Online Resource 1

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

Journal: *Language Resources and Evaluation*

Author: Emin Talip Demirkiran, Department of Computer Engineering, Eskisehir Technical University, Iki Eylul Campus, 26555 Eskisehir, Türkiye

Contact: etd@eskisehir.edu.tr; ORCID 0000-0001-6063-8178

S1 Analysis plan

The analysis plan was frozen on 2026-09-29 22:28 UTC, before the test contrast sets were generated. It fixed the hypotheses (H1a–H1c, H2a–H2b and the exploratory H3), the primary outcome (matched AUC per error type), the eligibility rule for confirmatory error types (at least 100 variants on the test set), the cluster bootstrap over source sentences (2,000 resamples, seed 20260930), Holm adjustment within H1b and within H1c, the decision rules, and a computational budget (estimate 4.6 h, hard limit 6 h on 12 CPU threads, no GPU, no model training, no text generation). SHA-256 hashes of the edit generator, the contrast-set builder, the metric wrappers, the SARI implementation and the validity diagnostic were recorded at freezing; the analysis code was frozen separately before any score was computed. The plan was not registered in an external registry. The complete plan, the hash record and the amendment log are included verbatim in the replication package (folder docs/).

S2 Amendments before test data existed

A1 (2026-09-29 22:39 UTC; file src/cmed/build_contrast_set.py). Reference strings identical to the base or to one of its REF variants were excluded from the references given to reference-based metrics, and references were de-duplicated. A data-quality check on the validation set had found such duplicates (identical simplifications by different ASSET annotators) for 4.9% of bases and 4.8% of REF variants; they would have given reference-based metrics an exact match on some legitimate variants. Reference-free metrics are unaffected.

A2 (2026-09-29 22:39 UTC; file src/cmed/validity.py). A type-stratified sampling mode (25% of each type) was added to the validity diagnostic so that it matches the plan; the frozen version supported only a fixed number of items per type. Diagnostics only.

A3 (2026-09-29 23:06 UTC; file src/cmed/metrics.py). The SummaC-ZS wrapper was corrected to load the vitc-base backbone specified in the plan (the frozen code still loaded the larger default), and entailment probabilities were cached per premise–hypothesis pair so that NLI-fwd, BiNLI and BiNLI-core share computations. The caching changes run time only.

S3 Operational deviations during execution

D1. The first launch ran as a detached background process and progressed at about half the speed observed in the trial run on the validation set; the cause was attributed to operating-system power management of background processes (inferred from low processor utilisation, not verified with an operating-system tool). The identical frozen command was relaunched from an attached shell about one hour after the first launch. Completed per-item scores were reused from the cache, the regenerated test contrast sets were byte-identical to the first generation, and the six-hour limit was counted from the original launch.

D2. To keep the confirmatory computations within the six-hour limit, the same frozen commands were executed in a different order: confirmatory metric computations first, followed by perplexity, the validity diagnostic, the exploratory BiNLI-core score and the H3 analysis. Scores are deterministic and independent of the order. No analysis choice was changed, and no outcome was inspected before all computations had finished. All computations completed within the limit.

S4 Terminology

The manuscript uses descriptive names; the code and data files of the replication package use the short identifiers below.

Table S0 Correspondence between manuscript terms and identifiers in the code and data

Manuscript	Code and data
NLI-fwd	NLI-P
BiNLI	D2-sym
BiNLI-core (exploratory)	D2-core
LENS (three references)	LENS_k3
AlignScore (base checkpoint)	AlignScore-base
meaning-altering edit types NEG, NUM, ENT, ANT, QMOD, OMIT, INS	I-NEG, I-NUM, I-ENT, I-ANT, I-QMOD, I-OMIT, I-INS
legitimate change types REF, PDEL, SPLIT	L-REF, L-PDEL, L-SPLIT
meaning-altering / legitimate	ILLEGIT / LEGIT
matched AUC (primary outcome, all pairs)	E1 (mAUC)
magnitude-matched contrast	E2
proportion of edits that lower the score	PA
omission vs peripheral deletion extension set	h2ext
analysis plan	protocol

S5 Additional results

Table S1 Matched AUC (primary outcome) with 95% cluster-bootstrap confidence intervals

Metric	Type	mAUC	95% CI	Pairs
BLEU	NEG	0.678	[0.651, 0.704]	1,425
BLEU	NUM	0.682	[0.636, 0.729]	764
BLEU	ENT	0.676	[0.642, 0.710]	995
BLEU	ANT	0.656	[0.613, 0.702]	539
BLEU	QMOD	0.642	[0.567, 0.719]	162
BLEU	OMIT	0.634	[0.576, 0.690]	381
BLEU	INS	0.866	[0.847, 0.883]	1,480
SARI	NEG	0.594	[0.564, 0.623]	1,425
SARI	NUM	0.712	[0.657, 0.768]	764
SARI	ENT	0.727	[0.696, 0.759]	995
SARI	ANT	0.622	[0.571, 0.669]	539
SARI	QMOD	0.574	[0.495, 0.657]	162
SARI	OMIT	0.793	[0.748, 0.835]	381
SARI	INS	0.545	[0.516, 0.574]	1,480
BERTScore	NEG	0.662	[0.633, 0.692]	1,425
BERTScore	NUM	0.554	[0.497, 0.608]	764
BERTScore	ENT	0.756	[0.725, 0.787]	995
BERTScore	ANT	0.581	[0.531, 0.633]	539
BERTScore	QMOD	0.556	[0.467, 0.650]	162
BERTScore	OMIT	0.790	[0.739, 0.838]	381
BERTScore	INS	0.882	[0.862, 0.902]	1,480
MeaningBERT	NEG	0.527	[0.496, 0.561]	1,425

Metric	Type	mAUC	95% CI	Pairs
MeaningBERT	NUM	0.548	[0.491, 0.603]	764
MeaningBERT	ENT	0.743	[0.708, 0.777]	995
MeaningBERT	ANT	0.494	[0.444, 0.548]	539
MeaningBERT	QMOD	0.556	[0.480, 0.639]	162
MeaningBERT	OMIT	0.701	[0.645, 0.755]	381
MeaningBERT	INS	0.795	[0.771, 0.819]	1,480
LENS	NEG	0.712	[0.681, 0.740]	1,425
LENS	NUM	0.594	[0.537, 0.648]	764
LENS	ENT	0.682	[0.649, 0.716]	995
LENS	ANT	0.657	[0.610, 0.701]	539
LENS	QMOD	0.679	[0.591, 0.758]	162
LENS	OMIT	0.449	[0.386, 0.513]	381
LENS	INS	0.759	[0.732, 0.785]	1,480
SummaC-ZS	NEG	0.985	[0.977, 0.992]	1,425
SummaC-ZS	NUM	0.991	[0.982, 0.998]	764
SummaC-ZS	ENT	0.950	[0.937, 0.962]	995
SummaC-ZS	ANT	0.942	[0.917, 0.965]	539
SummaC-ZS	QMOD	0.938	[0.889, 0.975]	162
SummaC-ZS	OMIT	0.472	[0.416, 0.532]	381
SummaC-ZS	INS	0.948	[0.935, 0.961]	1,480
AlignScore	NEG	0.981	[0.973, 0.988]	1,425
AlignScore	NUM	0.984	[0.973, 0.994]	764
AlignScore	ENT	0.974	[0.963, 0.984]	995
AlignScore	ANT	0.942	[0.918, 0.965]	539
AlignScore	QMOD	0.920	[0.862, 0.966]	162
AlignScore	OMIT	0.415	[0.361, 0.471]	381
AlignScore	INS	0.947	[0.935, 0.958]	1,480
NLI-fwd	NEG	0.993	[0.988, 0.997]	1,425
NLI-fwd	NUM	0.971	[0.953, 0.986]	764
NLI-fwd	ENT	0.983	[0.973, 0.992]	995
NLI-fwd	ANT	0.954	[0.928, 0.975]	539
NLI-fwd	QMOD	0.957	[0.906, 0.994]	162
NLI-fwd	OMIT	0.507	[0.447, 0.567]	381
NLI-fwd	INS	0.988	[0.981, 0.994]	1,480
BiNLI	NEG	0.986	[0.979, 0.992]	1,425
BiNLI	NUM	0.925	[0.895, 0.954]	764
BiNLI	ENT	0.954	[0.936, 0.970]	995
BiNLI	ANT	0.896	[0.858, 0.931]	539
BiNLI	QMOD	0.926	[0.871, 0.968]	162
BiNLI	OMIT	0.861	[0.820, 0.899]	381
BiNLI	INS	0.903	[0.884, 0.921]	1,480
BiNLI-core	NEG	0.988	[0.981, 0.994]	1,425
BiNLI-core	NUM	0.918	[0.885, 0.947]	764
BiNLI-core	ENT	0.957	[0.941, 0.972]	995
BiNLI-core	ANT	0.909	[0.875, 0.941]	539
BiNLI-core	QMOD	0.957	[0.914, 0.987]	162
BiNLI-core	OMIT	0.756	[0.703, 0.808]	381
BiNLI-core	INS	0.908	[0.891, 0.925]	1,480

QMOD is descriptive only (78 variants). BiNLI-core is exploratory.

Table S2 Magnitude-matched contrast (pairs whose normalized edit distances differ by at most 0.10)

Metric	Type	mAUC	95% CI	Pairs
BLEU	NEG	0.807	[0.722, 0.885]	88
BLEU	NUM	0.864	[0.714, 1.000]	44
BLEU	ENT	0.759	[0.658, 0.848]	79
BLEU	ANT	0.632	[0.375, 0.842]	19
BLEU	QMOD	0.500	[0.000, 1.000]	6
BLEU	OMIT	0.714	[0.606, 0.810]	91
BLEU	INS	0.918	[0.887, 0.946]	353

Metric	Type	mAUC	95% CI	Pairs
SARI	NEG	0.648	[0.538, 0.753]	88
SARI	NUM	0.727	[0.522, 0.900]	44
SARI	ENT	0.886	[0.806, 0.949]	79
SARI	ANT	0.737	[0.500, 0.929]	19
SARI	QMOD	0.833	[0.500, 1.000]	6
SARI	OMIT	0.868	[0.798, 0.932]	91
SARI	INS	0.564	[0.507, 0.617]	353
BERTScore	NEG	0.659	[0.558, 0.760]	88
BERTScore	NUM	0.432	[0.231, 0.632]	44
BERTScore	ENT	0.848	[0.768, 0.926]	79
BERTScore	ANT	0.474	[0.235, 0.750]	19
BERTScore	QMOD	0.167	[0.000, 0.500]	6
BERTScore	OMIT	0.879	[0.812, 0.938]	91
BERTScore	INS	0.952	[0.924, 0.976]	353
MeaningBERT	NEG	0.466	[0.358, 0.588]	88
MeaningBERT	NUM	0.511	[0.274, 0.731]	44
MeaningBERT	ENT	0.823	[0.738, 0.905]	79
MeaningBERT	ANT	0.263	[0.048, 0.500]	19
MeaningBERT	QMOD	0.500	[0.000, 1.000]	6
MeaningBERT	OMIT	0.736	[0.634, 0.832]	91
MeaningBERT	INS	0.839	[0.792, 0.882]	353
LENS	NEG	0.784	[0.684, 0.868]	88
LENS	NUM	0.750	[0.526, 0.933]	44
LENS	ENT	0.810	[0.729, 0.887]	79
LENS	ANT	0.789	[0.591, 0.957]	19
LENS	QMOD	1.000	[1.000, 1.000]	6
LENS	OMIT	0.407	[0.287, 0.522]	91
LENS	INS	0.768	[0.720, 0.814]	353
SummaC-ZS	NEG	0.989	[0.962, 1.000]	88
SummaC-ZS	NUM	1.000	[1.000, 1.000]	44
SummaC-ZS	ENT	0.975	[0.937, 1.000]	79
SummaC-ZS	ANT	0.947	[0.818, 1.000]	19
SummaC-ZS	QMOD	1.000	[1.000, 1.000]	6
SummaC-ZS	OMIT	0.516	[0.396, 0.629]	91
SummaC-ZS	INS	0.958	[0.934, 0.978]	353
AlignScore	NEG	0.977	[0.941, 1.000]	88
AlignScore	NUM	1.000	[1.000, 1.000]	44
AlignScore	ENT	0.987	[0.956, 1.000]	79
AlignScore	ANT	0.947	[0.818, 1.000]	19
AlignScore	QMOD	1.000	[1.000, 1.000]	6
AlignScore	OMIT	0.429	[0.318, 0.537]	91
AlignScore	INS	0.955	[0.928, 0.977]	353
NLI-fwd	NEG	1.000	[1.000, 1.000]	88
NLI-fwd	NUM	0.977	[0.921, 1.000]	44
NLI-fwd	ENT	0.987	[0.959, 1.000]	79
NLI-fwd	ANT	1.000	[1.000, 1.000]	19
NLI-fwd	QMOD	1.000	[1.000, 1.000]	6
NLI-fwd	OMIT	0.505	[0.397, 0.609]	91
NLI-fwd	INS	0.994	[0.986, 1.000]	353
BiNLI	NEG	0.989	[0.961, 1.000]	88
BiNLI	NUM	0.955	[0.875, 1.000]	44
BiNLI	ENT	0.962	[0.915, 1.000]	79
BiNLI	ANT	0.842	[0.647, 1.000]	19
BiNLI	QMOD	1.000	[1.000, 1.000]	6
BiNLI	OMIT	0.923	[0.855, 0.978]	91
BiNLI	INS	0.921	[0.887, 0.952]	353
BiNLI-core	NEG	1.000	[1.000, 1.000]	88
BiNLI-core	NUM	1.000	[1.000, 1.000]	44
BiNLI-core	ENT	0.987	[0.959, 1.000]	79
BiNLI-core	ANT	1.000	[1.000, 1.000]	19
BiNLI-core	QMOD	1.000	[1.000, 1.000]	6
BiNLI-core	OMIT	0.769	[0.667, 0.863]	91
BiNLI-core	INS	0.926	[0.895, 0.955]	353

Exploratory and descriptive: every cell contains fewer than 100 pairs.

Table S3 Proportion of meaning-altering edits that lower the metric score, P[$\Delta < 0$]

Metric	Type	Proportion	95% CI	Edits
BLEU	NEG	0.957	[0.942, 0.972]	692
BLEU	NUM	0.986	[0.973, 0.997]	364
BLEU	ENT	0.976	[0.964, 0.987]	480
BLEU	ANT	0.915	[0.889, 0.940]	260
BLEU	QMOD	0.859	[0.801, 0.908]	78
BLEU	OMIT	0.670	[0.591, 0.745]	179
BLEU	INS	0.999	[0.996, 1.000]	718
SARI	NEG	0.822	[0.794, 0.849]	692
SARI	NUM	0.901	[0.856, 0.941]	364
SARI	ENT	0.973	[0.958, 0.986]	480
SARI	ANT	0.848	[0.805, 0.888]	260
SARI	QMOD	0.788	[0.699, 0.864]	78
SARI	OMIT	0.978	[0.954, 0.995]	179
SARI	INS	0.633	[0.593, 0.672]	718
BERTScore	NEG	0.967	[0.953, 0.980]	692
BERTScore	NUM	0.920	[0.885, 0.951]	364
BERTScore	ENT	0.975	[0.959, 0.989]	480
BERTScore	ANT	0.892	[0.854, 0.929]	260
BERTScore	QMOD	0.795	[0.699, 0.890]	78
BERTScore	OMIT	0.927	[0.888, 0.964]	179
BERTScore	INS	1.000	[1.000, 1.000]	718
MeaningBERT	NEG	0.566	[0.525, 0.607]	692
MeaningBERT	NUM	0.709	[0.654, 0.759]	364
MeaningBERT	ENT	0.938	[0.912, 0.961]	480
MeaningBERT	ANT	0.723	[0.660, 0.785]	260
MeaningBERT	QMOD	0.635	[0.516, 0.753]	78
MeaningBERT	OMIT	0.832	[0.771, 0.885]	179
MeaningBERT	INS	0.950	[0.932, 0.967]	718
LENS	NEG	0.879	[0.851, 0.905]	692
LENS	NUM	0.717	[0.657, 0.775]	364
LENS	ENT	0.810	[0.774, 0.847]	480
LENS	ANT	0.823	[0.771, 0.875]	260
LENS	QMOD	0.897	[0.827, 0.958]	78
LENS	OMIT	0.503	[0.424, 0.581]	179
LENS	INS	0.811	[0.780, 0.841]	718
SummaC-ZS	NEG	0.984	[0.974, 0.993]	692
SummaC-ZS	NUM	0.992	[0.977, 1.000]	364
SummaC-ZS	ENT	0.977	[0.963, 0.990]	480
SummaC-ZS	ANT	0.965	[0.941, 0.985]	260
SummaC-ZS	QMOD	0.962	[0.912, 1.000]	78
SummaC-ZS	OMIT	0.419	[0.343, 0.494]	179
SummaC-ZS	INS	0.976	[0.965, 0.987]	718
AlignScore	NEG	0.968	[0.952, 0.981]	692
AlignScore	NUM	0.984	[0.968, 0.997]	364
AlignScore	ENT	0.990	[0.980, 0.998]	480
AlignScore	ANT	0.977	[0.958, 0.992]	260
AlignScore	QMOD	0.962	[0.915, 1.000]	78
AlignScore	OMIT	0.385	[0.312, 0.458]	179
AlignScore	INS	0.983	[0.972, 0.993]	718
NLI-fwd	NEG	0.986	[0.975, 0.994]	692
NLI-fwd	NUM	0.981	[0.964, 0.995]	364
NLI-fwd	ENT	0.988	[0.975, 0.998]	480
NLI-fwd	ANT	0.981	[0.963, 0.996]	260
NLI-fwd	QMOD	0.987	[0.959, 1.000]	78
NLI-fwd	OMIT	0.547	[0.478, 0.625]	179
NLI-fwd	INS	0.993	[0.986, 0.999]	718
BiNLI	NEG	0.988	[0.981, 0.996]	692
BiNLI	NUM	0.978	[0.963, 0.992]	364
BiNLI	ENT	0.985	[0.974, 0.996]	480

Metric	Type	Proportion	95% CI	Edits
BiNLI	ANT	0.965	[0.938, 0.989]	260
BiNLI	QMOD	0.962	[0.913, 1.000]	78
BiNLI	OMIT	0.944	[0.905, 0.978]	179
BiNLI	INS	0.989	[0.981, 0.996]	718
BiNLI-core	NEG	0.991	[0.984, 0.997]	692
BiNLI-core	NUM	0.973	[0.954, 0.989]	364
BiNLI-core	ENT	0.983	[0.972, 0.994]	480
BiNLI-core	ANT	0.965	[0.939, 0.988]	260
BiNLI-core	QMOD	0.974	[0.935, 1.000]	78
BiNLI-core	OMIT	0.849	[0.792, 0.901]	179
BiNLI-core	INS	0.990	[0.982, 0.997]	718

Exploratory.

Table S4 Matched AUC for small and large number changes

Metric	Small	95% CI	Large	95% CI
BLEU	0.682	[0.636, 0.729]	0.682	[0.636, 0.729]
SARI	0.712	[0.657, 0.768]	0.712	[0.657, 0.768]
BERTScore	0.516	[0.454, 0.574]	0.592	[0.533, 0.648]
MeaningBERT	0.535	[0.478, 0.592]	0.560	[0.501, 0.617]
LENS	0.568	[0.508, 0.625]	0.620	[0.560, 0.675]
SummaC-ZS	0.987	[0.972, 0.998]	0.995	[0.987, 1.000]
AlignScore	0.984	[0.972, 0.995]	0.984	[0.972, 0.995]
NLI-fwd	0.966	[0.945, 0.984]	0.976	[0.953, 0.993]
BiNLI	0.921	[0.880, 0.957]	0.929	[0.894, 0.960]
BiNLI-core	0.914	[0.874, 0.949]	0.921	[0.888, 0.951]

Small: ±5% or ±1 (years ±1–3); large: ×2, 3, 5 or 10 (years ±25–80). Exploratory.

Table S5 Matched AUC by type of legitimate change (number of pairs in parentheses)

Metric	Legit.	NEG	NUM	ENT	ANT	QMOD	OMIT	INS
BLEU	REF	0.68 (1380)	0.68 (726)	0.68 (958)	0.66 (519)	0.65 (156)	0.63 (358)	0.87 (1432)
BLEU	PDEL	0.62 (39)	0.72 (36)	0.65 (34)	0.53 (15)	0.50 (6)	0.65 (17)	0.86 (42)
BLEU	SPLIT	0.67 (6)	1.00 (2)	1.00 (3)	0.80 (5)	–	0.83 (6)	1.00 (6)
SARI	REF	0.60 (1380)	0.72 (726)	0.72 (958)	0.62 (519)	0.57 (156)	0.78 (358)	0.55 (1432)
SARI	PDEL	0.46 (39)	0.61 (36)	0.76 (34)	0.53 (15)	0.67 (6)	0.94 (17)	0.40 (42)
SARI	SPLIT	0.67 (6)	1.00 (2)	1.00 (3)	1.00 (5)	–	0.83 (6)	0.67 (6)
BERTScore	REF	0.68 (1380)	0.57 (726)	0.76 (958)	0.60 (519)	0.58 (156)	0.78 (358)	0.88 (1432)
BERTScore	PDEL	0.08 (39)	0.11 (36)	0.59 (34)	0.07 (15)	0.00 (6)	0.82 (17)	0.83 (42)
BERTScore	SPLIT	1.00 (6)	1.00 (2)	1.00 (3)	0.40 (5)	–	1.00 (6)	1.00 (6)
MeaningBERT	REF	0.53 (1380)	0.56 (726)	0.74 (958)	0.50 (519)	0.57 (156)	0.69 (358)	0.80 (1432)
MeaningBERT	PDEL	0.26 (39)	0.28 (36)	0.71 (34)	0.13 (15)	0.17 (6)	0.88 (17)	0.74 (42)
MeaningBERT	SPLIT	0.67 (6)	1.00 (2)	1.00 (3)	0.60 (5)	–	1.00 (6)	1.00 (6)
LENS	REF	0.71 (1380)	0.59 (726)	0.68 (958)	0.65 (519)	0.68 (156)	0.45 (358)	0.76 (1432)
LENS	PDEL	0.79 (39)	0.69 (36)	0.82 (34)	0.73 (15)	0.67 (6)	0.41 (17)	0.79 (42)
LENS	SPLIT	0.83 (6)	1.00 (2)	1.00 (3)	0.80 (5)	–	0.50 (6)	0.83 (6)
SummaC-ZS	REF	0.98 (1380)	0.99 (726)	0.95 (958)	0.94 (519)	0.94 (156)	0.49 (358)	0.95 (1432)
SummaC-ZS	PDEL	1.00 (39)	1.00 (36)	1.00 (34)	0.87 (15)	1.00 (6)	0.24 (17)	1.00 (42)
SummaC-ZS	SPLIT	1.00 (6)	1.00 (2)	1.00 (3)	1.00 (5)	–	0.17 (6)	1.00 (6)
AlignScore	REF	0.98 (1380)	0.98 (726)	0.97 (958)	0.94 (519)	0.92 (156)	0.42 (358)	0.94 (1432)
AlignScore	PDEL	1.00 (39)	1.00 (36)	1.00 (34)	0.93 (15)	1.00 (6)	0.24 (17)	1.00 (42)
AlignScore	SPLIT	1.00 (6)	1.00 (2)	1.00 (3)	1.00 (5)	–	0.50 (6)	1.00 (6)
NLI-fwd	REF	0.99 (1380)	0.97 (726)	0.98 (958)	0.95 (519)	0.96 (156)	0.50 (358)	0.99 (1432)
NLI-fwd	PDEL	1.00 (39)	1.00 (36)	1.00 (34)	1.00 (15)	1.00 (6)	0.71 (17)	1.00 (42)
NLI-fwd	SPLIT	1.00 (6)	0.50 (2)	1.00 (3)	1.00 (5)	–	0.33 (6)	1.00 (6)
BiNLI	REF	0.99 (1380)	0.93 (726)	0.96 (958)	0.90 (519)	0.92 (156)	0.86 (358)	0.91 (1432)

Metric	Legit.	NEG	NUM	ENT	ANT	QMOD	OMIT	INS
BiNLI	PDEL	0.97 (39)	0.81 (36)	0.91 (34)	0.73 (15)	1.00 (6)	0.94 (17)	0.81 (42)
BiNLI	SPLIT	1.00 (6)	1.00 (2)	1.00 (3)	1.00 (5)	–	0.83 (6)	1.00 (6)
BiNLI-core	REF	0.99 (1380)	0.92 (726)	0.96 (958)	0.91 (519)	0.96 (156)	0.75 (358)	0.91 (1432)
BiNLI-core	PDEL	0.97 (39)	0.92 (36)	0.97 (34)	0.87 (15)	1.00 (6)	0.82 (17)	0.93 (42)
BiNLI-core	SPLIT	1.00 (6)	1.00 (2)	1.00 (3)	1.00 (5)	–	0.83 (6)	1.00 (6)

Exploratory and descriptive; PDEL and SPLIT cells rest on few pairs; – no pairs.

Table S6 Omission versus peripheral deletion on the extension set (AUC_{OP})

Metric	AUC (OP)	95% CI	Omissions	Peripheral deletions
NLI-fwd	0.572	[0.528, 0.615]	839	242
BiNLI	0.639	[0.590, 0.687]	839	242
BiNLI-core	0.766	[0.724, 0.801]	839	242

D = AUC_{OP}(BiNLI) − AUC_{OP}(NLI-fwd) = 0.068 [0.014, 0.122]; change in insertion mAUC = −0.085 [−0.102, −0.067]. BiNLI-core is exploratory.

Table S7 External validity (H3, exploratory): Spearman correlation with released human meaning ratings

Metric	Mean mAUC	ρ ASSET	p	ρ Simplicity-DA	p
BLEU	0.699	0.591	$< .001$	0.556	$< .001$
SARI	0.665	0.612	$< .001$	0.493	$< .001$
BERTScore	0.704	0.865	$< .001$	0.767	$< .001$
MeaningBERT	0.635	0.701	$< .001$	0.854	$< .001$
LENS	0.642	0.675	$< .001$	0.651	$< .001$
SummaC-ZS	0.881	0.483	$< .001$	0.508	$< .001$
AlignScore	0.874	0.424	$< .001$	0.537	$< .001$
NLI-fwd	0.899	0.454	$< .001$	0.562	$< .001$
BiNLI	0.921	0.683	$< .001$	0.635	$< .001$
BiNLI-core	0.906	0.670	$< .001$	0.652	$< .001$

ASSET ratings: 100 system outputs; Simplicity-DA: 600 outputs. Mean mAUC over the eligible error types. Rank agreement between mean mAUC and rating correlation across 10 scores: $\rho = -0.26$ (p = .467) for ASSET and $\rho = -0.13$ (p = .726) for Simplicity-DA.

Table S8 GEE logistic models of correct preference: coefficient of the perplexity difference

Metric	Coefficient	p	Pairs
BLEU	+0.595	$< .001$	5,576
SARI	+0.684	$< .001$	5,583
BERTScore	+0.677	$< .001$	5,584
MeaningBERT	+0.351	$< .001$	5,582
LENS	+0.455	$< .001$	5,584
SummaC-ZS	+0.045	.610	5,584
AlignScore	+0.303	.010	5,584
NLI-fwd	+0.204	.033	5,584
BiNLI	+0.199	.023	5,584
BiNLI-core	+0.268	.002	5,584

One model per metric over all pairs of a legitimate change and an eligible meaning-altering edit of the same base (tied scores excluded). Outcome: the metric scores the legitimate change above the edit. Predictors: edit type (fixed

effect) and the difference between the log GPT-2 perplexity change of the edit and that of the legitimate change, each relative to the base. Sources as clusters. Exploratory.

Table S9 Matched AUC by tertile of the log GPT-2 perplexity change caused by the meaning-altering edit

Metric	Type	T1 (lowest)	T2	T3 (highest)
BLEU	NEG	0.690	0.686	0.618
BLEU	NUM	0.623	0.701	0.714
BLEU	ENT	0.646	0.655	0.703
BLEU	ANT	0.668	0.640	0.664
BLEU	OMIT	0.575	0.566	0.681
BLEU	INS	0.899	0.848	0.842
SARI	NEG	0.603	0.592	0.575
SARI	NUM	0.608	0.767	0.707
SARI	ENT	0.724	0.646	0.765
SARI	ANT	0.623	0.606	0.636
SARI	OMIT	0.736	0.842	0.798
SARI	INS	0.584	0.517	0.521
BERTScore	NEG	0.666	0.654	0.680
BERTScore	NUM	0.491	0.532	0.714
BERTScore	ENT	0.769	0.642	0.801
BERTScore	ANT	0.549	0.537	0.640
BERTScore	OMIT	0.770	0.750	0.812
BERTScore	INS	0.914	0.834	0.875
MeaningBERT	NEG	0.528	0.508	0.585
MeaningBERT	NUM	0.474	0.558	0.629
MeaningBERT	ENT	0.776	0.677	0.755
MeaningBERT	ANT	0.451	0.485	0.528
MeaningBERT	OMIT	0.747	0.579	0.725
MeaningBERT	INS	0.787	0.741	0.829
LENS	NEG	0.664	0.717	0.835
LENS	NUM	0.509	0.614	0.664
LENS	ENT	0.664	0.646	0.709
LENS	ANT	0.566	0.695	0.673
LENS	OMIT	0.345	0.487	0.477
LENS	INS	0.770	0.766	0.747
SummaC-ZS	NEG	0.988	0.989	0.965
SummaC-ZS	NUM	0.995	0.995	0.971
SummaC-ZS	ENT	0.948	0.965	0.944
SummaC-ZS	ANT	0.967	0.936	0.935
SummaC-ZS	OMIT	0.368	0.487	0.509
SummaC-ZS	INS	0.933	0.948	0.962
AlignScore	NEG	0.981	0.978	0.990
AlignScore	NUM	0.991	0.985	0.971
AlignScore	ENT	0.993	0.961	0.970
AlignScore	ANT	0.926	0.961	0.935
AlignScore	OMIT	0.310	0.421	0.454
AlignScore	INS	0.940	0.938	0.957
NLI-fwd	NEG	0.993	0.991	1.000
NLI-fwd	NUM	0.981	0.973	0.950
NLI-fwd	ENT	0.993	0.978	0.980
NLI-fwd	ANT	0.959	0.966	0.939
NLI-fwd	OMIT	0.471	0.461	0.537
NLI-fwd	INS	0.990	0.979	0.990
BiNLI	NEG	0.985	0.984	0.995
BiNLI	NUM	0.948	0.922	0.900
BiNLI	ENT	0.970	0.926	0.958
BiNLI	ANT	0.910	0.892	0.893
BiNLI	OMIT	0.828	0.829	0.885
BiNLI	INS	0.904	0.879	0.913
BiNLI-core	NEG	0.986	0.988	0.995
BiNLI-core	NUM	0.915	0.925	0.900
BiNLI-core	ENT	0.981	0.934	0.954
BiNLI-core	ANT	0.902	0.921	0.902

Metric	Type	T1 (lowest)	T2	T3 (highest)
BiNLI-core	OMIT	0.713	0.724	0.784
BiNLI-core	INS	0.912	0.872	0.921

Log-perplexity change = log PPL(edit) − log PPL(base); tertiles over all meaning-altering edits, cut points −0.018 and 0.239. Exploratory.

Table S10 Validity diagnostic: labels assigned by RoBERTa-large-MNLI to a stratified 25% sample

Type	Checked	Expected pattern	Source ⇒ variant	Variant ⇒ source
NEG	173	0.971	con 166, ent 5, neu 2	con 162, ent 2, neu 9
NUM	91	0.802	con 71, ent 18, neu 2	con 62, ent 14, neu 15
ENT	120	0.775	con 88, ent 27, neu 5	con 68, ent 20, neu 32
ANT	65	0.862	con 54, ent 9, neu 2	con 42, ent 9, neu 14
QMOD	20	0.850	con 17, ent 3	con 12, ent 5, neu 3
OMIT	45	1.000	ent 45	con 1, neu 44
INS	180	0.844	con 5, ent 28, neu 147	con 4, ent 121, neu 55
REF	358	0.975	con 7, ent 349, neu 2	con 6, ent 240, neu 112
PDEL	10	1.000	ent 10	ent 2, neu 8
SPLIT	2	1.000	ent 2	neu 2

con = contradiction, ent = entailment, neu = neutral. Expected pattern: the source does not entail substitution and insertion edits; omission edits do not entail the source; the source entails legitimate changes. The checker is itself an NLI model; no human validation was performed.

S6 Supplementary figures

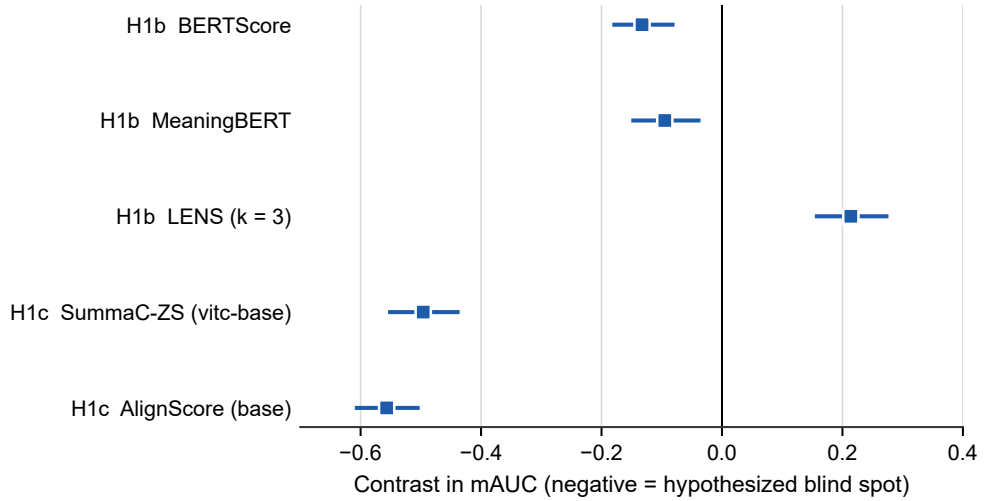


Fig. S1 Confirmatory contrasts C_b (H1b) and C_c (H1c) with 95% cluster-bootstrap confidence intervals. Negative values indicate the hypothesized blind spot

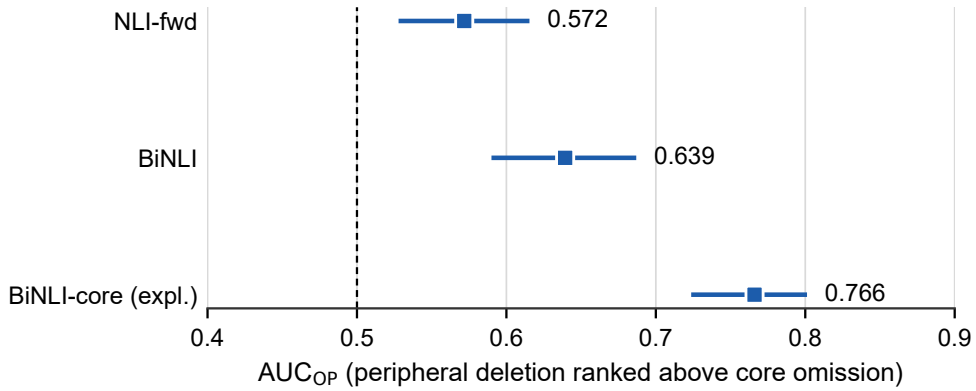


Fig. S2 AUC_{OP} (probability that a peripheral deletion is scored above the omission of a main assertion) on the extension set, with 95% confidence intervals. BiNLI-core is exploratory