Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# CMED: A Contrastive Diagnostic of Meaning-Preservation Metrics for Text Simplification

**Emin Talip Demirkiran**

etd@eskisehir.edu.tr

Eskisehir Technical University

**Research Article**

**Keywords:** text simplification, meaning preservation, evaluation metrics, contrast sets, natural language inference, large language models

**Posted Date:** November 13th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11219172/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# CMED: A Contrastive Diagnostic of Meaning-Preservation Metrics for Text Simplification

Emin Talip Demirkiran[1*]

[1*]Department of Computer Engineering, Eskisehir Technical University, Iki Eylul Campus, Eskisehir, 26555, Türkiye.

Corresponding author(s). E-mail(s): etd@eskisehir.edu.tr;

**Abstract**

Automatic metrics are routinely used to judge whether text simplifications, including those produced by large language models, preserve the meaning of the original, yet it is unclear which kinds of meaning errors these metrics detect. Because simplification legitimately deletes, paraphrases and splits content, a trustworthy metric must penalize meaning-altering edits more than such legitimate changes. We introduce CMED (Contrastive Meaning-Error Diagnostic), a resource of controlled contrast sets derived from the ASSET test set. It pairs 718 human simplifications with rule-generated meaning-altering edits of seven types (negation, number, entity, antonym, quantifier, omission and insertion) and with legitimate changes (alternative human simplifications, deletion of peripheral content and sentence splitting). Following an analysis plan fixed before the test data were generated, we measured, for nine widely used or recently proposed metrics, how often a legitimate change is scored above a meaning-altering edit of the same simplification. The metrics showed complementary blind spots. Factual-consistency metrics based on natural language inference detected substitutions and insertions almost perfectly (0.94–0.99) but were at or below chance on omissions (0.41–0.51). BERTScore was near chance on number changes, and MeaningBERT also on negation and antonyms. Only a bidirectional entailment score exceeded 0.70 on every error type, yet it separated omissions of main assertions from legitimate peripheral deletions only weakly (0.64). The edits are synthetic, English and sentence-level, and their validity was checked automatically rather than by human annotators. CMED and its code are released to support error-type-resolved evaluation of simplification metrics.

**Keywords:** text simplification, meaning preservation, evaluation metrics, contrast sets, natural language inference, large language models

# 1 Introduction

Text simplification rewrites a text so that it can be understood by a wider range of readers while retaining its meaning (Alva-Manchego et al., 2020a; Xu, Napoles, Pavlick, Chen, & Callison-Burch, 2016). Large language models (LLMs) now produce fluent simplifications that obtain high scores on standard automatic metrics (Feng, Qiang, Li, Yuan, & Zhu, 2023; Kew et al., 2023; Qiang et al., 2025). Whether such output can be trusted, however, depends on a property that the most widely reported metrics do not isolate: whether the simplified text still says what the original said. Expert evaluations of LLM-simplified texts in high-stakes domains report incorrect passages, missing information and potentially harmful content (Jeblick et al., 2024; Joseph et al., 2024), information loss in LLM simplification is dominated by oversimplification (Trienes et al., 2024), and in a human reading-comprehension evaluation, texts produced by simplification systems, including ChatGPT, left at least 14% of questions about the original unanswerable, compared with 12–14% for human-written texts (Agrawal & Carpuat, 2024).

Measuring meaning preservation in simplification is difficult because a simplification must change the text. Legitimate simplification deletes peripheral content, paraphrases and splits sentences (Alva-Manchego et al., 2020a). A metric that is to be trusted for meaning preservation must therefore penalize a meaning-altering edit, such as a flipped negation, a changed number, a swapped entity, a dropped main assertion or an inserted unsupported claim, *more* than it penalizes a legitimate change of the same text. Reference-based metrics are known to fail in characteristic ways: BLEU largely rewards conservative, copy-like output (Sulem, Abend, & Rappoport, 2018a), and SARI correlates weakly or even negatively with human meaning judgments (Korobeynikova, Battisti, Fischer, & Gao, 2026; Scialom, Martin, Staiano, Villemonte de la Clergerie, & Sagot, 2021; Sulem et al., 2018a; Sulem, Abend, & Rappoport, 2018b). Embedding-based and learned metrics correlate better with averaged human ratings (Beauchemin, Saggion, & Khoury, 2023; Maddela, Dou, Heineman, & Xu, 2023; Zhao, Durmus, & Yeung, 2023), yet studies that annotate individual errors in system outputs report that metric scores barely distinguish outputs with and without meaning errors (Vendeville, Ermakova, & De Loor, 2025; Wu & Arase, 2026) or correlate only weakly with error annotations (Heineman, Dou, Maddela, & Xu, 2023), and the entailment components of a hallucination detector for simplification performed close to chance (Marturi & Elwazzan, 2025).

This article argues that the question of *which* meaning errors a metric detects calls for a controlled diagnostic resource rather than another correlation study. We introduce CMED (Contrastive Meaning-Error Diagnostic), a set of contrast sets built from the human simplifications of ASSET (Alva-Manchego et al., 2020a). Following the contrast-set paradigm of Gardner et al. (2020), each human simplification is minimally perturbed; unlike manually authored contrast sets, our perturbations are generated by rules, typed by error class and contrasted with legitimate changes of the same simplification. We then ask, for each error type, how often a metric ranks the legitimate change above the meaning error. The design adapts the preference principle of MENLI (Chen & Eger, 2023) and typed perturbations closely related to those of MENLI and APPLS (Guo, August, Leroy, Cohen, & Wang, 2024) to the specific problem

of simplification, in which information may legitimately be removed. All metrics are computed on a standard CPU, without model training or text generation, reflecting the evaluation settings available to most practitioners.

The contributions are:

1. CMED, a released diagnostic resource of 4,969 scored items that crosses seven types of meaning-altering edits with three types of legitimate change, together with its generation code and automatic validity diagnostics (Section 3);
2. a per-error-type map of nine widely used or recently proposed meaning-preservation scores, tested against hypotheses and decision rules fixed before the test data were generated (Sections 4 and 5);
3. a test of whether adding source-direction entailment closes the omission blind spot of factual-consistency metrics without penalizing legitimate deletion (Section 5.3);
4. recommendations on which metrics to combine for which error types when evaluating simplification without large-model inference (Section 6.4).

## 2 Related work

***Reference-based metrics and the legitimacy confound.***

SARI was introduced to reward the addition, retention and deletion of words relative to multiple references and was validated on data restricted to paraphrase-style simplification (Xu et al., 2016). Subsequent studies showed that neither SARI nor BLEU tracks meaning when simplifications split or compress sentences: BLEU behaves mainly as a measure of conservatism, and its correlation with meaning judgments collapses once sentence-splitting systems are included (Sulem et al., 2018a), while SARI correlates weakly or negatively with meaning judgments (Alva-Manchego et al., 2020a; Sulem et al., 2018b). Reference-less overlap with the source predicts meaning ratings best but rewards copying (Martin et al., 2018). Meta-evaluations further showed that metric–human correlations depend on the quality range, the system type and the references used (Alva-Manchego, Scarton, & Specia, 2021b), and that correlations on system outputs are inflated by the strong correlation between human judgments of fluency and meaning (Scialom et al., 2021). A common thread is that meaning ratings are elicited as a single holistic score whose instructions allow the omission of unimportant information (Alva-Manchego et al., 2020a); acceptable deletion and information loss are therefore not separated in the evidence used to validate metrics.

***Learned and embedding-based meaning metrics.***

Several metrics have been proposed specifically for meaning preservation or overall simplification quality. MeaningBERT regresses human meaning-preservation ratings and was accompanied by two sanity checks, identical and unrelated sentence pairs (Beauchemin et al., 2023); LENS learns a holistic quality score from human ratings (Maddela et al., 2023); BETS defines meaning preservation as embedding recall of the input (Zhao et al., 2023). These metrics correlate markedly better with human ratings than n-gram metrics, but their validation remains correlational. Per-system analyses show that the agreement of LENS with adequacy judgments varies strongly between systems (Alfear, Kazakov, & Al-Khalifa, 2024), and a German study that

stratified results by simplification operation found all metrics weakest on paraphrase-type simplifications (Korobeynikova et al., 2026). None of these studies asks whether a metric penalizes a specific meaning error more than a specific legitimate change.

***Typed meaning errors and information loss.***

A second line of work annotates meaning errors in system outputs. Devaraj, Sheffield, Wallace, and Li (2022) distinguished insertion, deletion and substitution errors with severity levels and found that similarity metrics correlate with deletion but hardly with substitution; SALSA annotates edits and counts well-motivated deletion as a success rather than an error (Heineman et al., 2023); error-based evaluations of LLM simplifications (Wu & Arase, 2026) and a taxonomy of information distortion in scientific-text simplification (Vendeville et al., 2025) both report that current metrics barely detect the presence of errors. Information loss in LLM simplification has been characterized through question generation (Trienes et al., 2024) and through entity matching (Nandiraju, Leroy, Kauchak, & Ahmed, 2025), and meaning preservation has also been measured extrinsically through reading comprehension (Agrawal & Carpuat, 2024). These resources label naturally occurring errors, which makes them ecologically valid but leaves error types rare and unbalanced; inter-annotator agreement on such errors is limited, particularly for substitutions (Devaraj et al., 2022; Vendeville et al., 2025), and because the base text varies across items, a metric's failure cannot be attributed to a specific error type.

***Factual-consistency metrics.***

Metrics developed for the factual consistency of summaries score whether the output is supported by the source. SummaC aggregates sentence-level entailment (Laban, Schnabel, Bennett, & Hearst, 2022), and AlignScore scores claim-to-context alignment (Zha, Yang, Li, & Hu, 2023); by design, both evaluate the output against the source and do not measure what the output leaves out. When such metrics are transferred to plain-language summaries or simplifications, their agreement with expert judgments is moderate (Joseph et al., 2024), entailment-based hallucination detection is close to chance (Marturi & Elwazzan, 2025), and out of domain the metrics can rank systems opposite to human faithfulness judgments (Cripwell, Legrand, & Gardent, 2024). Cripwell et al. (2024) also computed source-direction (recall) variants of these metrics for document simplification, but these variants could not be compared with human judgments. Whether source-direction checking separates harmful omissions from the legitimate removal of peripheral content, the trade-off that every simplification makes, has not been tested.

***Perturbation tests of evaluation metrics.***

Controlled perturbations are an established way to probe metrics. Contrast sets perturb test instances minimally so that the gold label changes (Gardner et al., 2020). MENLI tests whether metrics for machine translation and summarization prefer a meaning-preserving paraphrase over a minimally corrupted candidate (Chen & Eger, 2023); APPLS applies graded perturbations to plain-language summaries and finds that number swaps go undetected by all metrics tested (Guo et al., 2024); a stress

**Table 1** CMED compared with related perturbation-based resources for testing evaluation metrics (as reported in the cited publications)

| | MENLI (Chen & Eger, 2023) | APPLS (Guo et al., 2024) | Stress test (Mujahid et al., 2026) | CMED (this work) |
|---|---|---|---|---|
| Task | MT, summarization | plain-language summarization | long-document summarization | sentence simplification |
| Perturbations | addition, omission, negation, number, name and others | graded; incl. number swaps | meaning-preserving only | 7 meaning-altering types |
| Legitimate counterpart | paraphrase | criterion-specific | (invariance test) | human rewrite, peripheral deletion, split |
| Omission | word dropping | information deletion | not tested | deletion of a main assertion |
| Factual-consistency metrics | NLI-based metrics | SummaC and AlignScore not tested | included | included |
| Inferential statistics | accuracy | none | descriptive | cluster bootstrap, Holm |

test of factual-consistency metrics uses meaning-preserving perturbations for long-document summarization (Mujahid, Wright, & Augenstein, 2026); and manipulations that degrade meaning were shown to leave SARI almost unchanged while making readability formulas arbitrarily favourable (Tanprasert & Kauchak, 2021). The edit types of CMED closely follow the perturbations of MENLI and APPLS. Table 1 summarizes how CMED differs from these resources: it contrasts typed meaning-altering edits with simplification-legitimate changes of the same human simplification, isolates the omission of main assertions, includes factual-consistency metrics, and reports inferential statistics. To our knowledge, within the literature covered by our searches (OpenAlex, arXiv and general web search, English-language, up to September 2026), no previous study combines these elements for text simplification.

# 3 The CMED resource

## 3.1 Design

Let $x$ be a complex source sentence and $r$ one of its human simplifications, which we call the *base*. CMED derives two kinds of variants from each base (Fig. 1). A *meaning-altering edit* changes $r$ minimally so that it no longer expresses the meaning of $x$. A *legitimate change* is a different simplification of $x$ that a careful human simplifier could produce. A metric that measures meaning preservation should score every legitimate change above every meaning-altering edit of the same base.

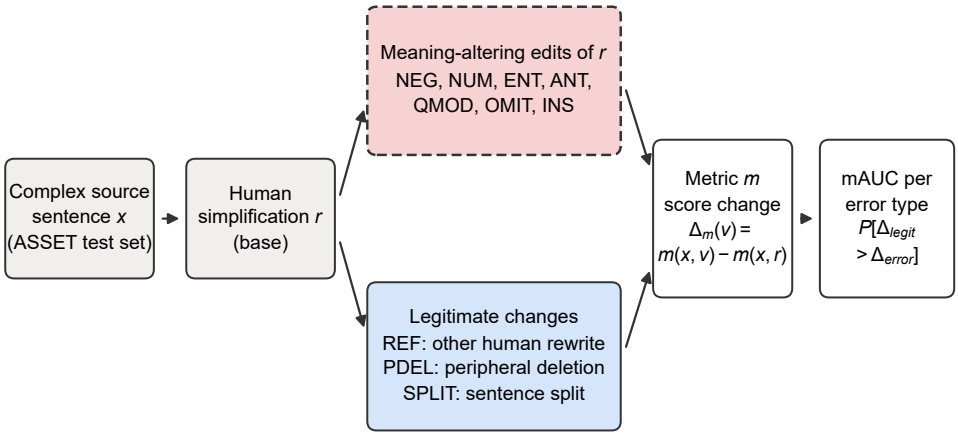



**Fig. 1** Design of CMED. Each human simplification (base) of an ASSET source sentence receives rule-generated meaning-altering edits and legitimate changes; a metric is scored by the probability that it assigns a smaller score decrease to a legitimate change than to a meaning-altering edit of the same base

## 3.2 Meaning-altering edits

Seven edit types were implemented as deterministic rules over dependency parses and named entities (spaCy 3.8.16 with `en_core_web_md`;Honnibal, Montani, Van Landeghem, & Boyd, 2020) and WordNet (Miller, 1995), operating on character offsets so that untouched text is left intact (Table 2):

- **NEG** inserts or removes sentential negation on the main verb.
- **NUM** changes a standalone numeral, either slightly (±5% or ±1; years ±1–3) or strongly (×2, 3, 5 or 10; years ±25–80).
- **ENT** replaces a person, place, organization or group name with a name of the same type from another source sentence.
- **ANT** replaces an adjective with a direct WordNet antonym of its most frequent sense.
- **QMOD** strengthens or reverses a quantifier or modal (e.g. *some*→*all*, *many*→*few*, *may*→*must*); logically entailed weakenings such as *all*→*some* are excluded because the edited sentence remains true.
- **OMIT** deletes a main assertion, either one of several sentences or an independent coordinated clause.
- **INS** appends an unrelated sentence taken from another source.

## 3.3 Legitimate changes

Three types of legitimate change were used. REF is another human simplification of the same source from ASSET (two per base). PDEL deletes a parenthetical, a short mid-sentence appositive or a non-restrictive relative clause, i.e. content that is

**Table 2** Examples of CMED variants. Rows NEG to REF are derived from one base in the test set (source: *Heavy rain fell across portions of Britain on October 5, causing localized accumulation of flood waters.*); PDEL and SPLIT come from other bases because they are rarely applicable

| Type | Text |
|---|---|
| Base | On October 5, there was heavy rain across parts of Britain. This caused flood waters to build up in some places. |
| NEG | On October 5, there was *not* heavy rain across parts of Britain. This caused ... |
| NUM | On October *15*, there was heavy rain across parts of Britain. This caused ... |
| ENT | On October 5, there was heavy rain across parts of *the United States*. This caused ... |
| ANT | On October 5, there was *light* rain across parts of Britain. This caused ... |
| QMOD | ... This caused flood waters to build up in *all* places. |
| OMIT | On October 5, there was heavy rain across parts of Britain. |
| INS | ... build up in some places. *He died there six weeks later in 888.* |
| REF | Heavy rain fell across parts of Britain on October 5. This caused localized floods. |
| PDEL | He was made the Companion of Honour (CH) in 1988. → He was made the Companion of Honour in 1988. |
| SPLIT | Their eyes are small, and their visual acuity is poor. → Their eyes are small. Their visual acuity is poor. |

not at issue and that the ASSET rating instructions allow simplifiers to drop (Alva-Manchego et al., 2020a). SPLIT splits a sentence at a comma followed by *and* and an independent clause. Lexical substitution by synonyms was deliberately not generated automatically, because automatic substitution risks real meaning change; paraphrase is represented by the human rewrites in REF.

## 3.4 Construction and composition

CMED is built from the ASSET test set (Alva-Manchego et al., 2020a, 2020b), which contains 359 English Wikipedia sentences, each with ten crowd-sourced simplifications. Simplifications 0 and 1 of each source serve as bases (718 bases), and simplifications $3 + b$ and $6 + b$ of base $b$ serve as REF variants. Reference-based metrics receive the remaining simplifications of the source as references, excluding any string identical to the base or to its REF variants (two to seven references per base). All edits are deterministic given a fixed seed, and edits that leave the text unchanged are discarded. Table 3 lists the resulting composition. Error types with fewer than 100 variants (QMOD) are reported descriptively only. For the analysis of omission versus peripheral deletion (Section 5.3), an extension set uses all ten simplifications of each source as bases and contains only OMIT and PDEL variants (839 and 242 variants).

The edit rules were developed and inspected only on the ASSET validation set. Two rules were revised on that set before the test data were generated: quantifier edits were restricted to non-entailed substitutions and antonym edits to the most frequent word sense, after an automatic validity check (Section 3.5) showed that entailed weakenings and rare senses often did not alter meaning.

## 3.5 Validity diagnostics

The label of every CMED item follows from its construction. As an additional diagnostic, not as ground truth, an NLI model from a different model family than the NLI scores under test (RoBERTa-large-MNLI) labelled a stratified 25% sample in

**Table 3** Composition of the CMED test resource and automatic validity diagnostics. The last column gives the proportion of a stratified 25% sample for which an independent NLI model (RoBERTa-large-MNLI) showed the pattern expected for the variant type (see Section 3.5)

| Class | Type | Variants | Checked | Expected pattern |
|---|---|---|---|---|
| Meaning-altering | NEG | 692 | 173 | 0.971 |
| | NUM | 364 | 91 | 0.802 |
| | ENT | 480 | 120 | 0.775 |
| | ANT | 260 | 65 | 0.862 |
| | QMOD[a] | 78 | 20 | 0.850 |
| | OMIT | 179 | 45 | 1.000 |
| | INS | 718 | 180 | 0.844 |
| Legitimate | REF | 1,432 | 358 | 0.975 |
| | PDEL | 42 | 10 | 1.000 |
| | SPLIT | 6 | 2 | 1.000 |

Bases: 718 from 359 sources; 4,969 items including bases. [a]Descriptive only (fewer than 100 variants).

both directions. The expected pattern was that the source does not entail substitution and insertion edits, that omission edits do not entail the source, and that the source entails legitimate changes. The pattern held for 78–100% of the meaning-altering and 98–100% of the legitimate variants (Table 3). The residual 15–22% for entity, number and insertion edits reflects either genuine label noise or blind spots of the checker; because the checker is itself an NLI model, the two cannot be separated, and no human validation was performed.

### 3.6 Format, licence and availability

Each CMED item is a JSON line recording the source sentence, the base, the variant text, its class, type and subtype, the references used by reference-based metrics, and the normalized token edit distance to the base. The resource, the per-item metric scores, the generation and analysis code and the analysis plan are archived on Zenodo (Demirkiran, 2026). Because CMED is derived from ASSET (CC BY-NC 4.0; Alva-Manchego et al., 2020b) and, for the external-validity items, from Simplicity-DA (CC BY-NC-SA 4.0; Alva-Manchego, Scarton, & Specia, 2021a), the data inherit these non-commercial licences; the code is released under the MIT License. Meaning-altering items are deliberately false statements, some of which mention real people and places, and are labelled as such.

## 4 Experimental setup

### 4.1 Metrics

We evaluated nine scores that are widely used or were recently proposed for meaning preservation or factual consistency, all computable on a CPU without training (Table 4). Because the full evaluation had to run within a fixed computational budget of about six hours on a 12-thread laptop processor, three metrics were used in

**Table 4** Metrics evaluated. Higher scores indicate better meaning preservation for all metrics

| Family | Metric | Configuration |
|---|---|---|
| Reference-based | BLEU (Papineni et al., 2002) | sacreBLEU (Post, 2018) sentence-level score against the references |
| | SARI (Xu et al., 2016) | re-implementation checked against a published reference value |
| Embedding / learned | BERTScore (Zhang et al., 2019) | F1 between source and output (RoBERTa-large, layer 17) |
| | MeaningBERT (Beauchemin et al., 2023) | released checkpoint; source and output |
| | LENS (Maddela et al., 2023) | released checkpoint; three references; rescaled |
| Factual consistency | SummaC-ZS (Laban et al., 2022) | sentence granularity; `vitc-base` backbone |
| | AlignScore (Zha et al., 2023) | base checkpoint; `nli_sp` mode; source as context |
| | NLI-fwd | $P(\text{entailment} \mid x \Rightarrow v)$, DeBERTa-large (He et al., 2021) fine-tuned on MNLI (Williams et al., 2018) |
| | BiNLI | $\frac{1}{2}[P(x \Rightarrow v) + P(v \Rightarrow x)]$, same model; bi-implication as in MENLI (Chen & Eger, 2023) |
| Exploratory | BiNLI-core | NLI-fwd averaged with the minimum entailment of the source's main-clause propositions by $v$ |

smaller configurations fixed before the test data were generated: SummaC-ZS with the `vitc-base` backbone (trained on VitaminC; Schuster, Fisch, & Barzilay, 2021) listed in the model map of the SummaC software instead of the larger default, AlignScore with its base checkpoint, and LENS with three references. Findings for these metrics therefore apply to these configurations. All model checkpoints were pinned to fixed revisions.

NLI-fwd and BiNLI are simple probes rather than separately published metrics. BiNLI follows the bi-implication variant of MENLI (Chen & Eger, 2023), which averages the entailment probabilities of the two directions and uses the same DeBERTa-large MNLI model. Both probes use the same entailment model, so that their comparison isolates the effect of adding the source direction. BiNLI-core restricts the source direction to main-clause propositions; because it relies on the same notion of a main clause that defines OMIT and PDEL, its performance on these types is partly built in, and it is reported as exploratory only.

## 4.2 Outcome measures

For a metric $m$ and a variant $v$ derived from base $r$ of source $x$, the score change is

$$\Delta_m(v) = m(x, v) - m(x, r). \tag{1}$$

For REF variants, $v$ is the alternative human simplification. For each meaning-altering type $t$, let $\mathcal{P}_t$ be the set of all pairs $(u, e)$ in which $u$ is a legitimate change and $e$ a type-$t$ edit of the same base. The primary outcome is the matched area under the ROC curve,

$$\mathrm{mAUC}_t(m) = \frac{1}{|\mathcal{P}_t|} \sum_{(u,e)\in\mathcal{P}_t} \Big( \mathbb{1}\big[\Delta_m(u) > \Delta_m(e)\big] + \tfrac{1}{2}\mathbb{1}\big[\Delta_m(u) = \Delta_m(e)\big] \Big), \tag{2}$$

the probability that the metric penalizes the meaning-altering edit more than a legitimate change of the same base. A value of 0.5 corresponds to chance, and values below 0.5 indicate that the metric prefers the erroneous text. Because meaning-altering edits are minimal (median normalized edit distance 0.06 on the validation set) while legitimate rewrites are larger (median 0.56), this contrast asks whether a metric prefers a realistic legitimate rewrite over a nearly identical but wrong text, as in MENLI (Chen & Eger, 2023). As secondary, exploratory outcomes we report a magnitude-matched version restricted to pairs whose edit distances differ by at most 0.10, and the proportion of edits that lower the score, $P[\Delta_m(e) < 0]$.

For omission, the key question is whether a metric can distinguish harmful omission from legitimate deletion. On the extension set, we therefore compute

$$\mathrm{AUC}_{\mathrm{OP}}(m) = P\big[\Delta_m(u) > \Delta_m(e)\big] + \tfrac{1}{2}\,P\big[\Delta_m(u) = \Delta_m(e)\big], \quad u \in \mathrm{PDEL},\ e \in \mathrm{OMIT}, \tag{3}$$

pooled over all PDEL and OMIT variants.

## 4.3 Hypotheses

The following hypotheses and decision rules were fixed before the test data were generated.

- **H1a** (no reliable metric): no metric attains a 95% confidence-interval lower bound of at least 0.70 on every eligible error type.
- **H1b** (substitution blind spot of embedding and learned metrics): for BERTScore, MeaningBERT and LENS,

$$C_b(m) = \overline{\mathrm{mAUC}}_{\{\mathrm{NUM,ENT,NEG}\}}(m) - \mathrm{mAUC}_{\mathrm{OMIT}}(m) < 0. \tag{4}$$

- **H1c** (omission blind spot of factual-consistency metrics): for SummaC-ZS and AlignScore,

$$C_c(m) = \mathrm{mAUC}_{\mathrm{OMIT}}(m) - \overline{\mathrm{mAUC}}_{\{\mathrm{INS,NUM,ENT,NEG}\}}(m) < 0. \tag{5}$$

- **H2a**: $D = \mathrm{AUC}_{\mathrm{OP}}(\mathrm{BiNLI}) - \mathrm{AUC}_{\mathrm{OP}}(\mathrm{NLI\text{-}fwd}) \geq 0.10$ with a 95% confidence-interval lower bound above zero. **H2b** (no harm): the lower bound of $\mathrm{mAUC}_{\mathrm{INS}}(\mathrm{BiNLI}) - \mathrm{mAUC}_{\mathrm{INS}}(\mathrm{NLI\text{-}fwd})$ exceeds $-0.05$.
- **H3** (exploratory): rankings of the metrics by diagnostic performance agree with rankings by correlation with released human meaning ratings.

A family of tests (H1b or H1c) is supported if every Holm-adjusted test is significant in the hypothesized direction and partially supported if at least one is.

## 4.4 Statistical analysis

Uncertainty was estimated with a cluster bootstrap (Efron & Tibshirani, 1994) over the 359 source sentences, so that all bases and variants of a source were resampled together (2,000 resamples, shared across metrics, percentile 95% intervals). Contrasts

were tested with two-sided bootstrap $p$-values (resolution 1/2,000) and adjusted with the Holm procedure (Holm, 1979) within H1b (three tests) and H1c (two tests). For H3 we computed Spearman correlations between each metric and the mean meaning ratings of 100 system outputs in the ASSET human ratings (Alva-Manchego et al., 2020a) and the meaning z-scores of 600 outputs in Simplicity-DA (Alva-Manchego et al., 2021a, 2021b). Fluency confounds were examined with GPT-2 perplexity (Radford et al., 2019), both by stratifying mAUC by perplexity tertiles and with a generalized estimating equation (GEE; Liang & Zeger, 1986) logistic model of correct preference with sources as clusters.

### 4.5 Analysis plan and deviations

The hypotheses, outcome measures, eligibility rule, statistical procedure and decision rules were written down and frozen, together with cryptographic hashes of the generation and analysis code, before the test contrast sets were generated. The plan was not registered in an external registry; it is summarized, together with all subsequent changes, in Online Resource 1, and the frozen plan and hash record are included in the replication package (Demirkiran, 2026). Three amendments were made after freezing but before any test data existed: reference strings identical to a base or its REF variants were excluded after a data-quality check found such duplicates in 4.9% of validation bases; a sampling mode for the validity diagnostic was added; and the SummaC wrapper was corrected to load the specified `vitc-base` backbone, together with caching of repeated NLI computations. During execution the computation was relaunched once and reordered so that confirmatory computations ran first; these operational changes did not alter any analysis choice. The complete evaluation took 5.3 hours. An independent re-implementation of the confirmatory analysis from the stored per-item scores reproduced every confirmatory point estimate exactly.

### 4.6 Use of AI tools

The study was conducted with extensive assistance from generative AI systems (Claude, Anthropic; ChatGPT, OpenAI) under the author's direction. AI assistance included literature screening and extraction, drafting of the study design, implementation of the contrast-set generator, metric wrappers and analysis code, execution of the experiments, statistical analysis, generation of figures and drafting of the manuscript text; the independent re-implementation of the confirmatory analysis mentioned above was also carried out by an AI system. The author verified the cited sources, the archived data, the analysis code, the numerical results, the methodological decisions and the interpretations, and takes full responsibility for the content.

## 5 Results

### 5.1 Discrimination by error type

Fig. 2 shows mAUC for every metric and error type; confidence intervals for all cells are given in Online Resource 1 (Table S1). Four patterns stand out. The factual-consistency metrics (SummaC-ZS, AlignScore and NLI-fwd) separate negation,

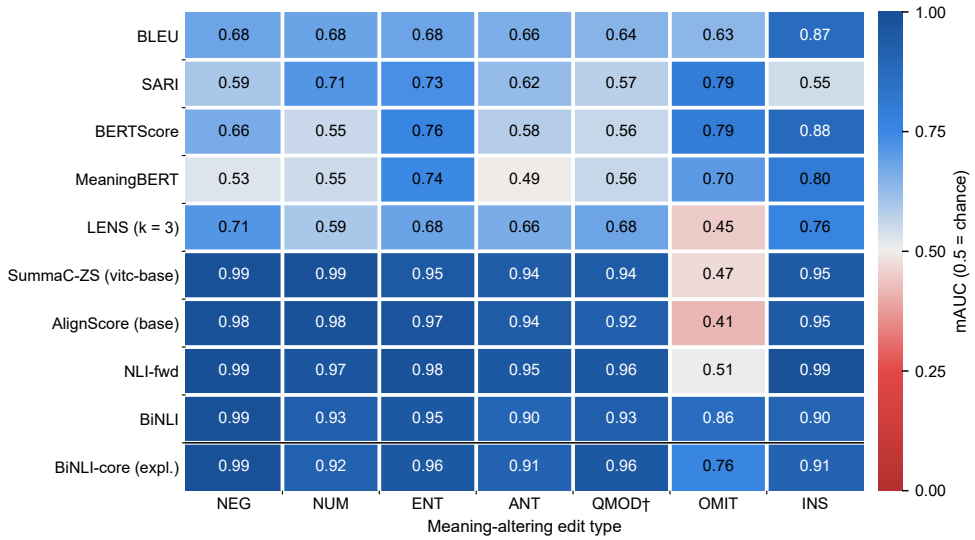



**Fig. 2** Matched AUC (mAUC, Eq. 2) for each metric and meaning-altering edit type on the CMED test resource. A value of 0.5 corresponds to chance; values below 0.5 indicate that the metric prefers the erroneous text. Values are printed in every cell. The horizontal line separates the exploratory BiNLI-core; †QMOD is descriptive only (78 variants)

number, entity, antonym and insertion edits from legitimate changes almost perfectly (0.94–0.99) but are at or below chance on omission (0.47, 0.41 and 0.51). BERTScore and MeaningBERT detect omission and insertion reasonably well (0.70–0.88) but are close to chance on number changes (0.55) and, for MeaningBERT, on negation (0.53) and antonyms (0.49). LENS is weakest on omission (0.45). SARI barely separates insertions from legitimate changes (0.55). BiNLI is the only score above 0.85 on every error type (lowest value 0.86, for omission).

## 5.2 Confirmatory tests

Table 5 summarizes the confirmatory tests. *H1a* was falsified: BiNLI had a confidence-interval lower bound of at least 0.70 on every eligible error type (lowest bound 0.82, for omission), whereas the lowest bounds of the other metrics ranged from 0.36 (AlignScore) to 0.58 (BLEU). *H1b* was partially supported. BERTScore ($C_b = -0.133$) and MeaningBERT ($-0.095$) discriminated substitutions significantly worse than omissions, whereas LENS showed a significant effect in the opposite direction ($+0.214$), because omission rather than substitution is its weakest type. *H1c* was supported for both factual-consistency metrics, with large effects ($C_c = -0.496$ for SummaC-ZS and $-0.557$ for AlignScore).

## 5.3 Omission versus peripheral deletion

On the extension set (839 omissions and 242 peripheral deletions), $\mathrm{AUC}_{\mathrm{OP}}$ was 0.572 [0.528, 0.615] for NLI-fwd and 0.639 [0.590, 0.687] for BiNLI (Online Resource 1,

**Table 5** Confirmatory results. Estimates with 95% cluster-bootstrap confidence intervals (2,000 resamples); $p$-values are Holm-adjusted within H1b and H1c (resolution 1/2,000)

| Hypothesis | Metric | Estimate | 95% CI | $p$ | Decision |
|---|---|---|---|---|---|
| H1b ($C_b$) | BERTScore | −0.133 | [−0.182, −0.079] | .0015 | partially supported |
| | MeaningBERT | −0.095 | [−0.151, −0.036] | .002 | |
| | LENS | +0.214 | [+0.154, +0.276] | .0015[a] | |
| H1c ($C_c$) | SummaC-ZS | −0.496 | [−0.554, −0.436] | .001 | supported |
| | AlignScore | −0.557 | [−0.610, −0.502] | .001 | |
| H2a ($D$) | BiNLI − NLI-fwd | +0.068 | [+0.014, +0.122] | .01[b] | not supported |
| H2b ($\Delta_{\text{INS}}$) | BiNLI − NLI-fwd | −0.085 | [−0.102, −0.067] | – | not supported |
| H1a | BiNLI | lowest lower bound 0.820[c] | | – | falsified |

[a]Significant in the direction opposite to H1b. [b]Unadjusted; the decision rule requires $D \geq 0.10$. [c]Lowest 95% confidence-interval lower bound across the eligible error types; H1a is falsified because the bound is at least 0.70 on every eligible type.

Fig. S2). The difference $D = +0.068$ [+0.014, +0.122] excludes zero, but its point estimate is below the pre-specified threshold of 0.10, so H2a is not supported; because the interval includes 0.10, an improvement of the targeted size cannot be ruled out. Adding the source direction also lowered insertion discrimination from 0.988 to 0.903 (−0.085 [−0.102, −0.067]), a larger loss than the pre-specified margin, so H2b is not supported either. The exploratory BiNLI-core reached 0.766 [0.724, 0.801], a value that is partly built in by its construction. By contrast, against the rewrite-dominated pool of the primary outcome, adding the source direction raised omission discrimination from 0.507 to 0.861 (Fig. 2; descriptive comparison).

## 5.4 Exploratory analyses

The magnitude-matched contrast is consistent with the primary pattern, although every cell contains fewer than 100 pairs (at most 91 for the types discussed here) and the results are therefore descriptive: SummaC-ZS, AlignScore and NLI-fwd obtained 0.52, 0.43 and 0.51 on omission, and MeaningBERT 0.47 on negation. Small and large number changes were detected almost equally by every metric; BERTScore reached only 0.59 even for large changes. In a small subset of 17 within-base pairs, SummaC-ZS and AlignScore scored omissions above peripheral deletions in about three of four pairs (0.24).

Fluency interacts with metric behaviour. In GEE models, correct preferences by BLEU, SARI, BERTScore, LENS and MeaningBERT depended strongly on how much the meaning-altering edit increased GPT-2 perplexity relative to the legitimate change (coefficients +0.35 to +0.68, $p < .001$). The dependence was smaller for NLI-fwd and BiNLI (+0.20, $p \approx .02$–$.03$), BiNLI-core (+0.27, $p = .002$) and AlignScore (+0.30, $p = .01$), and absent for SummaC-ZS (+0.05, $p = .61$).

On released human meaning ratings (H3), BERTScore (Spearman $\rho = 0.865$ on the ASSET ratings and 0.767 on Simplicity-DA) and MeaningBERT (0.701 and 0.854) correlated best, the factual-consistency metrics SummaC-ZS, AlignScore and NLI-fwd least (0.42–0.56), and the bidirectional scores in between (0.64–0.68). The rank

correlation between diagnostic performance (mean mAUC over eligible types) and rating correlation was slightly negative and not significant ($\rho = -0.26$, $p = .47$, and $\rho = -0.13$, $p = .73$; ten scores).

# 6 Discussion

## 6.1 Complementary blind spots follow from how metrics compare texts

The central finding is that the metrics fail in different, largely predictable ways rather than to different degrees. Factual-consistency metrics ask whether the output is supported by the source. An output that silently drops a main assertion remains fully supported, so these metrics cannot see omission; this is a consequence of their design (Laban et al., 2022; Zha et al., 2023), and CMED quantifies it for simplification. The same design makes them sensitive to local substitutions, which contradict the source even when almost all words are shared. Embedding and learned metrics behave in the opposite way: they respond to how much content changes, which makes omission and insertion visible, but a changed number or an inserted negation alters a single token in an otherwise identical sentence and moves the embedding comparison very little. This pattern was anticipated by earlier work: MENLI found standard metrics weak on name and number errors but competitive on omission and addition, and NLI-based metrics much stronger on names, negation and numbers (Chen & Eger, 2023); BERTScore correlated with annotated deletion errors but only weakly with substitution errors (Devaraj et al., 2022); and no metric detected number swaps in plain-language summaries (Guo et al., 2024). CMED confirms this pattern for simplification with typed contrasts against legitimate changes. The CMED results suggest an error-type explanation for the near-chance error detection observed in naturally occurring outputs (Vendeville et al., 2025; Wu & Arase, 2026): a metric can appear uninformative on mixed natural errors while being highly informative for some error types and blind to others.

An alternative explanation for the success of some metrics is that they detect disfluency rather than meaning change, since rule-based edits can make a sentence less natural. The exploratory GEE results support this reading for BLEU, SARI, BERTScore, LENS and MeaningBERT, whose correct preferences depended strongly on the perplexity increase caused by an edit, but much less so for the entailment-based scores. Part of the apparent sensitivity of surface and embedding metrics may therefore not transfer to fluent meaning errors of the kind that LLMs produce. This interpretation is correlational and should be tested with fluency-matched edits.

## 6.2 Correlation with ratings is not error detection

The metrics that correlate best with averaged human meaning ratings were not the ones that best detected meaning errors in CMED. Although this analysis was exploratory and the rank correlations were not significant, the pattern is consistent with two known properties of rating-based validation: meaning ratings are entangled with fluency and simplicity judgments (Cripwell, Legrand, & Gardent, 2023; Scialom

et al., 2021), and correlations with them depend on the quality range and system type (Alfear et al., 2024; Alva-Manchego et al., 2021b). A high correlation with averaged ratings therefore does not guarantee that a metric flags the specific error that makes a simplification untrustworthy. Targeted, error-typed diagnostics complement correlation-based meta-evaluation rather than replace it: correlations indicate how a metric tracks overall perceived quality, and diagnostics indicate which errors it can and cannot see.

### 6.3 The legitimacy problem remains open

Bidirectional entailment was the strongest single score against alternative human rewrites, but it separated the omission of a main assertion from the deletion of peripheral content only moderately and lost sensitivity to inserted content. In our interpretation, the first result reflects a central difficulty of evaluating simplification, namely deciding what may be dropped and what must be kept: information loss is inherent to simplification (Trienes et al., 2024), and judging the relevance of deleted content is inherently ambiguous (Heineman et al., 2023). The source direction of entailment penalizes any deletion, peripheral or not, and the exploratory BiNLI-core suggests that an explicit notion of main-clause content helps, although its advantage here is partly built in. The second result arises because averaging both directions dilutes the strong penalty that forward entailment assigns to unsupported content. Neither result depends on large models, which suggests that the limitations lie in how meaning preservation is operationalized rather than in model capacity; whether larger NLI models or LLM-based judges behave differently was outside the computational scope of this study.

### 6.4 Recommendations

Within the scope of this study, four recommendations follow. First, SARI and BERTScore alone should not be used to certify meaning preservation. Second, evaluations should pair a forward factual-consistency metric, which covers substitutions and insertions, with a measure that is sensitive to omission, such as bidirectional entailment (Chen & Eger, 2023) or question-based coverage (Trienes et al., 2024); where resources permit, human reading-comprehension evaluation offers a complementary extrinsic check (Agrawal & Carpuat, 2024). Third, in high-stakes domains, numbers and negations deserve targeted checks, because several widely used metrics are close to chance on them. Fourth, papers that propose meaning-preservation metrics should report performance per error type, for example on CMED, in addition to correlations with ratings.

### 6.5 Limitations

The meaning-altering edits are synthetic, rule-based and applied to English Wikipedia sentences; they approximate, but do not reproduce, the errors made by LLMs, and naturally occurring error annotations (Vendeville et al., 2025; Wu & Arase, 2026) bound the generalization of our findings. The legitimate comparison pool is dominated by alternative human simplifications (96.8% of legitimate variants), and the samples

of peripheral deletions and splits in the main set are small, so conclusions about these legitimate operations are less certain; human references themselves can contain errors (Devaraj et al., 2022; Ma, Seneviratne, & Daskalaki, 2022). The validity of the edits was checked only automatically, by an NLI model; 15–22% of entity, number and insertion edits were judged entailed, which biases the corresponding estimates toward chance. Omission is represented by 179 test variants. Three metrics were used in smaller configurations for computational reasons, and LLM-based judges, document-level, domain-specific and multilingual simplification were not evaluated.

# 7 Conclusion

CMED asks whether a meaning-preservation metric ranks a legitimate simplification above a minimally edited, meaning-altering one, separately for each type of meaning error. Across nine CPU-feasible scores, most metrics showed complementary blind spots: factual-consistency metrics were nearly perfect on substitutions and insertions but at or below chance on omission, BERTScore missed number changes, and MeaningBERT missed numbers, negation and antonyms. A bidirectional entailment score was the only one to exceed 0.70 on every error type against alternative human rewrites, yet it separated legitimate from harmful deletion only weakly and lost sensitivity to insertion. Trustworthy evaluation of simplification, including simplification by large language models, should therefore combine metrics chosen for their coverage of error types and should validate new metrics with targeted diagnostics in addition to correlations with human ratings.

**Supplementary information.** Online Resource 1 contains a summary of the analysis plan with its amendments and deviations, the correspondence between the terms used here and the identifiers in the code, confidence intervals for all results (Table S1), the magnitude-matched, preservation-proportion, number-magnitude, legitimate-type and omission-versus-deletion results, the external-validity, GEE and perplexity-tertile results, the validity-diagnostic details, and Figs. S1–S2.

# Statements and Declarations

**Funding.** No funding was received for conducting this study.

**Competing interests.** The author has no relevant financial or non-financial interests to disclose.

**Ethics approval and consent to participate.** This study did not involve the recruitment of human participants, an intervention, or the collection of new personal data. The analyses used publicly available, previously released text-simplification datasets (ASSET and Simplicity-DA) and their published human-generated simplifications and ratings. No participant identifiers were used and no attempt was made to re-identify contributors. Accordingly, no new informed consent was obtained for this secondary computational analysis, and no new human-subject data were generated.

**Data availability.** The study used publicly available datasets from their original providers: ASSET (Alva-Manchego et al., 2020b) for the primary contrastive diagnostic and Simplicity-DA (Alva-Manchego et al., 2021a) for the external-validity analysis. ASSET is licensed under CC BY-NC 4.0 and Simplicity-DA-derived material under CC BY-NC-SA 4.0. The original datasets are not redistributed as independent copies. The derived contrast sets, per-item metric scores, analysis outputs and reproducibility materials are archived in the accompanying Zenodo replication package (Demirkiran, 2026), subject to the applicable source-data licences.

**Code availability.** The contrast-set generation code, metric wrappers, analysis code, audit scripts, tests and reproduction instructions are archived on Zenodo (Demirkiran, 2026). Original software developed for this study is released under the MIT License. Third-party software, pretrained models and derived datasets remain subject to their respective original licences.

**Author contributions.** Emin Talip Demirkiran: Conceptualization, Supervision, Project administration, Validation, and Writing – review & editing. The author is the sole author, directed the study throughout, and takes responsibility for the integrity of the work, for the verification of its sources and results, and for the approval of the final version.

**Use of generative AI.** Generative AI tools (Claude and ChatGPT) were used throughout this work. Beyond manuscript preparation (literature exploration, organisation and restructuring, language refinement and reviewer-style critique), an AI system, under the direction of the author, implemented the analysis code, carried out the experiments and produced the reported analyses and figures. The author independently checked the cited sources, the archived data, the analysis code, the numerical results, the methodological decisions and the interpretations, and takes full responsibility for the final manuscript. Generative AI was not treated as an author.

## References


Agrawal, S., & Carpuat, M. (2024). Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, *12*, 432–448, https://doi.org/10.1162/tacl_a_00653

Alfear, N.A., Kazakov, D., Al-Khalifa, H. (2024). Meta-evaluation of sentence simplification metrics. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 11229–11235). Torino, Italia: ELRA and ICCL. Retrieved from https://aclanthology.org/2024.lrec-main.981/

Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., Specia, L. (2020a). ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. *Proceedings of the 58th Annual Meeting*

*of the Association for Computational Linguistics* (pp. 4668–4679). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.424/

Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., Specia, L. (2020b). *ASSET [data set].* GitHub repository, commit 9d65904. Retrieved from https://github.com/facebookresearch/asset

Alva-Manchego, F., Scarton, C., Specia, L. (2021a). *Simplicity-DA [data set].* GitHub repository metaeval-simplification, commit 7703c46. Retrieved from https://github.com/feralvam/metaeval-simplification

Alva-Manchego, F., Scarton, C., Specia, L. (2021b). The (Un)Suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, *47*(4), 861–889, https://doi.org/10.1162/coli_a_00418

Beauchemin, D., Saggion, H., Khoury, R. (2023). MeaningBERT: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, *6*, 1223924, https://doi.org/10.3389/frai.2023.1223924

Chen, Y., & Eger, S. (2023). MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, *11*, 804–825, https://doi.org/10.1162/tacl_a_00576

Cripwell, L., Legrand, J., Gardent, C. (2023). Simplicity level estimate (SLE): A learned reference-less metric for sentence simplification. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12053–12059). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.739/

Cripwell, L., Legrand, J., Gardent, C. (2024). Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation. *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024* (pp. 1–14). European Language Resources Association (ELRA) and ICCL.

Demirkiran, E.T. (2026). *Replication package for "CMED: A contrastive diagnostic of meaning-preservation metrics for text simplification" (version 1.0.0) [software and data set].* Zenodo.

Devaraj, A., Sheffield, W., Wallace, B.C., Li, J.J. (2022). Evaluating factuality in text simplification. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7331–7345). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://

aclanthology.org/2022.acl-long.506/

Efron, B., & Tibshirani, R.J. (1994). *An introduction to the bootstrap.* Chapman and Hall/CRC.

Feng, Y., Qiang, J., Li, Y., Yuan, Y., Zhu, Y. (2023). *Sentence simplification via large language models.* arXiv preprint arXiv:2302.11957. Retrieved from https://arxiv.org/abs/2302.11957

Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., ... Zhou, B. (2020, November). Evaluating models' local decision boundaries via contrast sets. T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1307–1323). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.117/

Guo, Y., August, T., Leroy, G., Cohen, T., Wang, L.L. (2024). APPLS: Evaluating evaluation metrics for plain language summarization. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 9194–9211). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.emnlp-main.519/

He, P., Liu, X., Gao, J., Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. *International Conference on Learning Representations.* Retrieved from https://arxiv.org/abs/2006.03654

Heineman, D., Dou, Y., Maddela, M., Xu, W. (2023). Dancing between success and failure: Edit-level simplification evaluation using SALSA. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 3466–3495). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.211/

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70, https://doi.org/10.2307/4615733

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python.* Zenodo.

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A.T., Topalis, J., ... Ingrisch, M. (2024). ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European Radiology*, *34*(5), 2817–2825, https://doi.org/10.1007/s00330-023-10213-1

Joseph, S., Chen, L., Trienes, J., Göke, H., Coers, M., Xu, W., ... Li, J.J. (2024). FactPICO: Factuality evaluation for plain language summarization of medical

evidence. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8437–8464). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.acl-long.459/

Kew, T., Chi, A., Vásquez-Rodríguez, L., Agrawal, S., Aumiller, D., Alva-Manchego, F., Shardlow, M. (2023). BLESS: Benchmarking large language models on sentence simplification. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 13291–13309). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.821/

Korobeynikova, M., Battisti, A., Fischer, L., Gao, Y. (2026). DETECT: Determining ease and textual clarity of German text simplifications. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2852–2882). Rabat, Morocco: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2026.eacl-long.131/

Laban, P., Schnabel, T., Bennett, P.N., Hearst, M.A. (2022). SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, *10*, 163–177, https://doi.org/10.1162/tacl_a_00453

Liang, K.-Y., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22, https://doi.org/10.1093/biomet/73.1.13

Ma, Y., Seneviratne, S., Daskalaki, E. (2022, December). Improving text simplification with factuality error detection. S. Štajner et al. (Eds.), *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)* (pp. 173–178). Abu Dhabi, United Arab Emirates (Virtual): Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.tsar-1.16/

Maddela, M., Dou, Y., Heineman, D., Xu, W. (2023). LENS: A learnable evaluation metric for text simplification. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 16383–16408). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.acl-long.905/

Martin, L., Humeau, S., Mazaré, P.-E., de La Clergerie, É., Bordes, A., Sagot, B. (2018, November). Reference-less quality estimation of text simplification systems. A. Jönsson, E. Rennes, H. Saggion, S. Stajner, & V. Yaneva (Eds.), *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)* (pp. 29–38). Tilburg, the Netherlands: Association for Computational Linguistics.

Retrieved from https://aclanthology.org/W18-7005/

Marturi, K.C., & Elwazzan, H.H. (2025). Hallucination detection and mitigation in scientific text simplification using ensemble approaches: DS@GT at CLEF 2025 SimpleText. G. Faggioli, N. Ferro, P. Rosso, & D. Spina (Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)* (Vol. 4038, pp. 4324–4337). Madrid, Spain: CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-4038/paper_356.pdf

Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41, https://doi.org/10.1145/219717.219748

Mujahid, Z.M., Wright, D., Augenstein, I. (2026). Stress testing factual consistency metrics for long-document summarization. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 31914–31933). San Diego, California, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2026.acl-long.1472/

Nandiraju, A.K.S.K., Leroy, G., Kauchak, D., Ahmed, A. (2025). *Automated feedback loops to protect text simplification with generative AI from information loss.* arXiv preprint arXiv:2505.16172.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002, July). Bleu: a method for automatic evaluation of machine translation. P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P02-1040/

Post, M. (2018, October). A call for clarity in reporting BLEU scores. O. Bojar et al. (Eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-6319/

Qiang, J., Huang, M., Zhu, Y., Yuan, Y., Zhang, C., Yu, K. (2025). *Redefining simplicity: Benchmarking large language models from lexical to document simplification.* arXiv preprint arXiv:2502.08281. Retrieved from https://arxiv.org/abs/2502.08281

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). *Language models are unsupervised multitask learners* (Tech. Rep.). OpenAI. Retrieved from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Schuster, T., Fisch, A., Barzilay, R. (2021, June). Get your vitamin C! robust fact verification with contrastive evidence. K. Toutanova et al. (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 624–643). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.52/

Scialom, T., Martin, L., Staiano, J., Villemonte de la Clergerie, É., Sagot, B. (2021). *Rethinking automatic evaluation in sentence simplification.* arXiv preprint arXiv:2104.07560. Retrieved from https://arxiv.org/abs/2104.07560

Sulem, E., Abend, O., Rappoport, A. (2018a). BLEU is not suitable for the evaluation of text simplification. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 738–744). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1081/

Sulem, E., Abend, O., Rappoport, A. (2018b). Semantic structural evaluation for text simplification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 685–696). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-1063/

Tanprasert, T., & Kauchak, D. (2021). Flesch-Kincaid is not a text simplification evaluation metric. *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)* (pp. 1–14). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.gem-1.1/

Trienes, J., Joseph, S., Schlötterer, J., Seifert, C., Lo, K., Xu, W., … Li, J.J. (2024). InfoLossQA: Characterizing and recovering information loss in text simplification. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4263–4294). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.acl-long.234/

Vendeville, B., Ermakova, L., De Loor, P. (2025). Resource for error analysis in text simplification: New taxonomy and test collection. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3723–3732). Padua, Italy: Association for Computing Machinery.

Williams, A., Nangia, N., Bowman, S.R. (2018, June). A broad-coverage challenge corpus for sentence understanding through inference. M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

*Volume 1 (Long Papers)* (pp. 1112–1122). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-1101/

Wu, X., & Arase, Y. (2026). An in-depth evaluation of large language models in sentence simplification with error-based human assessment. *ACM Transactions on Intelligent Systems and Technology*, *17*(4), 1–25, https://doi.org/10.1145/3744744

Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, *4*, 401–415, https://doi.org/10.1162/tacl_a_00107

Zha, Y., Yang, Y., Li, R., Hu, Z. (2023). AlignScore: Evaluating factual consistency with a unified alignment function. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11328–11348). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.acl-long.634/

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y. (2019). *BERTScore: Evaluating text generation with BERT.* arXiv preprint arXiv:1904.09675.

Zhao, X., Durmus, E., Yeung, D.-Y. (2023). Towards reference-free text simplification evaluation with a BERT Siamese network architecture. *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13250–13264). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.findings-acl.838/

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf