M²PRESERVE: A Multidimensional Framework for Evaluating Meaning Preservation in Text Simplification

Abdullah Barayan[1,2], Jose Camacho-Collados[1] and Fernando Alva-Manchego[1]
[1] School of Computational and Mathematical Sciences, Cardiff University, UK
[2] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
{barayanas,camachocolladosj,alvamanchegof}@cardiff.ac.uk

Abstract

Evaluating meaning preservation in text simplification remains an open challenge: human evaluations rely on coarse-grained holistic judgments, while automatic metrics frequently misalign with human assessments and fail to capture simplification-specific edits. We introduce M²PRESERVE, a multidimensional fine-grained framework that decomposes meaning preservation into two complementary dimensions, *Completeness* and *Faithfulness*, defined over key facts aligned between original and simplified texts. To support evaluation, we construct M²PRESERVEEVAL, a paragraph-level dataset with 8,773 human-annotated key-fact instances and high inter-annotator agreement. With this dataset, we benchmark a wide range of automatic metrics and evaluate M²PRESERVE-LLM as an automated evaluator. Results show that conventional metrics are weak indicators of *Completeness*, whereas M²PRESERVE-LLM produces more structured and interpretable assessments. Further analysis reveals that simplification systems vary more in how much source content they retain than in how faithfully they render the facts they include, pointing to content selection and fact retention as central challenges in simplification.

AbdullahSalemBar/M2Preserve

1 Introduction

Text simplification (TS) aims to reduce the linguistic and conceptual complexity of a text to facilitate accessibility and comprehension for a target audience while preserving its original meaning and relevant information (Al-Thanyyan and Azmi, 2021; Alva-Manchego et al., 2020b). It is essential for educational technology, aiding readers with cognitive or language impairments (Madjidi and Crick, 2025) and content adaptation for second-language learners (Alva-Manchego et al., 2025). Although substantial progress has been made in the development of automatic simplification systems, reliably evaluating whether a simplified text preserves meaning remains a fundamental challenge.

Human evaluation of meaning preservation typically relies on holistic ratings that offer limited diagnostic insight into *how* meaning is preserved or lost, and may suffer from subjectivity and inconsistency (Muñoz-García et al., 2025; Grabar and Saggion, 2022). On the automatic side, evaluation commonly relies on metrics adapted from machine translation and summarization. However, these metrics often misalign with human judgments in TS (You and Guo, 2026; Agrawal and Carpuat, 2024; Beauchemin et al., 2023; Devaraj et al., 2022) and misinterpret simplification-specific operations, such as paraphrasing, elaboration, or appropriate content omission, producing misleading assessments of meaning preservation (You and Guo, 2026; Barayan et al., 2025; Kew et al., 2023).

Prior work in summarisation and factuality evaluation has moved beyond holistic judgments toward fine-grained, alignment-based assessment using atomic facts (Liu et al., 2023b; Min et al., 2023). Multidimensional frameworks such as FineSurE (Song et al., 2024) have further shown the value of decomposing evaluation into distinct, interpretable dimensions. Building on these advances, we introduce M²PRESERVE, a framework that brings this fine-grained, multidimensional perspective to the evaluation of meaning preservation in TS. M²PRESERVE decomposes meaning preservation into two dimensions: *Completeness*, which measures whether essential source content is retained, and *Faithfulness*, which assesses whether the simplified text remains accurate and free of unsupported or distorted information. Both dimensions are operationalised through *key-fact alignment*: atomic facts are extracted from both the original and simplified texts and explicitly aligned, while unaligned facts are classified into fine-grained categories that distinguish omissions, valid elaborations, and factual errors. In TS, source-unsupported

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 113–138
October 17-21, 2026 ©2026 Association for Computational Linguistics

additions are not necessarily errors, as definitions, examples, or background information may improve accessibility through elaborative simplification (Srikanth and Li, 2021). Tracking elaborations alongside errors allows the framework to separate legitimate expansions of content from genuine inaccuracies.

Building on this framework, we construct M^{2}PRESERVEEVAL, a human-annotated 8.7k key-fact benchmark for evaluating paragraph-level meaning preservation in TS, and analyze its annotations to show that systems differ more in source-fact retention than in factual grounding (Section 4). We then introduce M^{2}PRESERVE-LLM, an LLM-based adaptation of the framework under single-judge and jury-based paradigms, providing more structured and interpretable assessments than conventional metrics (Section 5). Finally, we use M^{2}PRESERVEEVAL to benchmark automatic metrics, showing that conventional metrics are weak indicators of *Completeness*, and further validate M^{2}PRESERVE-LLM on plain language summarization, a related task with elaborative content, showing that it is more robust to legitimate elaborations than factuality metrics (Section 6).

2 Related Work

Human Evaluation. Human evaluation of meaning preservation in TS has typically relied on holistic ratings, often using Likert-style or direct assessment scales (Wu and Arase, 2026; Sottana et al., 2023; Maddela et al., 2021; Sun et al., 2021; Alva-Manchego et al., 2020b). While easy to collect, such ratings can be subjective and inconsistent, as annotators may apply scales differently (Muñoz-García et al., 2025; Grabar and Saggion, 2022). Binary judgments reduce ambiguity but sacrifice granularity (Cripwell et al., 2023, 2024), especially beyond the sentence level, where information may be preserved, omitted, or altered unevenly. More diagnostic approaches, such as reading-comprehension-based evaluation (Agrawal and Carpuat, 2024), provide richer evidence but increase annotator burden and may be influenced by readers' comprehension ability. Moreover, it does not provide a unified fine-grained characterization of meaning change that distinguishes omissions from additions.

Automatic Evaluation Metrics. Automatic evaluation of meaning preservation in TS commonly relies on metrics adapted from machine translation and summarization, including lexical overlap metrics such as BLEU (Papineni et al., 2002), semantic similarity metrics such as BERTScore (Zhang et al., 2020), and factual consistency metrics such as AlignScore (Zha et al., 2023). However, these metrics often misalign with human judgments in TS (You and Guo, 2026; Agrawal and Carpuat, 2024; Beauchemin et al., 2023; Devaraj et al., 2022) because they are not designed for simplification-specific edits such as paraphrasing, elaboration, or appropriate content omission (You and Guo, 2026; Barayan et al., 2025; Kew et al., 2023). Recent metrics more directly target meaning preservation or factuality, including MeaningBERT (Beauchemin et al., 2023), LogProp (Pauli et al., 2025), and PlainQAFact (You and Guo, 2026). While these approaches are important advances, they remain holistic, sentence-level, or primarily factuality-focused, and therefore do not provide a unified fine-grained assessment of both completeness and faithfulness.

3 M^{2}PRESERVE Framework

M^{2}PRESERVE operationalises meaning preservation through *key-fact alignment*: atomic facts, defined as discrete, contextually independent units of meaning, are extracted from both the original and simplified texts and explicitly aligned with their counterparts. Unaligned facts are then classified into fine-grained categories that distinguish omissions, valid elaborations, and factual errors, which in turn determine the scoring of *Completeness* and *Faithfulness*. Figure 1 provides an overview of the framework. In the following, we detail each step of this annotation and scoring process.[1]

3.1 Completeness Evaluation

In this evaluation setting, annotators are shown (i) the original complex text, (ii) a list of key facts extracted from it, and (iii) the simplified text. For each key fact, annotators first determine whether it is inferable from the simplified text. If the fact is inferable, the annotator links it to the supporting sentence(s), making the alignment explicit. If the fact is not inferable, annotators classify it into one of the following categories:

- **Wrong Key Fact:** The fact was not correctly extracted from the original text. This category serves as a data quality filter rather than a simplification judgment.

[1]The annotation interface is shown in Appendix A.

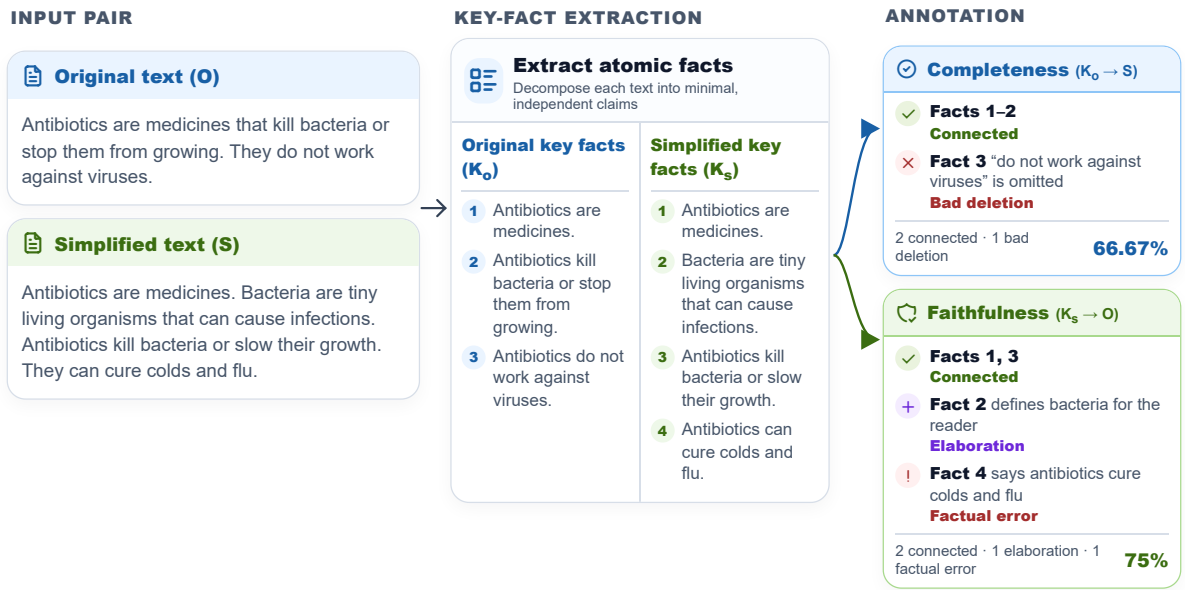


Figure 1: M^{2}PRESERVE framework: Atomic facts are extracted, aligned, and classified to score *Completeness* and *Faithfulness*. In this example, the *Completeness* score is 66.67%, since two out of the three original key facts are preserved, while one important fact is omitted; the *Faithfulness* score is 75%, since three out of the four simplified key facts are acceptable, including two connected facts and one explanatory elaboration.

- **Bad Deletion:** The fact is absent from the simplified text, and its omission obscures the main idea of the original.
- **Good Deletion:** The fact is absent from the simplified text, but its omission does not obscure the main idea of the original.

3.2 Faithfulness Evaluation

In the reverse setup, annotators are shown (i) the simplified text, (ii) a list of key facts extracted from it, and (iii) the original complex text. For each key fact, annotators determine whether it is accurately supported by the original text. If supported, the annotator links it to the supporting sentence(s), making the alignment explicit. Otherwise, the fact is classified into one of the following categories:

- **Wrong Key Fact:** The fact was not correctly extracted from the simplified text. As in the completeness evaluation, this category serves as a data quality filter.
- **Factual Error:** The fact introduces misleading, incorrect, or fabricated information not present in the original text.
- **Elaboration:** The fact provides additional details (e.g., examples, background information, or definitions) that are accurate but not explicitly mentioned in the original text. Elaborations are not penalised in the faithfulness score, as they represent legitimate additions rather than inaccuracies.

3.3 Scoring

Both scores are computed as proportions of key facts that are either explicitly aligned or assigned a category that does not penalise the system, scaled to a 0–100 range.

Formulation. Let **Connected**(k, T), **GoodDel**(k) and **Elab**(k) be indicator predicates that are true when key fact (k) is aligned to text (T), classified as Good Deletion, or classified as Elaboration, respectively. Then:

$$C_{\text{pres}} = \left|\{k \in K_O : \textbf{Connected}(k, S) \vee \textbf{GoodDel}(k)\}\right|$$

$$F_{\text{sup}} = \left|\{k' \in K_S : \textbf{Connected}(k', O) \vee \textbf{Elab}(k')\}\right|$$

where:

- C_{pres} — number of original key facts preserved in the simplification,
- F_{sup} — number of simplified-text key facts supported by the original,
- K_O and K_S — key-fact sets extracted from the original and simplified texts, respectively.

The final scores are then computed as:

$$\text{Completeness} = \frac{C_{\text{pres}}}{|K_O| - |\mathsf{WKF}_O|} \times 100$$

$$\text{Faithfulness} = \frac{F_{\text{sup}}}{|K_S| - |\mathsf{WKF}_S|} \times 100$$

where:

- WKF_O and WKF_S — key facts flagged as **Wrong Key Fact**, excluded from the calculation to avoid penalizing extraction errors.

4 M²PRESERVEEVAL Dataset

4.1 Source Text Collection

We constructed M²PRESERVEEVAL using 60 source paragraphs drawn from four publicly available English readability and simplification datasets: (1) *OneStopEnglish* (Vajjala and Lučić, 2018), (2) *ELG-CEFR-En* (Breuker, 2022), (3) *Cambridge-Exams* (Xia et al., 2016), both drawn from the English reference subset of *UniversalCEFR* (Imperial et al., 2025),[2] and (4) the TSAR 2025 Shared Task source paragraphs (Alva-Manchego et al., 2025), drawn from the British Council's LearnEnglish website.[3] For the latter three sources, we extracted paragraphs from texts at CEFR level B2 and above, while *OneStopEnglish* paragraphs were taken from advanced-level texts. We selected these resources because they provide proficiency-labelled English passages suitable for challenging source-text selection, while also offering diversity in topic, writing style, and source type. While *OneStopEnglish* and TSAR are associated with human simplifications, *ELG-CEFR-En* and *Cambridge-Exams* contain no paired simplification references. Including these unpaired sources reduces the risk that evaluated models reproduce memorized reference simplifications, following the motivation of SIMPEVAL2022 (Maddela et al., 2023). On average, M²PRESERVEEVAL paragraphs contain 99 words and 4–5 sentences, with a Flesch–Kincaid grade level of around 12 and predicted CEFR levels mostly in the B2–C1 range, providing challenging source material for evaluating TS.[4] This profile is similar in length and difficulty to existing datasets, while covering a more diverse set of sources. Detailed source-text statistics and comparisons with existing datasets are provided in Appendix B.

4.2 Simplification Generation

We generated simplifications for the 60 source paragraphs using five models: four instruction-tuned LLMs (GPT-4o, Gemini 2.0 Flash, Gemma 3-12B (Gemma Team, 2025), and Qwen 2.5-14B (Qwen Team, 2024; Yang et al., 2024)) and BART-Swipe (Laban et al., 2023), which is a document-level simplification model fine-tuned on SWiPE. For the instruction-tuned LLMs, we used a prompt template commonly adopted in TS literature (Kew et al., 2023; Maddela et al., 2023; Sottana et al., 2023), based on the ASSET crowdworker guidelines (Alva-Manchego et al., 2020a).[5]

4.3 Key-Fact Extraction and Filtering

Key-Fact Extraction. Key facts were automatically extracted using GPT-4o, guided by the Russellian–Neo-Davidsonian (R-ND) in-context learning (ICL) prompt introduced by Wanner et al. (2024). We selected this configuration after comparing two instruction-tuned LLMs, GPT-4o and Gemini 2.0 Flash, under two ICL prompting strategies: R-ND and FactScore (Min et al., 2023).[6] Table 1 reports the number of extracted key facts for the original texts and each simplification system. Original texts contain the most facts on average (15.9), while simplified outputs generally contain fewer facts, reflecting content deletion or consolidation. BART-Swipe has the lowest average number of facts (10.8), suggesting more aggressive simplification, while Gemma-3-12b-it has the highest average among the simplification systems (15.1), closely approaching the original texts.

Key-Fact Filtering. To identify unsupported (i.e. hallucinated) key facts, we validated each extracted fact against its corresponding text using three factuality detection models: MiniCheck (Tang et al., 2024), HHEM-2.1-Open (Bao et al., 2024), and AlignScore-large (Zha et al., 2023). Facts rejected by the majority of models were flagged for manual inspection. Of 4,965 unique facts, only 40 were flagged as potentially unsupported (0.81%). After manual review, 39 were confirmed valid and only one required adjustment. Annotators were still instructed to validate key facts during annotation, as

[2] In *UniversalCEFR*, *reference* denotes texts created by experts, teachers, or language-learning professionals.

[3] https://learnenglish.britishcouncil.org/

[4] Flesch–Kincaid scores were computed using `textstat`, and CEFR levels were predicted using the TSAR 2025 Shared Task evaluation model.

[5] The prompt is provided in Appendix B.2.

[6] Further details are provided in Appendix B.3.

Text	Total	Avg. Facts
Original	952	15.9
gemini-2.0-flash	789	13.2
gemma-3-12b-it	905	15.1
GPT-4o	807	13.5
Qwen-2.5-14b	863	14.4
BART-Swipe	649	10.8
Total Facts	4,965	

Table 1: Fact extraction statistics for the original text and model simplifications.

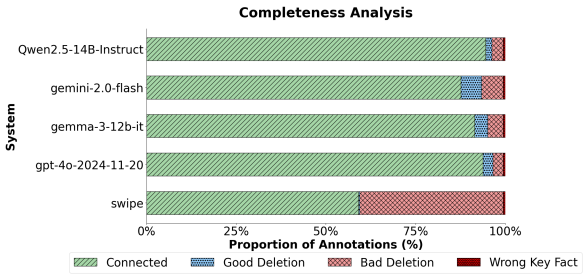


Figure 2: Distribution of *Completeness* key-fact annotation labels across systems: Connected, Good Deletion, Bad Deletion, and Wrong Key Fact.

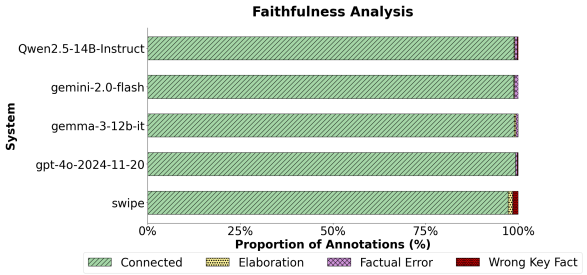


Figure 3: Distribution of *Faithfulness* key-fact annotation labels across systems: Connected, Elaboration, Factual Error, and Wrong Key Fact.

described in Section 3.

4.4 Annotation Procedure

We recruited fluent English speakers to evaluate the text-pair samples. To ensure annotators understood the framework, they first completed a 10-pair qualification task and were required to achieve at least 90% accuracy. Only those who qualified annotated the main dataset, with each instance receiving three independent annotations. The final label for each key fact was assigned by majority vote. In cases where no majority decision was reached, one of the study authors acted as an adjudicator to review the discrepancy and assign the final label.

Type	IAA	Completeness	Faithfulness
Score-level	Gwet's AC2	0.985	0.987
	Krippendorff's α (interval)	0.937	0.496
	ICC(2,k)	0.978	0.748
key fact-level	Gwet's AC1	0.933	0.988
	Krippendorff's α (nominal)	0.777	0.57
	Fleiss' κ	0.777	0.57
Majority Agreement	Proportion (%)	99.64%	99.95%

Table 2: Inter-annotator agreement for *Completeness* and *Faithfulness*, reported at the score level, key fact-level, and majority agreement proportions.

4.5 Annotation Quality Assessment

To assess annotation reliability, we measure inter-annotator agreement (IAA) for both *Completeness* and *Faithfulness*. Because the label distribution is highly skewed, with most key facts marked as *Connected* and many instances receiving perfect scores, especially for *Faithfulness*, we use Gwet's coefficients as our primary reliability measures (Lee et al., 2024). Specifically, we report AC2 for interval-scale scores and AC1 for categorical key-fact operations. Additionally, we report Krippendorff's α, ICC(2,k) and Fleiss' κ for completeness with previous work. Table 2 shows high agreement across both dimensions. At the score level, the agreement is almost perfect for both *Completeness* (AC2 = 0.985) and *Faithfulness* (AC2 = 0.987). At the key fact-level, agreement also remains high, with AC1 of 0.933 for *Completeness* and 0.988 for *Faithfulness*. In addition, strict majority agreement is almost universal, covering 99.64% of *Completeness* key facts and 99.95% of *Faithfulness* key facts. These results indicate that the M^{2}PRESERVE framework enables highly consistent annotations, supporting its goal of reducing subjectivity in meaning preservation evaluation.

4.6 Analysis of Human Annotations

Figures 2 and 3 present the distribution of human key-fact annotation labels across systems for *Completeness* and *Faithfulness*, respectively.

Completeness. Instruction-tuned LLMs exhibit consistently high *Completeness*, with most source facts preserved as *Connected*. Among these systems, Qwen2.5-14B-Instruct and GPT-4o achieve the highest preservation rates, closely followed by Gemma-3-12b-it, while Gemini-2.0-Flash shows slightly lower *Completeness*. The proportion of *Bad Deletions* remains relatively low for LLM-based systems, suggesting that most content reductions reflect acceptable simplification decisions rather than substantial meaning loss. In contrast,

BART-Swipe shows substantially lower *Completeness*, with only around 60% of the facts labeled as *Connected* and a much higher proportion of *Bad Deletions*, reflecting its more aggressive compression behavior and the lower number of facts observed in Table 1.

Faithfulness. All systems demonstrate near-perfect *Faithfulness*, with almost all simplified facts mapped back to the source (*Connected*). Factual errors are extremely rare across models. Minor differences appear in the rate of *Elaborations*, with BART-Swipe producing slightly more additions than the instruction-tuned LLMs, while GPT-4o shows virtually no factual errors.

Factual error analysis. Manual inspection shows that some factual errors are due to *quote rewriting* rather than substantive hallucination. For example, models sometimes paraphrase quoted speech, such as rewriting *"This is very encouraging"* as *"This is very positive"* or *"This is really good news"*. Although semantically similar, these variants are not supported verbatim by the source and are therefore labeled as *Factual Error*. This behavior could arise from the prompt not explicitly specifying constraints on quoted speech.

Fact Retention vs. Factual Grounding. Overall, systems differ more in how they retain source facts than in whether their generated facts are grounded in the source. While *Faithfulness* remains near ceiling, *Completeness* reveals variation in which facts are preserved, acceptably omitted, or incorrectly deleted. This suggests that, for the systems studied here, meaning-preservation differences are less driven by hallucination and more by content-selection decisions. Future research could explore *fact-aware* simplification systems that incorporate automatic identification of crucial source facts as a signal to guide content retention during the simplification process.

5 M²PRESERVE-LLM

We explore whether LLMs can approximate human judgments of meaning preservation under the M²PRESERVE framework. In this setting, key facts are provided, and the LLM performs only alignment and labeling, acting as a metric that mirrors the human annotation process.

We consider two evaluation paradigms. In the **LLM-as-a-Judge** paradigm, a single LLM performs key-fact alignment and assigns operation labels for *Completeness* and *Faithfulness*. While straightforward, a single judge may exhibit systematic biases in deletion or elaboration decisions. Therefore, in the **LLMs-as-a-Jury** paradigm, multiple LLM instances perform the same task independently and their outputs are aggregated by majority voting. This paradigm mirrors multi-annotator setups and examines whether collective LLM decisions better approximate human agreement. We construct the jury over structurally reliable models, using a single model as a fallback adjudicator when no strict majority is reached.[7]

6 Alignment with Human Judgments

Having established human annotations for M²PRESERVEEVAL (Section 4), we evaluate how well automatic metrics, including M²PRESERVE-LLM, align with human judgments.

6.1 Automatic Evaluation Metrics

We compare M²PRESERVE-LLM against diverse automatic metrics commonly used or proposed to assess meaning preservation, adequacy or factuality on M²PRESERVEEVAL. These cover lexical similarity (BLEU (Papineni et al., 2002)), semantic similarity (BERTScore (Zhang et al., 2020)), meaning preservation (MeaningBERT (Beauchemin et al., 2023), LogProp (Pauli et al., 2025)), factual consistency (AlignScore (Zha et al., 2023), MiniCheck (Tang et al., 2024), HHEM-2.1-Open (Bao et al., 2024), PlainQAFact (You and Guo, 2026)), and LLM-based evaluation (G-Eval (Liu et al., 2023a)). We also include the Agg-Metric framework (Maddela and Alva-Manchego, 2025), which adapts sentence-level metrics to document-level simplification.[8]

6.2 Evaluation Setup

Following WMT24 (Freitag et al., 2024), we evaluate metric alignment with human judgments at the system and segment levels. At the **system level**, we use *Soft Pairwise Accuracy (SPA)* (Thompson et al., 2024) to assess whether metrics correctly rank simplification systems. At the **segment level**, we use *groupby-item segment-level accuracy with tie calibration* ($\mathrm{acc}^{*}_{\mathrm{eq}}$) (Deutsch et al., 2023) to test whether metrics correctly rank simplifications of the same source segment. As an addi-

[7] Details, prompts, and evaluated LLMs are provided in Appendix C.1.

[8] Further details on the metrics, prompts, and implementation settings are provided in Appendix C.1.

	Metric	PDP	acc^*_{eq}	SPA	bAcc
COMPLETENESS	Agg-BERTScore	0.21	0.48	0.59	–
	G-Eval: Qwen3-4B-Instruct	0.18	0.53	0.89	–
	G-Eval: Qwen3-4B-Thinking	0.66	0.60	0.67	–
	G-Eval: gemma-3-4b-it	0.57	0.59	**0.98**	–
	G-Eval: gpt-4.1	**0.95**	**0.78**	0.90	–
	G-Eval: gpt-5.4	0.93	0.76	0.89	–
	G-Eval: gemini-3.1-pro	0.93	**0.78**	0.76	–
	M^2P: Qwen3-4B-Instruct	0.89	0.74	0.88	0.89
	M^2P: Qwen3-4B-Thinking	0.65	0.63	0.93	0.92
	M^2P: gpt-4.1	0.82	0.72	0.92	**0.95**
	M^2P: gpt-5.4	0.60	0.66	0.76	0.94
	M^2P: gemini-3.1-pro	0.66	0.63	0.77	**0.95**
	M^2P: LLM-as-a-jury	0.93	0.77	0.93	**0.95**
FAITHFULNESS	MiniCheck	0.04	0.90	0.84	–
	MiniCheck- by sent.- prob	0.04	0.90	0.81	–
	HHEM-2.1- by sent.	0.13	0.90	0.77	–
	G-Eval: Qwen3-4B-Thinking	-0.04	0.90	0.54	–
	G-Eval: gpt-5.4	0.40	0.92	0.63	–
	M^2P: Qwen3-4B-Thinking	0.29	0.90	0.73	0.78
	M^2P: gpt-4.1	0.05	0.90	0.79	**0.83**
	M^2P: gpt-5.4	0.38	0.90	**0.86**	0.81
	M^2P: gemini-3.1-pro	**0.71**	**0.93**	0.79	0.80
	M^2P: LLM-as-a-jury	0.52	0.90	0.80	0.72

Table 3: Automatic metric alignment with human evaluation on M^2PRESERVEEVAL. M^2P denotes M^2PRESERVE-LLM. *prob* denotes using the model's predicted probability as the score, while *by sent.* indicates sentence-level scoring followed by aggregation.

tional segment-level analysis, we report *Pairwise Difference Pearson (PDP)* (DiIanni and Deutsch, 2025), which evaluates whether differences in metric scores between simplifications reflect the magnitude of differences in human judgments. Finally, for key fact-level evaluation, we use *Balanced Accuracy (bAcc)* (Brodersen et al., 2010) to measure connected vs. non-connected classification while accounting for class imbalance.

6.3 Results

Table 3 summarizes the main results, while the full results are reported in Appendix Table 11.

Completeness. Results show that conventional metrics are weak indicators of *Completeness*. Agg-BERTScore is the strongest conventional metric in the summary table, but still reaches only 0.21 PDP, 0.48 acc^*_{eq}, and 0.59 SPA. The full results in Appendix Table 11 confirm this pattern: across most conventional metrics, segment-level performance is tightly compressed, with acc^*_{eq} mostly between 0.46 and 0.49 and PDP values near zero or negative, while system-level ranking is also weak. In contrast, G-Eval is already a strong baseline for *Completeness*, even with smaller open backbones, though its performance still varies across backbones and evaluation views. M^2PRESERVE-LLM does not remove this variation, but it makes the signal more structured by evaluating explicit coverage over key facts rather than assigning a single holistic score. This yields more interpretable *Completeness* judgments and supports key fact-level evaluation while remaining competitive with G-Eval. The per-class results in Table 4 show that the main challenge for M^2PRESERVE-LLM is calibrating deletion decisions rather than recognizing *Connected* facts. Most judges identify preserved facts reliably, whereas their behavior differs substantially on *Good Deletion* and *Bad Deletion* classification. `Qwen3-4B-Instruct` tends to be conservative, achieving the strongest performance on *Bad Deletion* while rarely predicting *Good Deletion*. By contrast, `gpt-5.4` and `gemini-3.1-pro` are more permissive, achieving very high recall on *Good Deletion* but substantially weaker performance on *Bad Deletion*.

Faithfulness. *Faithfulness* shows a different pattern from *Completeness*. Because *Faithfulness* is near ceiling, a large share of pairwise comparisons are effectively ties. Since acc^*_{eq} rewards metrics for correctly predicting both rankings and ties after tie calibration, many metrics end up with similarly high values, making acc^*_{eq} not very discriminative. PDP and SPA therefore provide a clearer view in this setting. Conventional metrics, especially factuality metrics, show competitive performance, with MiniCheck achieving the strongest SPA among them (0.84). By contrast, M^2PRESERVE-LLM provides more informative *Faithfulness* signals overall. Among single-model judges, `gpt-5.4` achieves the highest SPA (0.86) and strong key fact-level performance (bAcc = 0.81), while `gemini-3.1-pro` achieves the strongest segment-level correlation, with the highest PDP (0.71) and acc^*_{eq} (0.93). However, factuality metrics can be sensitive to elaborative explanations that add external content to improve readability (You and Guo, 2026; Barayan et al., 2025). Since such cases are rare in M^2PRESERVEEVAL, we further test this in a plain language summarization setting where elaboration is more frequent (Section 6.4).

Format Reliability. Most M^2PRESERVE-LLM judges followed the expected output format. The main exception was `phi-4`, which returned a mismatched number of labels for 37.7% of *Complete-*

Model	Faithfulness				Completeness			
	Conn.	Elab.	Fact Err.	WKF	Conn.	Bad Del.	Good Del.	WKF
Qwen3-4B-Inst.	0.96/0.93	0.02/0.06	0.09/0.40	0.13/0.15	0.94/0.91	0.71/0.91	0.03/0.02	0.02/0.07
Qwen3-4B-Th.	0.98/0.97	0.21/0.31	0.24/0.56	0.00/0.00	0.95/0.92	0.52/0.46	0.24/0.59	0.00/0.00
gpt-4.1	0.98/0.96	0.18/0.69	0.20/0.32	0.00/0.00	0.96/0.94	0.64/0.72	0.23/0.34	0.00/0.00
gpt-5.4	0.97/0.95	0.20/0.50	0.24/0.52	0.02/0.08	0.96/0.93	0.32/0.20	0.26/0.93	0.35/0.27
gemini-3.1-pro	0.99/0.99	0.40/0.63	0.63/0.80	0.00/0.00	0.97/0.95	0.31/0.19	0.28/0.94	0.00/0.00
claude-sonnet-4.6	0.98/0.98	0.31/0.44	0.15/0.56	0.00/0.00	0.97/0.94	0.57/0.43	0.32/0.87	0.00/0.00
LLM-as-a-jury	0.99/0.98	0.29/0.50	0.45/0.68	0.04/0.07	0.97/0.95	0.50/0.35	0.30/0.89	0.11/0.07

Table 4: Per-class F1/Recall scores for M^{2}PRESERVE-LLM. Conn. = Connected, Elab. = Elaboration, Fact Err. = Factual Error, WKF = Wrong Key Fact, Bad Del. = Bad Deletion, and Good Del. = Good Deletion.

ness instances and 35.0% of *Faithfulness* instances. The issue was much rarer for `gemma-3` (3.0% and 2.3%, respectively). These errors mainly involved missing, extra, or empty labels.

LLM-as-a-Jury. The M^{2}PRESERVE-LLM results suggest that relying on a single LLM judge is not always optimal, since different backbones show complementary strengths across dimensions and labels. This motivates an LLM-as-a-jury setting, in which multiple judges contribute operation labels, and the final score is computed from their aggregated decisions. As a simple baseline, we construct a majority-vote jury over six structurally reliable models, excluding `phi-4` and `gemma-3` due to annotation-format reliability issues, and use `gpt-5.4` as a fallback adjudicator when no strict majority is reached. Table 3 shows that the jury is more beneficial for *Completeness* than for *Faithfulness*. Within the M^{2}PRESERVE-LLM setting, the jury achieves the strongest completeness PDP (0.93) and acc^*_{eq} (0.77), and ties the best SPA (0.93) and bAcc (0.95), indicating that aggregation improves robustness in all evaluation metrics. For *Faithfulness*, the jury also performs competitively (PDP = 0.52, SPA = 0.80), but the improvement over single-model judges is less pronounced than for *Completeness*. The per-class results in Table 4 help explain this pattern. For *Completeness*, individual judges show different deletion-label biases: conservative models tend to favor *Bad Deletion*, whereas more permissive models tend to favor *Good Deletion*. Majority voting reduces the influence of any single bias, yielding a more balanced signal. However, because `gpt-5.4` serves as the fallback adjudicator, unresolved cases are still often resolved toward *Good Deletion*. For *Faithfulness*, the jury remains strong on *Connected* facts and competitive on *Factual Error*, but still struggles on *Elaboration* and *Wrong Key Fact*. This suggests that jury aggregation is less helpful for faithfulness because the remaining problem is not broad bias reduction but distinguishing fine-grained labels.

Metric	τ	r	AUC (99% CI)	r_s ext.
MiniCheck- by sent.- prob	**0.66**	0.79	**0.966** [0.945, 0.983]	-0.59***
PlainQAFact	**0.66**	0.81	0.963 [0.935, 0.985]	-0.33***
M^{2}P: gpt-5.4	**0.66**	**0.82**	0.962 [0.935, 0.984]	**-0.09**
MiniCheck- by sent.	**0.66**	0.79	0.961 [0.938, 0.980]	-0.57***
MiniCheck- prob	0.51	0.63	0.857 [0.804, 0.902]	-0.17*
MiniCheck	0.58	0.58	0.785 [0.730, 0.837]	-0.14
HHEM-2.1- by sent.- prob	0.40	0.48	0.783 [0.722, 0.838]	-0.65***
HHEM-2.1- by sent.	0.39	0.48	0.775 [0.714, 0.831]	-0.64***
HHEM-2.1- prob	0.22	0.27	0.653 [0.582, 0.724]	-0.50***
HHEM-2.1	0.20	0.20	0.550 [0.522, 0.582]	-0.29***

Table 5: Meta-evaluation results on PLAINFACT. r_s ext. denotes the correlation between factual-summary metric scores and the external-sentence ratio; more negative values indicate stronger penalization of elaborative content. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

6.4 Elaboration Sensitivity

Because elaborations are relatively rare in M^{2}PRESERVEEVAL, we further test *Faithfulness* in a setting where elaborative content is more frequent. We use PLAINFACT (You and Guo, 2026), a plain language summarization benchmark containing 400 summary–abstract pairs (200 factual and 200 non-factual), with 44% of PLS sentences annotated as elaborative explanations. Following You and Guo (2026) for comparability, we report Kendall's τ, Pearson, and AUC-ROC as our main meta-evaluation metrics. We additionally report Spearman correlation between metric scores for factual summaries and the elaborative-sentence ratio to quantify sensitivity to elaboration. Table 5 shows that MiniCheck sentence-level variants, PlainQAFact, and M^{2}PRESERVE-gpt-5.4 achieve similar top factual/non-factual discrimination, with Kendall's τ around 0.66 and AUC

between 0.961 and 0.966. However, MiniCheck and HHEM sentence-level variants show strong negative correlations with elaborative content, suggesting over-penalization of valid elaboration. In contrast, M^{2}PRESERVE-gpt-5.4 shows only weak elaboration sensitivity ($r_s = -0.09$), while maintaining near-top discrimination. This suggests a better balance between factuality detection and robustness to valid elaboration.

7 Conclusion

We introduced M^{2}PRESERVE, a framework for evaluating meaning preservation in TS via key-fact alignment, decomposing the problem into *Completeness* and *Faithfulness*. To support evaluation, we constructed M^{2}PRESERVEEVAL, a paragraph-level dataset with 8.7k human-annotated key-fact instances and high inter-annotator agreement. Analysis reveals that systems differ more in source-fact retention than in factual grounding, pointing to content selection as a challenge in simplification evaluation. Metric benchmarking further shows that conventional metrics are largely uninformative of *Completeness*, whereas M^{2}PRESERVE-LLM provides more structured and interpretable assessments. Validation on plain language summarization further shows that M^{2}PRESERVE-LLM is more robust to legitimate elaborations than factuality metrics, suggesting that fine-grained, alignment-based evaluation is necessary for capturing the full range of meaning-preservation phenomena in simplification.

Limitations

Four limitations of the current work are worth noting. First, M^{2}PRESERVE depends on the quality of key-fact extraction. In the current evaluation pipeline, key facts are automatically extracted with GPT-4o using the R-ND in-context learning prompt, filtered with factuality models, and then validated during human annotation. Although this process worked well in our experiments, errors or biases in key-fact decomposition may still affect downstream scoring and interpretation. Future work could reduce this dependence through stronger decomposition models, improved prompting strategies, and more robust validation pipelines.

Second, M^{2}PRESERVEEVAL remains limited in scale and coverage. Although it contains 8.7k human-annotated key facts, these are derived from only 60 English source paragraphs and simplifications produced by five systems. As a result, the benchmark may not fully capture the diversity of simplification phenomena across domains, languages, document lengths, and model behaviors.

Third, the current M^{2}PRESERVE-LLM framework relies solely on the internal knowledge of the underlying LLM and does not incorporate external knowledge retrieval. Although M^{2}PRESERVE-LLM performs strongly on both our main benchmark and the additional plain language summarization setting, retrieval may become important when evaluating elaborations that depend on domain-specific background knowledge. Future work could investigate retrieval-augmented variants of M^{2}PRESERVE-LLM that ground evaluation in external evidence and better support the assessment of valid explanatory content.

Fourth, validation of M^{2}PRESERVE-LLM against independently collected human judgments is currently limited to the *Faithfulness* dimension. Although PLAINFACT enables out-of-domain evaluation of the distinction between valid elaborations and factual errors, existing TS datasets provide only holistic meaning-preservation scores and therefore cannot independently validate the proposed decomposition into *Completeness* and *Faithfulness*. Comparable external validation of *Completeness* remains open.

Despite these limitations, we hope our work will provide valuable insights into text simplification evaluation and support the development of more reliable and interpretable evaluation methods.

Ethics Statement

The collection of manual annotations received a favourable opinion from the Ethics Committee of the School of Computer Science and Informatics at Cardiff University. Annotators were informed about the purpose of the study, how their annotations would be used for research and data set creation, and that no personally identifiable information would be released. We actively addressed the annotators' queries throughout the annotation process, ensuring clear and transparent communication at all stages. Annotators were compensated £25 for every 50 completed rating instances, up to a total of £300 for all 600 instances, and received bonuses for consistent, high-quality work. Our released dataset excludes any information that could disclose annotators' personal details. The source texts were drawn from publicly available or research-licensed resources: *ELG-CEFR-En* and

Cambridge-Exams are distributed under CC BY-NC-SA 4.0, *OneStopEnglish* under CC BY-SA 4.0, and the TSAR 2025 source texts were made available for research use as part of the shared task with permission from the British Council.

Supplementary Materials Availability Statement: The M^{2}PRESERVEEVAL dataset, annotation interface, and code used in this work are publicly available at `https://github.com/AbdullahSalemBar/M2Preserve`.

References

Marah Abdin, Jyoti Aneja, Harkirat Singh Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C'esar Teodoro Mendes, Anh Hong Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *ArXiv*, abs/2412.08905.

Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2).

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020a. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020b. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 116–130, Suzhou, China. Association for Computational Linguistics.

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

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.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. Meaningbert: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Mark Breuker. 2022. CEFR labelling and assessment services. In *European Language Grid: A Language Technology Platform for Multilingual Europe*, pages 277–282. Springer International Publishing Cham.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, page 3121–3124, USA. IEEE Computer Society.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

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.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Colten DiIanni and Daniel Deutsch. 2025. Don't sweat the small stuff: Segment-level meta-evaluation based on pairwise difference correlation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25062–25070, Suzhou, China. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 453–463, Avignon, France. ATALA.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. 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.

Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3941–3960, Miami, Florida, USA. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Mounica Maddela and Fernando Alva-Manchego. 2025. Adapting sentence-level automatic metrics for document-level simplification evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6444–6459, Albuquerque, New Mexico. Association for Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

Elham Madjidi and Christopher Crick. 2025. Towards inclusive reading: A neural text generation framework for dyslexia accessibility. In *Proceedings of*

the 11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion, DSAI '24, page 360–367, New York, NY, USA. Association for Computing Machinery.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Kushan Mitra, Dan Zhang, Sajjadur Rahman, and Estevam Hruschka. 2025. FactLens: Benchmarking fine-grained fact verification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18085–18096, Vienna, Austria. Association for Computational Linguistics.

Victoria Muñoz-García, Paloma Moreda, and Manuel Palomar. 2025. Exploring evaluation methods on text simplification: A systematic review. In *2025 7th International Conference on Natural Language Processing (ICNLP)*, pages 277–281.

OpenAI. 2025. GPT-5 System Card. *Computing Research Repository*, arXiv:2601.03267.

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.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2025. Mind the style gap: Meta-evaluation of style and attribute transfer metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21550–21564, Suzhou, China. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.

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.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Xuanxin Wu and Yuki Arase. 2026. An in-depth evaluation of large language models in sentence simplification with error-based human assessment. *ACM Trans. Intell. Syst. Technol.*, 17(4).

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhiwen You and Yue Guo. 2026. Plainqafact: Retrieval-augmented factual consistency evaluation metric for biomedical plain language summarization. *Journal of Biomedical Informatics*, 178:105019.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Annotation Interface.

Annotations were performed in an easy-to-use custom web-based tool that allows annotators to classify each key fact and visually connect it to relevant sentence(s). Each text pair (source, simplified) is evaluated twice – once for completeness and once for faithfulness. Figure 4 and Figure 5 show screenshots of the evaluation of completeness and faithfulness, respectively.

B M^{2}PRESERVEEVAL Dataset

B.1 Source Text Statistics

Table 6 summarizes source-text statistics for M^{2}PRESERVEEVAL and compares them with existing human-evaluated document- and paragraph-level datasets.

B.2 Simplification Generation

Table 7 shows the prompt provided to LLMs to generate simplifications.

B.3 Key-Fact Extraction

We compare two instruction-tuned LLMs, **GPT-4o** and **Gemini 2.0 Flash**, each paired with two key-fact decomposition prompts, the Russellian–Neo-Davidsonian (R-ND) prompt and the FactScore prompt. Tables 16–17 show the R-ND prompt, and Tables 18–19 show the FactScore prompt used for key-fact extraction. Evaluation is conducted on two benchmark datasets with gold human key-fact annotations, namely **REALSumm** (Bhandari et al., 2020) and **ROSE** (Liu et al., 2023b). We adopt the evaluation framework of Mitra et al. (2025), which measures decomposition quality in terms of *Atomicity*, *Sufficiency*, *Coverage*, *Fabrication*, and *Redundancy*. Tables 8 and 9 report detailed metrics for each model–prompt setup, together with human reference decomposition, for both datasets. Across both datasets, setup differences are relatively modest, with all models demonstrating generally comparable levels of fabrication control and high coverage. However, key differences emerge in atomicity. Notably, R-ND prompts tend to produce more atomic facts per document than FactScore, particularly with Gemini. **GPT-4o + R-ND** emerges as a balanced and effective configuration. While Gemini + R-ND generates the highest number of atomic facts, GPT-4o + R-ND maintains competitive atomicity and sufficiency scores while exhibiting lower redundancy and fabrication.

C Evaluating Correlation with Human Judgment

C.1 Automatic Evaluation Metrics

BLEU (Papineni et al., 2002) is a precision-oriented metric that measures n-gram overlap between a candidate and a reference, with a brevity penalty. In our setting, we compute BLEU for each source–simplification pair using SacreBLEU (Post, 2018), treating the source paragraph as the reference.

BERTScore (Zhang et al., 2020) is a semantic similarity metric that computes token-level similarity using contextual embeddings of BERT (Devlin et al., 2019). In our experiments, we compute BERTScore F_1 for each source–simplification pair using the `roberta-large` model, treating the source text as the reference, with `idf=False` and `rescale_with_baseline=True`.

MeaningBERT (Beauchemin et al., 2023) is a reference-less supervised metric based on BERT (Devlin et al., 2019), trained on human meaning preservation judgments, that predicts a meaning preservation score given a complex sentence and its simplified version.

LogProp (Pauli et al., 2025) is a reference-less style-aware evaluation metric that uses token likelihood estimates from language models under different contextual instructions (style rewrite, paraphrase, repetition) to assess both content preservation and style strength. In our experiments, we use the original LogProp implementation with the `meta-llama/Llama-3.1-8B-Instruct` backbone and report the *content preservation* component for each source–simplification pair, using *simplified* as the target style.[9]

AlignScore (Zha et al., 2023) is a factual consistency metric that leverages a RoBERTa-based alignment model to assess the alignment between the claims in generated text and their corresponding source documents. In our experiments, we used the BASE and LARGE variants.[10]

MiniCheck (Tang et al., 2024) is a sentence-level fact-checking model based on Flan-T5-Large that

[9] https://github.com/AmaliePauli/style_transfer_evaluation

[10] AlignScore checkpoints: https://github.com/yuh-zha/AlignScore?tab=readme-ov-file#checkpoints

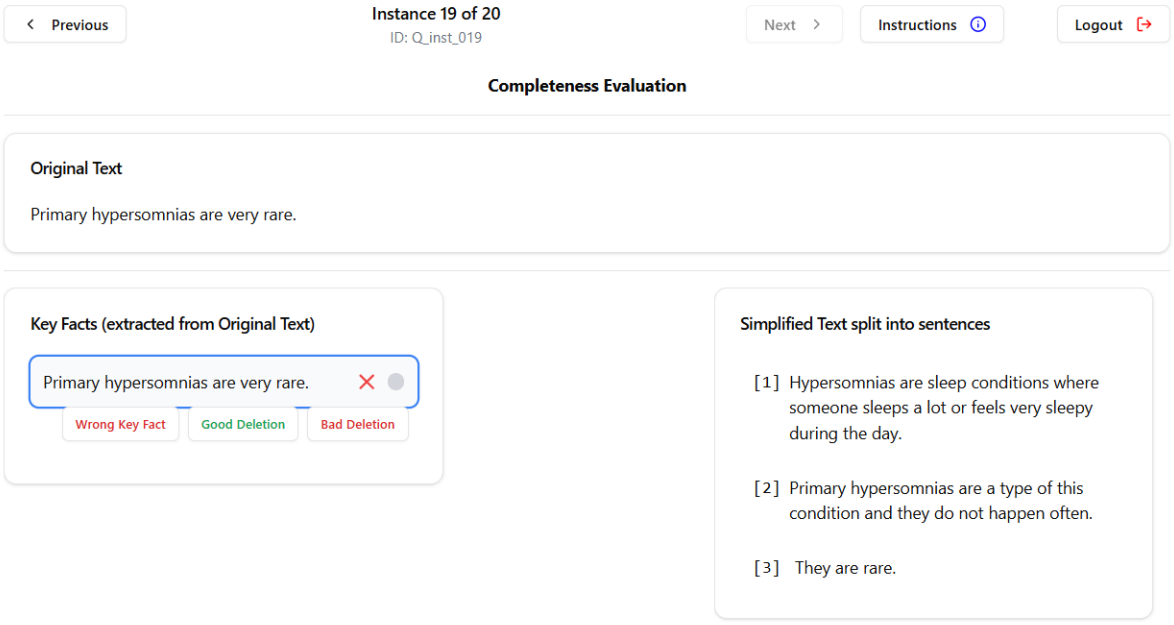


Figure 4: Completeness Evaluation Annotation Interface.

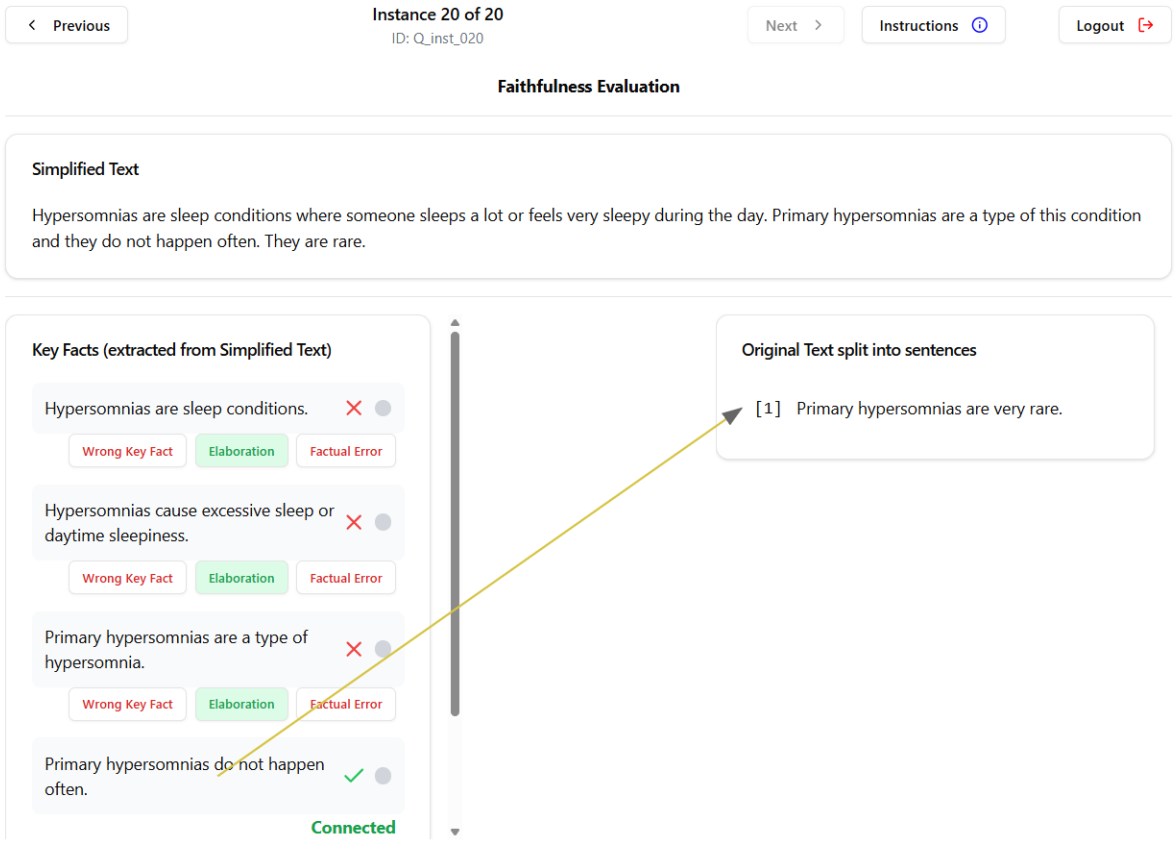


Figure 5: Faithfulness Evaluation Annotation Interface.

takes a document and a sentence as input and predicts whether the sentence is supported (1) or unsupported (0) by the document. In our experiments, we use the `flan-t5-large` variant and evaluate MiniCheck both sentence-by-sentence, by splitting each simplification into sentences and aggregating the resulting scores, and at the full-text level by scoring the entire simplification against the source.

Dataset	#Docs	Words ($\mu \pm \sigma$)	Sents ($\mu \pm \sigma$)	FK ($\mu \pm \sigma$)	CEFR (median; $\mu \pm \sigma$)
Sun et al. (2021)	100	149.15 ± 133.11	5.81 ± 4.91	11.58 ± 3.21	C1 (4.34 ± 1.09)
Cripwell et al. (2023)	198	48.59 ± 21.45	2.53 ± 0.79	9.74 ± 3.05	B2 (3.54 ± 0.96)
Cripwell et al. (2024)	250	80.23 ± 28.19	3.86 ± 0.92	10.81 ± 3.3	B2 (4.08 ± 0.97)
Agrawal and Carpuat (2024)	60	120.87 ± 30.2	5.42 ± 2.01	11.61 ± 2.68	B2 (4.23 ± 0.53)
M^{2}PRESERVEEVAL (Ours)	60	98.92 ± 19.09	4.52 ± 1.61	12.12 ± 3.5	B2 (4.02 ± 0.57)
ELG-CEFR-En	10	103.7 ± 15.61	6.4 ± 1.71	9.68 ± 1.12	B2 (3.9 ± 0.57)
OneStopEnglish	20	103.35 ± 19.52	3.4 ± 1.05	14.72 ± 4.2	B2 (4.15 ± 0.49)
British Council	20	93.05 ± 19.83	4.6 ± 1.27	11.12 ± 2.51	B2 (3.8 ± 0.62)
Cambridge-Exams	10	97.0 ± 18.99	4.7 ± 1.25	11.33 ± 1.96	B2 (4.3 ± 0.48)

Table 6: Statistics of source texts in **M^{2}PRESERVEEVAL** and comparison datasets.

Prompt

Please rewrite the following complex text in order to make it easier to understand by non-native speakers of English. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified text needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning.

Complex Text: {*Complex Text*}
Simplified Text:

Table 7: Prompt provided to LLMs to generate simplifications.

Setup	Atomicity	Sufficiency	Fabrication	Coverage	Redundancy	Avg. Atomic Facts
Human	2.34	2.58	1.07	2.97	1.95	10.43
GPT-4o + FactScore	1.99	2.80	1.05	3.00	1.69	7.61
Gemini + FactScore	2.27	2.74	1.05	3.00	1.83	9.28
GPT-4o + R-ND	2.15	2.78	1.05	2.99	1.78	9.66
Gemini + R-ND	2.37	2.62	1.06	3.00	1.94	12.44

Table 8: Key-fact decomposition results on REALSumm using the FactLens metrics, including the human reference decomposition.

Setup	Atomicity	Sufficiency	Fabrication	Coverage	Redundancy	Avg. Atomic Facts
Human	2.39	2.53	1.11	2.99	2.04	11.47
GPT-4o + FactScore	2.06	2.76	1.07	2.99	1.77	8.46
Gemini + FactScore	2.27	2.66	1.07	2.99	1.91	10.32
GPT-4o + R-ND	2.20	2.70	1.07	2.99	1.86	10.31
Gemini + R-ND	2.39	2.56	1.08	2.99	1.95	12.81

Table 9: Key-fact decomposition results on ROSE using the FactLens metrics, including the human reference decomposition.

We report both a probability-based variant, which averages the model's raw support scores, and a thresholded variant, which aggregates binary support decisions.

HHEM-2.1-Open (Bao et al., 2024) is a hallucination detection metric based on a Flan-T5-Base backbone. It takes a pair (premise, hypothesis) as input and outputs a score between 0 and 1 indicating the degree to which the hypothesis is supported by the premise. In our experiments, we evaluate HHEM-2.1-Open both at the sentence level, by scoring each simplified sentence against the source and aggregating the resulting scores, and at the full-text level by scoring the entire simplification against the source. We report both a probability-based variant, which averages the model's raw support scores, and a thresholded variant, which con-

verts scores into binary support decisions before aggregation.

PlainQAFact (You and Guo, 2026) is a retrieval-augmented and question-answering-based factuality assessment framework for evaluating the factuality of biomedical plain language summarization. It is designed to handle elaborative explanations by first distinguishing simplification and explanation sentences, then applying tailored scoring with external knowledge retrieval. In our experiments, we use the official PlainQAFact implementation with the learned classifier, `Llama-3.1-8B-Instruct` for answer extraction, `uzw/bart-large-question-generation` for question generation, the original QA answering model from QAFactEval (Fabbri et al., 2022) and the `combined` knowledge base.[11]

Agg-Metric (Maddela and Alva-Manchego, 2025) is a document-level simplification evaluation framework that adapts sentence-level metrics by computing scores at the sentence level and then aggregating them across all sentences in a document to yield an overall score. In our experiments, we applied Agg-Metric over the sentence-level metrics BERTScore and MeaningBERT.

G-Eval (Liu et al., 2023a) is an LLM-based evaluation framework that guides a model through reasoning steps via chain-of-thought prompting and produces a rating on a 1–5 Likert scale using a form-filling paradigm. In our experiments, we follow the G-Eval protocol and adapt the prompts from Lee et al. (2024) and Liu et al. (2023a) to the text simplification task to evaluate completeness and faithfulness, respectively. The prompts for completeness and faithfulness are provided in Table 14 and Table 15, respectively.

M^{2}PRESERVE-LLM adapts our human annotation framework into an LLM-based evaluation procedure. For each evaluation instance, the LLM is given the original text, the simplified text, and the relevant extracted key facts. For *Completeness*, it aligns and labels key facts extracted from the original text; for *Faithfulness*, it aligns and labels key facts extracted from the simplified text, using the annotation definitions in Section 3. The prompts specify the task, label definitions, alignment criteria, and required output format. The *Completeness* and *Faithfulness* prompts are provided in Tables 13 and 12, respectively.

LLMs as Evaluators We evaluate a diverse set of proprietary and open-weight LLMs as automatic judges under both G-Eval and M^{2}PRESERVE-LLM, including `Qwen3-4B-Instruct` (Qwen Team, 2025), `Qwen3-4B-Thinking` (Qwen Team, 2025), `Phi-4` (Abdin et al., 2024), `gpt-4.1`, `gpt-5.4` (OpenAI, 2025), `gemma-3-4b-it` (Gemma Team, 2025), `gemini-3.1-pro`, and `claude-sonnet-4.6`. Table 10 lists the model checkpoints.

C.2 Results

Table 11 demonstrates the full results on M^{2}PRESERVEEVAL for *Completeness* and *Faithfulness*.

D Model Checkpoints

Table 10 lists the model checkpoints used for simplification generation, key-fact extraction, and LLM-as-evaluator experiments. Open-weight models were downloaded from Hugging Face and run locally on a single NVIDIA GeForce RTX 4090 GPU with 24GB memory, while proprietary models were accessed through the OpenAI or OpenRouter APIs.

Model	Checkpoint
GPT-4o$^{\triangle}$	`gpt-4o-2024-11-20`
Gemini 2.0 Flash*	`google/gemini-2.0-flash-001`
Gemma 3-12B$^{\diamond}$	`google/gemma-3-12b-it`
Qwen 2.5-14B$^{\diamond}$	`Qwen/Qwen2.5-14B-Instruct`
BART-Swipe$^{\diamond}$	`Salesforce/bart-large-swipe-clean`
Qwen3-4B-Instruct$^{\diamond}$	`Qwen/Qwen3-4B-Instruct-2507`
Qwen3-4B-Thinking$^{\diamond}$	`Qwen/Qwen3-4B-Thinking-2507`
Gemma 3-4B$^{\diamond}$	`google/gemma-3-4b-it`
Phi-4$^{\diamond}$	`microsoft/phi-4`
GPT-4.1$^{\triangle}$	`gpt-4.1-2025-04-14`
GPT-5.4$^{\triangle}$	`gpt-5.4-2026-03-05`
Gemini 3.1 Pro*	`google/gemini-3.1-pro-preview`
Claude Sonnet 4.6*	`anthropic/claude-sonnet-4.6`

Table 10: Checkpoints for models used in simplification generation, key-fact extraction, and LLM-as-evaluator experiments. $^{\triangle}$ = OpenAI API, $^{\diamond}$ = Hugging Face, and * = OpenRouter API.

E Use of LLMs

In producing this work, we used Grammarly for minor grammar and spelling corrections and ChatGPT for LaTeX formatting and code/debugging support. All tool-generated suggestions were carefully reviewed and verified by the authors before incorporation into the paper.

[11] https://github.com/zhiwenyou103/PlainQAFact

	COMPLETENESS				FAITHFULNESS			
	PDP	acc^*_{eq}	SPA	bAcc	PDP	acc^*_{eq}	SPA	bAcc
BLEU	-0.06	0.46	0.52	-	0.07	0.90	0.79	-
BERTScore	0.03	0.46	0.49	-	0.08	0.90	0.76	-
MeaningBERT	-0.02	0.47	0.51	-	0.17	0.90	0.79	-
LogProp	-0.04	0.49	0.44	-	-0.05	0.90	0.67	-
AlignScore-base	-0.13	0.46	0.47	-	0.08	0.90	0.77	-
AlignScore-large	-0.13	0.46	0.50	-	0.09	0.91	0.72	-
HHEM-2.1- by sent.- prob	-0.38	0.46	0.39	-	0.12	0.90	0.74	-
HHEM-2.1- by sent.	-0.23	0.46	0.42	-	0.13	0.90	0.77	-
HHEM-2.1- prob	-0.27	0.46	0.37	-	0.11	0.90	0.72	-
HHEM-2.1	-0.07	0.46	0.43	-	0.08	0.90	0.76	-
MiniCheck- by sent.- prob	-0.12	0.46	0.43	-	0.04	0.90	0.81	-
MiniCheck- by sent.	-0.11	0.46	0.46	-	0.06	0.90	0.79	-
MiniCheck- prob	-0.17	0.47	0.36	-	0.16	0.91	0.80	-
MiniCheck	-0.14	0.46	0.36	-	0.04	0.90	0.84	-
PlainQAFact	-0.17	0.47	0.48	-	0.07	0.90	0.76	-
Agg - BERTScore	0.21	0.48	0.59	-	0.03	0.90	0.64	-
Agg - MeaningBERT	-0.07	0.46	0.46	-	0.09	0.90	0.74	-
G-Eval								
Qwen3-4B-Instruct	0.18	0.53	0.89	-	0.06	0.90	0.77	-
Qwen3-4B-Thinking	0.66	0.60	0.67	-	-0.04	0.90	0.54	-
gemma-3-4b-it	0.57	0.59	0.98	-	-0.05	0.90	0.53	-
Phi-4	-0.14	0.46	0.34	-	0.12	0.90	0.68	-
gpt-4.1	0.95	0.78	0.90	-	0.47	0.92	0.55	-
gpt-5.4	0.93	0.76	0.89	-	0.40	0.92	0.63	-
gemini-3.1-pro	0.93	0.78	0.76	-	0.37	0.92	0.46	-
claude-sonnet-4.6	0.88	0.74	0.93	-	-0.13	0.91	0.52	-
M^{2}PRESERVE								
Qwen3-4B-Instruct	0.89	0.74	0.88	0.89	0.14	0.90	0.71	0.67
Qwen3-4B-Thinking	0.65	0.63	0.93	0.92	0.29	0.90	0.73	0.78
gemma-3-4b-it	0.26	0.51	0.83	-	0.12	0.91	0.57	-
Phi-4	0.54	0.60	0.93	-	0.16	0.90	0.68	-
gpt-4.1	0.82	0.72	0.92	0.95	0.05	0.90	0.79	0.83
gpt-5.4	0.60	0.66	0.76	0.94	0.38	0.90	0.86	0.81
gemini-3.1-pro	0.66	0.63	0.77	0.95	0.71	0.93	0.79	0.80
claude-sonnet-4.6	0.82	0.73	0.91	0.94	0.08	0.90	0.69	0.77
LLM-as-a-jury	0.93	0.77	0.93	0.95	0.52	0.90	0.80	0.72

Table 11: Performance of automatic evaluation metrics on M^{2}PRESERVEEVAL. Metrics are evaluated on *Completeness* and *Faithfulness*. We report Pairwise Difference Pearson (PDP), segment-level accuracy with tie calibration (acc^*_{eq}), system-level Soft Pairwise Accuracy (SPA), and key fact-level Balanced Accuracy (bAcc); higher is better. *prob* denotes using the model's predicted probability as the score, while *by sent.* indicates that scores are computed for each sentence and then aggregated to obtain the final score.

Prompt

Faithfulness Evaluation.

You will be provided with:
. A simplified text.
. A list of key facts extracted from the simplified text.
. The original text, split into sentences.

Your task is to determine whether each key fact from the simplified text is accurately supported by the original text.

Steps:
1. Check if the key fact is accurately represented in the original text (either explicitly or implicitly, even if paraphrased):
. If accurately represented, connect the key fact to the corresponding sentence(s) in the original text.
. If the fact is NOT accurately represented, classify it as either:
. Wrong Key Fact: The key fact was NOT correctly extracted from the simplified text.
. factual error: The key fact introduces misleading, incorrect, or fabricated information does not present in the original text.
. elaboration: The key fact provides additional details (e.g., examples, background information, or definitions) that are accurate but not explicitly mentioned in the original text.

Provide your answer in JSON format. The answer should be a list of dictionaries whose keys are "key fact", "response", "Operation", and "line number":

["key fact": "first key fact", "response": "No", "Operation": "Wrong Key Fact", "line number": [], "key fact": "second key fact", "response": "Yes", "Operation": "Connected", "line number": [1,2], "key fact": "third key fact", "response": "No", "Operation": "factual error", "line number": [], "key fact": "fourth key fact", "response": "No", "Operation": "elaboration", "line number": []]

Simplified Text:

{*Simplified Text*}

Key Facts (extracted from Simplified Text):

{*Key Fact*}

{*Key Fact*}

Original Text split into sentences:

{*[1] sentence*}

{*[2] sentence*}

Table 12: M^{2}PRESERVE-LLM Prompt for Faithfulness Evaluation.

Prompt

Completeness Evaluation.

You will be provided with:
. The original text.
. A list of key facts extracted from the original text.
. The simplified text, split into sentences.

Your task is to assess whether each key fact from the original text can be clearly inferred from the simplified text.

Steps:
1. Check if the key fact can be clearly inferred from the simplified text (either explicitly or implicitly, even if paraphrased):
. If the key fact can be clearly inferred, connect it to the corresponding sentence(s) in the simplified text.
. If the key fact cannot be clearly inferred (i.e. it is missing or misrepresented), classify it as either:
. Wrong Key Fact: The key fact was NOT correctly extracted from the original text.
. Bad Deletion: The key fact is omitted, and this information is critical; its absence does obfuscate the overall main idea of the original text.
. Good Deletion: The key fact is omitted, but the omitted information is not essential; its absence does not obfuscate the overall main idea of the original text.

Provide your answer in JSON format. The answer should be a list of dictionaries whose keys are "key fact", "response", "Operation", and "line number":

["key fact": "first key fact", "response": "No", "Operation": "Wrong Key Fact", "line number": [], "key fact": "second key fact", "response": "Yes", "Operation": "Connected", "line number": [1,2], "key fact": "third key fact", "response": "No", "Operation": "Bad Deletion", "line number": [], "key fact": "fourth key fact", "response": "No", "Operation": "Good Deletion", "line number": []]

Original Text:

{*Original Text*}

Key Facts (extracted from Original Text):

{*Key Fact*}

{*Key Fact*}

Simplified Text split into sentences:

{*[1] sentence*}

{*[2] sentence*}

Table 13: M^{2}PRESERVE-LLM Prompt for Completeness Evaluation.

Prompt

You will be given a source text and its simplified version.
Your task is to rate the simplified text on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Completeness (1-5) - the degree to which the simplification includes all key information present in the source document. A complete simplification accurately captures the main points, ideas, and relevant details without omitting crucial elements.

Evaluation Steps:

1. Read the article carefully and identify the main points, key information, and relevant details.
2. Read the simplification and compare it to the article. Check if the simplification captures all essential facts, main ideas, and pertinent details presented in the original article.
3. Assign a score from 1 to 5 for completeness based on the Evaluation Criteria.

Source Text:
{*Complex Text*}

Simplified Text:
{*Simplified Text*}

Evaluation Form (scores ONLY):
- Completeness:

Table 14: G-Eval Prompt for Completeness Evaluation.

Prompt

You will be given a source text and its simplified version.
Your task is to rate the simplified text on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Consistency (1-5) - the factual alignment between the simplification and the source. A factually consistent simplification contains only statements that are entailed by the source document. Additionally, annotators did not penalize benign elaborations (e.g., examples, background information, or definitions) that were accurate but not explicitly mentioned in the original text, provided that they were factually correct, did not contradict the source, and did not introduce new specific claims about the entities or events described in the source. Annotators were asked to penalize simplifications that contained hallucinated facts (i.e., unsupported or contradictory claims).

Evaluation Steps:

1. Read the Source text carefully and identify the main facts and details it presents.
2. Read the simplification and compare it to the Source text. Check if the simplification contains any factual errors that are not supported by the Source text.
3. Assign a score for consistency based on the Evaluation Criteria.

Source Text:
{*Complex Text*}

Simplified Text:
{*Simplified Text*}

Evaluation Form (scores ONLY):
- Consistency:

Table 15: G-Eval Prompt for Faithfulness Evaluation.

R-ND Prompt, Part 1

Please decompose the following text into individual facts: He made his acting debut in the film The Moon is the Sun's Dream (1992), and continued to appear in small and supporting roles throughout the 1990s.
- He has an acting debut.
- He acted in a film.
- His acting debut was in a film.
- His acting debut was in The Moon is the Sun's Dream.
- He acted in The Moon is the Sun's Dream.
- The Moon is the Sun's Dream is a film.
- The Moon is the Sun's Dream was released in 1992.
- His acting debut occurred in 1992.
- He appeared in small roles.
- He appeared in supporting roles.
- His small roles occurred throughout the 1990s.
- His supporting roles occurred throughout the 1990s.
- His appearance in small roles occurred after his acting debut.
- His appearance in supporting roles occurred after his acting debut.

Please decompose the following text into individual facts: He is also a successful producer and engineer, having worked with a wide variety of artists, including Willie Nelson, Tim McGraw, and Taylor Swift.
- He is a producer.
- He is successful at being a producer.
- He is an engineer.
- He is successful at being an engineer.
- He has worked with a wide variety of artists.
- Willie Nelson is an artist.
- He has worked with Willie Nelson.
- Tim McGraw is an artist.
- He has worked with Tim McGraw.
- Taylor Swift is an artist.
- He has worked with Taylor Swift.

Please decompose the following text into individual facts: In 1963, Collins became one of the third group of astronauts selected by NASA and he served as the back-up Command Module Pilot for the Gemini 7 mission.
- NASA selected a third group of astronauts.
- Collins belonged to the third group of astronauts.
- Collins was selected by NASA.
- Collins's selection by NASA occurred in 1963.
- The Gemini 7 mission has a back-up Command Module Pilot.
- Collins's role in the Gemini 7 mission was as the back-up Command Module Pilot.
- Collins participated in the Gemini 7 mission.

Please decompose the following text into individual facts: In addition to his acting roles, Bateman has written and directed two short films and is currently in development on his feature debut.
- Bateman has acting roles.
- Bateman has written short films.
- The number of short films Bateman has written is two.
- Bateman has directed short films.
- The number of short films Bateman has directed is two.
- Bateman is currently in development on his feature debut.
- The two short films were made before his feature debut.
- His acting roles came before his feature debut.

Table 16: R-ND Prompt provided to LLMs for key-fact extraction, part 1 of 2.

R-ND Prompt, Part 2

Please decompose the following text into individual facts: Michael Collins (born October 31, 1930) is a retired American astronaut and test pilot who was the Command Module Pilot for the Apollo 11 mission in 1969.
- Michael Collins was born in October.
- Michael Collins was born on the 31st day of a month.
- Michael Collins was born in 1930.
- Michael Collins is retired.
- Michael Collins is American.
- Michael Collins was an astronaut.
- Michael Collins was a test pilot.
- Michael Collins participated in the Apollo 11 mission.
- Michael Collins's participation in the Apollo 11 mission occurred in 1969.
- The Apollo 11 mission was active in 1969.
- The day of Michael Collins's birth occurred before his year of participation in the Apollo 11 mission.
- The Apollo 11 mission had a Command Module Pilot.
- Michael Collins's role in the Apollo 11 mission was as the Command Module Pilot.

Please decompose the following text into individual facts: He was an American composer, conductor, and musical director.

- He was American.
- He was a composer.
- He was a conductor.
- He was a musical director

Please decompose the following text into individual facts: She currently stars in the romantic comedy series, Love and Destiny, which premiered in 2019.

- She stars in Love and Destiny.
- Love and Destiny is a series.
- Love and Destiny is a romantic comedy.
- Love and Destiny premiered in 2019.

Please decompose the following text into individual facts: His music has been described as a mix of traditional Mexican and Latin American styles, as well as jazz, folk, and rock.
- He has music.
- His music has been described.
- His music has been described as a mix of styles.
- His music has been described as containing elements of traditional styles of music.
- His music has been described as containing elements of Mexican style of music.
- His music has been described as containing elements of Latin American style of music.
- His music has been described as containing elements of jazz music.
- His music has been described as containing elements of folk music.
- His music has been described as containing elements of rock music.

Please decompose the following text into individual facts: He also serves as an ambassador for the charity Leonard Cheshire Disability.
- He has a role in Leonard Cheshire Disability.
- His role in Leonard Cheshire Disability is as an ambassador.
- Leonard Cheshire Disability is a charity.

Please decompose the following text into individual facts: {*input*}

Table 17: R-ND Prompt provided to LLMs for key-fact extraction, part 2 of 2.

FACTSCORE Prompt, Part 1

Please breakdown the following text into independent facts: He made his acting debut in the film The Moon is the Sun's Dream (1992), and continued to appear in small and supporting roles throughout the 1990s.
- He made his acting debut in the film.
- He made his acting debut in The Moon is the Sun's Dream.
- The Moon is the Sun's Dream is a film.
- The Moon is the Sun's Dream was released in 1992.
- After his acting debut, he appeared in small and supporting roles.
- After his acting debut, he appeared in small and supporting roles throughout the 1990s.

Please breakdown the following text into independent facts: He is also a successful producer and engineer, having worked with a wide variety of artists, including Willie Nelson, Tim McGraw, and Taylor Swift.
- He is successful.
- He is a producer.
- He is a engineer.
- He has worked with a wide variety of artists.
- Willie Nelson is an artist.
- He has worked with Willie Nelson.
- Tim McGraw is an artist.
- He has worked with Tim McGraw.
- Taylor Swift is an artist.
- He has worked with Taylor Swift.

Please breakdown the following text into independent facts: In 1963, Collins became one of the third group of astronauts selected by NASA and he served as the back-up Command Module Pilot for the Gemini 7 mission.
- Collins became an astronaut.
- Collins became one of the third group of astronauts.
- Collins became one of the third group of astronauts selected.
- Collins became one of the third group of astronauts selected by NASA.
- Collins became one of the third group of astronauts selected by NASA in 1963.
- He served as the Command Module Pilot.
- He served as the back-up Command Module Pilot.
- He served as the Command Module Pilot for the Gemini 7 mission.

Please breakdown the following text into independent facts: In addition to his acting roles, Bateman has written and directed two short films and is currently in development on his feature debut.
- Bateman has acting roles.
- Bateman has written two short films.
- Bateman has directed two short films.
- Bateman has written and directed two short films.
- Bateman is currently in development on his feature debut.

Table 18: FACTSCORE Prompt provided to LLMs for key-fact extraction, part 1 of 2.

FACTSCORE Prompt, Part 2

Please breakdown the following text into independent facts: Michael Collins (born October 31, 1930) is a retired American astronaut and test pilot who was the Command Module Pilot for the Apollo 11 mission in 1969.
- Michael Collins was born on October 31, 1930.
- Michael Collins is retired.
- Michael Collins is an American.
- Michael Collins was an astronaut.
- Michael Collins was a test pilot.
- Michael Collins was the Command Module Pilot.
- Michael Collins was the Command Module Pilot for the Apollo 11 mission.
- Michael Collins was the Command Module Pilot for the Apollo 11 mission in 1969.

Please breakdown the following text into independent facts: He was an American composer, conductor, and musical director.
- He was an American.
- He was a composer.
- He was a conductor.
- He was a musical director.

Please breakdown the following text into independent facts: She currently stars in the romantic comedy series, Love and Destiny, which premiered in 2019.
- She currently stars in Love and Destiny.
- Love and Destiny is a romantic comedy series.
- Love and Destiny premiered in 2019.

Please breakdown the following text into independent facts: During his professional career, McCoy played for the Broncos, the San Diego Chargers, the Minnesota Vikings, and the Jacksonville Jaguars.

- McCoy played for the Broncos.
- McCoy played for the Broncos during his professional career.
- McCoy played for the San Diego Chargers.
- McCoy played for the San Diego Chargers during his professional career.
- McCoy played for the Minnesota Vikings.
- McCoy played for the Minnesota Vikings during his professional career.
- McCoy played for the Jacksonville Jaguars.
- McCoy played for the Jacksonville Jaguars during his professional career.

Please breakdown the following text into independent facts: {*input*}

Table 19: FACTSCORE Prompt provided to LLMs for key-fact extraction, part 2 of 2.