if they are given very specific instructions. We also found that "informativeness" is much harder to agree on than "cohesion," simply because what one person finds helpful, another might find unnecessary. This work provides a better roadmap for building AI tools that can explain complex ideas clearly and reliably.

6. Bibliographical References

Noof Abdullah Alfear, Dimitar Kazakov, and Hend Al-Khalifa. 2024. Meta-evaluation of sentence simplification metrics. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11229–11235, Torino, Italia. ELRA and ICCL.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Rohan Anil and Andrew M. Dai et al. 2023. Palm 2 technical report.

Sumit Asthana, Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2024. Evaluating LLMs for targeted concept simplification for domain-specific texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6208–6226, Miami, Florida, USA. Association for Computational Linguistics.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

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, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey 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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Michael Chmielewski and Sarah C. Kucker. 2020. An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2024. Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 1–14, Torino, Italia. ELRA and ICCL.

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.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9194–9211, Miami, Florida, USA. Association for Computational Linguistics.

David Heineman, Yao Dou, and Wei Xu. 2023. Thresh: A unified, customizable and deployable platform for fine-grained text evaluation. pages 336–345.

Freya Hewett, Hadi Asghari, and Manfred Stede. 2024. Elaborative simplification for German-language texts. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 29–39, Kyoto, Japan. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Klaus Krippendorff. 1980. Reliability. In *Content Analysis: An Introduction to Its Methodology*, pages 277–360.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A dataset for document-level simplification of Wikipedia pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695, Toronto, Canada. Association for Computational Linguistics.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA. Association for Computational Linguistics.

Michael H. Long and Steven J. Ross. 1993. Modifications that preserve language and content. *EDRS*, pages 29–52.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Huyen Nguyen, Haihua Chen, Lavanya Pobbathi, and Junhua Ding. 2024. A comparative study of quality evaluation methods for text summarization.

OpenAI. 2024. GPT-4 technical report.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014*, 4(1).

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, Online. Association for Computational Linguistics.

Mykola Trokhymovych, Indira Sen, and Martin Gerlach. 2024. An open multilingual system for scoring readability of Wikipedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6296–6311, Bangkok, Thailand. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ruiqi Wang, Jiyu Guo, Cuiyun Gao, Guodong Fan, Chun Yong Chong, and Xin Xia. 2025. Can LLMs replace human evaluators? an empirical study of LLM-as-a-judge in Software Engineering. *Proc. ACM Softw. Eng.*, 2(ISSTA).

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

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.

Jeffrey T. Yarbro and Andrew M. Olney. 2021. Contextual definition generation. In *Proceedings of the Third International Workshop on Intelligent Textbooks 2021*, pages 74–83. CEUR-WS.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A. Annotation guidelines

Task Description

To make a text easier to understand, adding more information can be helpful for readers. This added content is called an elaboration. There are various kinds of elaborations, such as explanations, definitions, and other types of background information. For example: **Context** *"Photosynthesis is a foundational biological process that sustains nearly all ecosystems on the planet. Understanding it is vital for studying plant life as its efficiency impacts global atmospheric composition."* **Elaboration** *"This process, whereby plants convert light energy into chemical energy, also generates the oxygen we breathe."* In this task, you are asked to evaluate elaborations based on **two criteria**: *cohesion* and *informativeness*. They are defined as follows:

1. Cohesion:
A cohesive elaboration should:

- be relevant to the context;
- be free from errors and logical inconsistencies;
- not contain any misleading examples; and
- not simply repeat parts of the original text.

Cohesion is evaluated as **a binary property**, i.e. either an elaboration is *cohesive*, or it is *incohesive*. For example: **Context** "The government is split into two parties that often have different political beliefs."

Elaborations

- *"For example, the Labor Party and the Conservative Party."* **Cohesive.** The elaboration expands upon the original context with relevant and logically consistent information.
- *"This division often leads to lengthy debates and legislative gridlock, as each party attempts to push its own agenda."* **Cohesive.** The elaboration expands upon the original context with relevant and logically consistent information.
- *"The weather outside is very sunny."* **Incohesive.** The elaboration is irrelevant to the context provided.
- *"Therefore, the government is a single, unified entity with no internal disagreements."* **Incohesive.** The elaboration introduces logical inconsistencies for the context.
- *"The government is primarily focused on the production of chocolate and ice cream."* **Incohesive.** The elaboration introduces irrelevant information for the context.
- *"For example, the wedding party and the birthday party."* **Incohesive.** The elaboration introduces misleading examples.
- *"The government is split into two parties"* **Incohesive.** The elaboration repeats part of the original text.

2. Informativeness:
Informativeness evaluates the quality of the information provided in the elaboration based on the context. You will use a **scale from 1 to 3** based on how informative the elaboration is, as follows:

1. **Uninformative:** Nobody would benefit from the elaboration; it does not provide helpful information for understanding the context.
2. **Somewhat Informative:** Some people would benefit from the elaboration; it provides some basic information related to the context.
3. **Informative:** Most people would benefit from the elaboration; it goes beyond basic information by offering a more in-depth perspective or less obvious details.

For example: **Context** *"But not many countries support Obama's plan to fire missiles at Syria."*

Elaborations

1. Uninformative: *"Missiles can travel at speeds that exceed Mach 3."* The elaboration provides a general fact about missiles, but it does not help the reader understand the context of Obama's plan to fire missiles at Syria or the level of international support for it.
2. Somewhat Informative: *"Countries were concerned about the potential for civilian casualties."* The elaboration adds a relevant perspective by focusing on concerns from countries. It provides a basic reason for the opposition (civilian casualties), but lacks deeper insight into the political dynamics or the specific views of key stakeholders.
3. Informative: *"Allies of the Syrian government, might be drawn into the conflict, leading to a dangerous escalation."* The elaboration explains a key geopolitical risk, helping readers understand why countries opposed the plan. It adds depth by highlighting the potential for escalation, making it highly informative.

B. Additional Experimental Results

Table 11 shows the full classification metrics for cohesion evaluations, to complement the accuracy scores reported in the main paper. In particular, the table also shows the total number of instances predicted as cohesive and incohesive, as well as the precision, recall and F1 scores.

To evaluate LLM predictions of informativeness, we compared two approaches: asking the models to use a 7-point scale and asking for a degree between 0 and 100. Table 12 compares these two approaches.

Prompt & Models	Cohesive	Incohesive	Accuracy	Precision	Recall	F1
Full Guidelines						
gpt4o	231	231	0.844	0.844	0.844	0.844
qwen3-next-80b	192	270	0.825	0.891	0.740	0.809
gpt5	287	175	0.788	0.732	0.909	0.811
deepseek-chat-v3.1	120	342	0.751	0.983	0.511	0.672
gpt4o-mini	122	340	0.643	0.770	0.407	0.533
falcon-3-7b	351	111	0.580	0.552	0.841	0.667
gemma-3-4b-it	436	26	0.530	0.516	0.970	0.674
llama-3.1-8b	330	132	0.509	0.506	0.720	0.594
mistral-7b	111	351	0.485	0.469	0.228	0.307
Concise Prompt + CoT						
gpt4o	181	281	0.840	0.934	0.732	0.820
gpt5	250	212	0.816	0.792	0.857	0.823
qwen3-next-80b	130	332	0.751	0.946	0.532	0.681
deepseek-chat-v3.1	101	361	0.710	0.98	0.429	0.596
gpt4o-mini	107	355	0.640	0.804	0.372	0.509
falcon-3-7b	347	115	0.610	0.573	0.861	0.689
gemma-3-4b-it	382	80	0.578	0.547	0.905	0.682
llama-3.1-8b	342	120	0.543	0.529	0.784	0.632
mistral-7b	0	462	0.500	0.000	0.000	0.000
Full Guidelines + CoT						
gpt5	240	222	0.864	0.850	0.883	0.866
gpt4o	189	273	0.844	0.921	0.753	0.829
qwen3-next-80b	127	335	0.758	0.969	0.532	0.687
deepseek-chat-v3.1	96	366	0.704	0.990	0.411	0.581
gpt4o-mini	112	350	0.669	0.848	0.411	0.554
falcon-3-7b	359	103	0.593	0.560	0.87	0.681
llama-3.1-8b	338	124	0.565	0.544	0.797	0.647
gemma-3-4b-it	430	32	0.530	0.516	0.961	0.672
mistral-7b	255	207	0.476	0.478	0.528	0.502

Table 11: Full LLM Cohesion Evaluations

C. Experimental Details

The smallest LLMs were run locally on our test machines (falcon-3-7b[11], mistral-7b[12], gemma-3-4b-it[13], llama-3.1-8b[14]). For experiments with qwen3-next-80b and deepseek-chat-v3.1, we relied on the Openrouter[15] platform. Finally, gpt5, gpt4o and GPT-4o-mini were evaluated using the OpenAI API[16]. We used default hyperparameter values for the models.

[11] https://huggingface.co/tiiuae/Falcon3-7B-Instruct

[12] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[13] https://huggingface.co/google/gemma-3-4b-it

[14] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[15] https://openrouter.ai/

[16] https://platform.openai.com

D. Discussion of Incohesive Golden Elaborations

In this section, we analyze the five 'gold' elaborations from the ElabQUD dataset that were rated as incohesive by our annotators out of the 100 total gold elaboration. As illustrated in Table 13 , with the exception of instance 3 in the table, which provided accurate information regarding the role of a jury, the remaining four instances are indeed incohesive. The faults are characterized by contextual irrelevance or logical inconsistencies. We attribute these errors to the extraction methodology used to identify elaborations within the original Newsela articles, which exclusively considers sentences immediately following a specified context. This approach occasionally captures textual fragments or the introductory phrases of unrelated sentences, leading to incoherent elaborations.

	7-point Scale Spearman ρ	0 to 100 Scale Spearman ρ
Full Guidelines		
gpt4o	0.50 $\pm$ 0.27	0.53 $\pm$ 0.25
qwen3-next-80b	0.36 $\pm$ 0.30	0.40 $\pm$ 0.29
gpt5	0.46 $\pm$ 0.25	0.49 $\pm$ 0.25
deepseek-chat-v3.1	0.46 $\pm$ 0.25	0.47 $\pm$ 0.28
gpt4o-mini	**0.52** $\pm$ 0.22	**0.54** $\pm$ 0.23
falcon-3-7b	0.42 $\pm$ 0.26	0.48 $\pm$ 0.24
gemma-3-4b-it	0.14 $\pm$ 0.33	0.27 $\pm$ 0.29
llama-3.1-8b	0.40 $\pm$ 0.27	0.41 $\pm$ 0.28
mistral-7b	0.25 $\pm$ 0.32	0.16 $\pm$ 0.35
Concise Prompt + CoT		
gpt5	0.49 $\pm$ 0.25	0.43 $\pm$ 0.27
gpt4o	0.48 $\pm$ 0.25	**0.55** $\pm$ 0.24
qwen3-next-80b	0.40 $\pm$ 0.27	0.46 $\pm$ 0.28
deepseek-chat-v3.1	0.38 $\pm$ 0.28	0.48 $\pm$ 0.23
gpt4o-mini	**0.50** $\pm$ 0.24	0.52 $\pm$ 0.23
falcon-3-7b	0.44 $\pm$ 0.28	0.48 $\pm$ 0.25
llama-3.1-8b	0.38 $\pm$ 0.27	0.45 $\pm$ 0.27
gemma-3-4b-it	0.12 $\pm$ 0.33	0.46 $\pm$ 0.24
mistral-7b	0.35 $\pm$ 0.33	0.37 $\pm$ 0.27
Full Guidelines + CoT		
gpt4o	0.42 $\pm$ 0.27	**0.59** $\pm$ 0.22
gpt5	0.47 $\pm$ 0.26	0.44 $\pm$ 0.26
qwen3-next-80b	0.39 $\pm$ 0.29	0.46 $\pm$ 0.29
deepseek-chat-v3.1	0.45 $\pm$ 0.26	0.41 $\pm$ 0.27
gpt4o-mini	**0.54** $\pm$ 0.22	0.50 $\pm$ 0.22
falcon-3-7b	0.48 $\pm$ 0.25	0.48 $\pm$ 0.24
gemma-3-4b-it	0.29 $\pm$ 0.31	0.43 $\pm$ 0.26
llama-3.1-8b	0.29 $\pm$ 0.31	0.45 $\pm$ 0.26
mistral-7b	0.43 $\pm$ 0.28	0.37 $\pm$ 0.28

Table 12: Informativeness Evaluation approaches comparison

	Context	Elaboration
1	Finn came up with a different explanation: Students cannot hide in the back of the classroom in smaller classes. They behave better and are more involved. He saw the change himself visiting classrooms in Buffalo, New York. Smaller, quieter classes may have their biggest effect on kids who do not pay attention and try to avoid looking the teacher in the eye. That's because they cannot hide.	But there was another question.
2	Steele is a scuba diver who has scooped up the animals from the seafloor since the 1970s. He sells the urchins to sushi restaurants. Steele heard about the sea getting more acidic. He saw right away what it could mean for his business and the ocean he loves. So Steele told Hofmann about the urchins.	The scientist started looking into his worries.
3	People listen to her, said Pastor Henry Logan, who has been working with her since August. "She does it one meal at a time." Last Monday, violent protests broke out again in Ferguson. A grand jury decided not to charge the officer who shot Brown. A grand jury is made up of a group of people.	They decide if a person should be charged with a crime.
4	He and his team named her Lucy after a song by the Beatles. The song played over and over the night her bones were found. Johanson is now the head of the Institute of Human Origins at Arizona State University. He spoke about how Lucy's discovery changed what scientists thought about early humans. He also discussed what he hopes to find next.	Newsela has adapted the answers given by Johanson.
5	They only made $27 more than what the city spent to put them on the streets. The Pasadena meters did not cost the city money. They were designed by college students. They were paid for by money that companies gave to the city, Huang said. City leaders say the money collected by the meters might help in finding people homes.	The charities have already proven to be very helpful.

Table 13: Incohesive Golden Elaborations