LLM as a Meta-Judge: Synthetic Data for NLP Evaluation Metric Validation

Lukáš Eigler[1,2] **and Jindřich Libovický**[1] **and David Hurych**[2]
[1]Faculty of Mathematics and Physics, Charles University, Czech Republic
[2]valeo.ai
eiglerlukas@gmail.com libovicky@ufal.mff.cuni.cz david.hurych@gmail.com

Abstract

Validating evaluation metrics for NLG typically relies on expensive and time-consuming human annotations, which predominantly exist only for English datasets. We propose *LLM as a Meta-Judge*, a scalable framework that utilizes LLMs to generate synthetic evaluation datasets via controlled semantic degradation of real data, replacing human judgment. We validate our approach using *meta-correlation*, measuring the alignment between metric rankings derived from synthetic data and those from standard human benchmarks. Experiments across Machine Translation, Question Answering, and Summarization demonstrate that synthetic validation serves as a reliable proxy for human judgment, achieving meta-correlations exceeding 0.9 in multilingual QA, and is a viable alternative when human judgments are unavailable or too expensive to obtain. Our code and data are publicly available at https://github.com/eiglerl/meta-judge.

1 Introduction

Evaluating natural language generation (NLG) is challenging because semantically equivalent content may have many valid surface realizations. The standard approach to validating evaluation metrics requires collecting human judgments on system outputs and computing correlation with metric scores (Graham et al., 2013). This creates a bottleneck as human-annotated datasets are expensive, predominantly English-only, and require renewal as systems evolve.

Few public datasets with human judgment exist: WMT for translation (Callison-Burch et al., 2008), RoSE for summarization (Liu et al., 2023b), and MOCHA for question answering (Chen et al., 2020). These remain largely English-centric, and metrics validated on one task or language do not necessarily transfer to other tasks or languages.

We propose replacing human judgment with LLM-generated outputs of controlled quality. Our approach prompts an LLM to produce semantically degraded versions of reference texts at specified severity levels, creating synthetic data where quality ordering is known by construction (Figure 1).

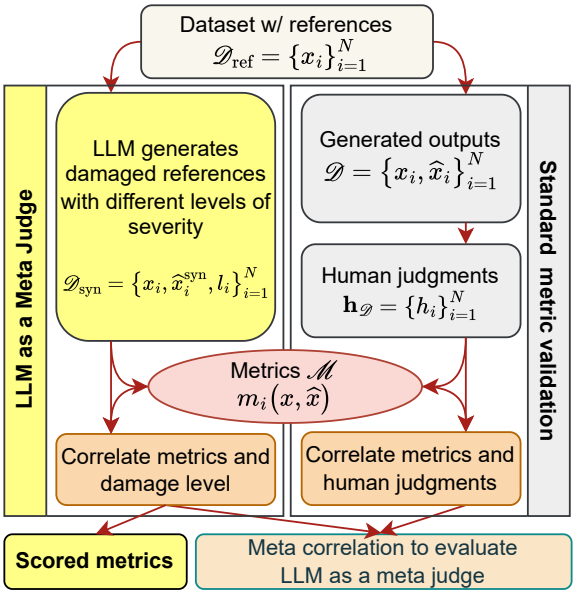

Figure 1: LLM as a Meta-Judge contrasted with standard metric validation with human judgment: LLMs generate damaged reference sentences, and we validate the NLG metrics by correlation of the damage level with metric values. We validate our protocol via meta-correlation, i.e., correlation with the standard metric validation.

We make the following contributions: (1) *Meta-Judge*, a protocol for validating NLG metrics without human judgment, using LLM-generated text with controlled degradation of real reference text as a proxy for system outputs. (2) *meta-correlation*: correlation between metric rankings on synthetic data and on human-annotated benchmarks, measuring proxy reliability. (3) Empirical validation across machine translation, question answering, and summarization in multiple languages, including low-resource languages.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1417–1435

2 Evaluation Metric Validation

Evaluation metrics assess generated text quality. We define a metric as a function $m(x, \hat{x})$ computing a real number for reference x and generated string $\hat{x}$. NLG metrics fall into several paradigms.

String overlap metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), compute string-level overlap with interpretability and efficiency, but limited semantic sensitivity.

Embedding-based metrics leverage contextual representations: BERTScore (Zhang et al., 2020) computes token-wise similarity via BERT embeddings, while YiSi (Lo, 2019) extends this with semantic role labeling.

Learned metrics trained on human judgments achieve the best correlations: COMET (Rei et al., 2020) employs cross-lingual encoders, BLEURT (Sellam et al., 2020) leverages BERT pre-training with synthetic degradation. Recent approaches use LLMs as zero-shot evaluators (Zheng et al., 2023; Liu et al., 2023a), though this introduces circularity when evaluating LLM-generated text (Shen et al., 2023).

Since NLG tasks lack formal specifications, human judgment remains the gold standard. Let $\mathcal{M} = (m_1, \ldots, m_k)$ be evaluation metrics, $\mathcal{D} = \{x_i, \hat{x}_i\}_{i=1}^{N}$ a dataset with references and candidates, and $\mathbf{h}_\mathcal{D}$ human judgments. For metric m_i, we define scores as a vector $\mathbf{s}_\mathcal{D}^{m_i} = [m_i(x, \hat{x}) : (x, \hat{x}) \in \mathcal{D}]$. Metric quality is measured as correlation with human judgment:

$$r_{\text{hum}}^{i} = \rho(\mathbf{s}_\mathcal{D}^{m_i}, \mathbf{h}_\mathcal{D}) \tag{1}$$

where ρ is Spearman rank correlation.

WMT has systematically evaluated translation metrics since 2008 (Callison-Burch et al., 2008), evolving from pairwise comparisons to direct assessment (Graham et al., 2013) and Multidimensional Quality Metrics (Freitag et al., 2021). Datasets exist for summarization (RoSE, 22k annotations; Liu et al., 2023b) and QA (MOCHA, 40k judgments; Chen et al., 2020), though English-only. CUS-QA (Libovický et al., 2025) is a rare multilingual exception. Correlations vary from $r \approx 0.4$ for lexical metrics on summarization to $r > 0.9$ for learned metrics on high-resource translation (Freitag et al., 2022).

3 LLM as a Meta-Judge

We propose a framework for validating evaluation metrics using an LLM, which we call *Meta-Judge*. Unlike *LLM-as-a-Judge*, where an LLM scores outputs as an evaluation metric, the LLM in the Meta-Judge framework is solely used for data generation to validate evaluation metrics. The protocol is defined as a function $\mathcal{D} \times \mathcal{M} \to \mathcal{S}$, where we use only reference texts x from dataset $\mathcal{D}$ in this case, $\mathcal{M}$ is a set of evaluation metrics, and $\mathcal{S}$ are metric validation scores.

The protocol consists of three steps: (1) We prompt the LLM to semantically degrade references at known severity levels (damage levels), creating a synthetic dataset with references, damaged references, and pseudo-labels describing the damage. (2) Metrics compute scores for each damaged reference. (3) Correlation between metric scores and pseudo-labels provides metric validation.

Standard metric validation, by contrast, requires candidate outputs and human judgments: $\mathcal{D} \times \mathcal{H} \times \mathcal{M} \to \mathcal{S}$, where $\mathcal{D}$ contains references and candidates, $\mathcal{H}$ are human judgments, and validation is obtained by correlating metric scores with human judgments.

3.1 Metric Validation Using Meta-Judge

The primary application of our approach is to assess metric quality without human judgments. We generate a synthetic dataset $\mathcal{D}_{\text{syn}}$ using an LLM and an existing dataset. The LLM receives a reference text (and optional context) x, along with a specification of damage types through the prompt. We instruct the LLM to generate synthetic text $\hat{x}^{\text{syn}}$ corresponding to a damage level $l \in \{0, \ldots, L_{\max}\}$. These levels serve as pseudo-labels, representing monotonic quality degradation from paraphrasing ($l = 0$) to severe hallucinations ($l = L_{\max}$).

The synthetic dataset is $\mathcal{D}_{\text{syn}} = \{x_i, \hat{x}_i^{\text{syn}}, l_i\}_{i=1}^{N}$, where x_i denotes input, optional context, and reference, $\hat{x}_i^{\text{syn}}$ is the damaged text, and l_i is the damage level. Thus $\hat{x}_i^{\text{syn}}$ acts as generated output and l_i as pseudo-label replacing human judgment. Collected pseudo-labels are $\mathbf{p}_{\mathcal{D}_{\text{syn}}} = [l : (x, \hat{x}^{\text{syn}}, l) \in \mathcal{D}_{\text{syn}}]$.

To validate metric m_i on $\mathcal{D}_{\text{syn}}$, we compute segment-level scores $\mathbf{s}_{\mathcal{D}_{\text{syn}}}^{m_i}$ for all damaged texts. Since metrics measure quality (higher is better) while damage levels measure error (higher is worse), we negate the pseudo-labels:

$$r_{\text{syn}}^{i} = \rho(\mathbf{s}_{\mathcal{D}_{\text{syn}}}^{m_i}, -\mathbf{p}_{\mathcal{D}_{\text{syn}}}) \tag{2}$$

3.2 Meta-Correlation Analysis

For the synthetic dataset to serve as a useful proxy for human judgments, the synthetic correlations

r_{syn} must accurately estimate the metric's performance. We call this second-order correlation *meta-correlation*.

The validation proceeds in three steps:

(1) Human correlations r_{hum}. We compute metric scores $\mathbf{s}_{\mathcal{D}}^{m_i}$ for each metric $m_i \in \mathcal{M}$ on dataset $\mathcal{D}$ with human judgments $\mathbf{h}_{\mathcal{D}}$, then compute Spearman correlation using Equation 1.

(2) Synthetic correlations r_{syn}. We compute metric scores $\mathbf{s}_{\mathcal{D}_{\text{syn}}}^{m_i}$ on the synthetic dataset and correlate with negative pseudo-labels $\mathbf{p}_{\mathcal{D}_{\text{syn}}}$ using Equation 2.

(3) Meta-correlation *MC*. We compute Spearman correlation between the vector of human correlations $\mathbf{r}_h = [r_{\text{hum}}^i : m_i \in \mathcal{M}]$ and synthetic correlations $\mathbf{r}_s = [r_{\text{syn}}^i : m_i \in \mathcal{M}]$:

$$MC = \rho(\mathbf{r}_h, \mathbf{r}_s) \tag{3}$$

A high positive meta-correlation indicates that the synthetic dataset is a reliable proxy for human judgment.

4 Experiments

4.1 Tasks

We evaluate our method across three tasks and multiple languages, including low-resource settings. All selected tasks have datasets with human judgments necessary for computing meta-correlation.

Question Answering. CUS-QA (Libovický et al., 2025) covers region-specific knowledge in Czech, Slovak, Ukrainian, and their English translations. MOCHA (Chen et al., 2020) tests various reasoning forms compiled from multiple sources, with judgments collected on a 1–5 scale and averaged across annotators.

Summarization. RoSE (Liu et al., 2023b) provides human judgments on news article summarization via Atomic Content Unit matching, aggregated over three annotators per summary.

Machine Translation. We use language pairs from WMT 2021 (Akhbardeh et al., 2021) (English (en) to Hausa (ha), Xhosa (xh) to Zulu (zu)), which uses Direct Assessment with z-normalized scores, and WMT 2024 (Kocmi et al., 2024) (Czech (cs) to Ukrainian (uk), English (en) to Czech (cs), English (en) to Icelandic (is)), which uses Error Span Annotation, covering both high-resource and low-resource settings.

4.2 Automatic Metrics

We use a diverse set of seven evaluation metrics spanning multiple paradigms: string overlap metrics (BLEU, ROUGE, chrF, METEOR), embedding-based metrics (BERTScore), and model-based approaches (COMET, BLEURT). To get a sufficient number of data points for robust correlation estimation, we evaluate each metric under multiple parameter configurations. Details of metric parameters are provided in Appendix A.

4.3 Synthetic Data Generation

To generate synthetic datasets containing semantically damaged texts, we use three open-source LLMs with varying parameter counts and architectures. We prompt the LLMs to produce semantically damaged outputs based on the given input and discrete damage level, utilizing greedy decoding to ensure deterministic generation. We use six damage levels ($L_{\max} = 5$, levels 0–5) as a design choice. While any range can be used, the prompt needs to describe every level of damage, and defining many distinct levels is difficult due to potential overlap and nuance. We compare three Meta-Judge models: *Llama 4 Scout* (Meta AI, 2025), *Llama 3.3 70B* (Team, 2024), and *Qwen 3 30B* (Yang et al., 2025).

To investigate the impact of in-context learning on the consistency of these pseudo-labels, we employ two prompting strategies: (1) *Zero-Shot:* The model relies solely on the instruction and damage definitions, and (2) *Few-Shot:* The model is provided with domain-specific examples of damage levels to better steer the degradation process. The complete prompts for all tasks under both prompting conditions are provided in Appendix B. The few-shot examples were manually prepared according to the damage level specifications.

4.4 Results

Table 1 reports meta-correlation results across all datasets and tasks using Spearman rank correlation (Kendall correlation in Table 3; detailed per-metric correlations in Tables 4–7 in the Appendix).

The strongest and most consistent results are observed in CUS-QA (question answering). The Meta-Judge protocol achieves high meta-correlation across all tested languages (Czech, Slovak, Ukrainian) and their English translations, with values exceeding 0.9 in several configurations. Performance is generally higher for original languages

Model	Shot	CUS-QA (en)			CUS-QA (orig.)			MOCHA	RoSE	WMT 21		WMT 24		
		cs	sk	uk	cs	sk	uk			en-ha	xh-zu	cs-uk	en-cs	en-is
Llama 4 Scout	Few	.859	.829	.731	.895	.913	.716	.302	.473	.543	.523	.945	.276	.299
	Zero	.793	.774	.669	.827	.875	.717	.211	.453	.303	.473	.920	.472	.413
Llama 3.3 70B	Few	.808	.788	.751	.917	.922	.759	.678	.043	.490	.454	.854	.410	.419
	Zero	.621	.652	.583	.777	.774	.605	.486	.327	.432	.447	.794	.491	.513
Qwen 3 30B	Few	.796	.792	.651	.871	.885	.684	.726	.833	.334	.285	.918	.325	.370
	Zero	.917	.776	.754	.956	.955	.821	.872	.675	.032	.308	.937	.286	.269

Table 1: Meta-correlation between LLM as Meta-Judge and standard metric validation with human judgment measured by Spearman correlation. The values are black if they are significant at the 0.01 confidence, dark gray if they are only significant at 0.05 confidence, and light gray if they are not significantly different from zero.

Model	Shot	Spearman	Kendall
Llama 4 Scout	Few	$.905 \pm .010$	$.756 \pm .020$
	Zero	$.884 \pm .056$	$.710 \pm .087$
Llama 3.3 70B	Few	$.914 \pm .019$	$.750 \pm .024$
	Zero	$.862 \pm .051$	$.733 \pm .046$
Qwen 3 30B	Few	$.890 \pm .029$	$.731 \pm .049$
	Zero	$.958 \pm .022$	$.827 \pm .035$

Table 2: Mean $\pm$ standard deviation of the meta correlation for the Czech subset of CUS-QA with several prompts.

than translations. Results on MOCHA are similarly strong, with Qwen 3 reaching 0.87 in zero-shot mode, though Llama 4 Scout below other models.

Note that few-shot results are not always better than zero-shot. This was revealed by multiple benchmarks on the internet, but, as far as we know, no scientific work studied this problem thoroughly for our tasks. Similar behavior of Qwen and Llama models was observed for the chain-of-thought prompting by Cheng et al. (2025).

Meta-correlation results for RoSE (summarization) and WMT (machine translation) are more variable. WMT 2024 Czech–Ukrainian achieves high meta-correlation across all models, but this is partially an artifact: the default ROUGE tokenizer discards Cyrillic characters, causing ROUGE to fail on Ukrainian text and produce clearly poor metrics that make the overall ranking easier. For other language pairs, we attribute variability to differences in system output variance: English–Czech is a long-standing WMT task where systems consistently achieve high performance, making ranking difficult, while lower-resource pairs (Icelandic, Zulu, Hausa) exhibit greater output variance that facilitates metric discrimination. Compared to Llama models, Qwen performs notably worse on low-resource WMT 2021 languages.

Per-metric analysis (Tables 4–7) reveals consistent patterns: BLEU shows low or negative correlations with both human judgments and synthetic damage levels, with the correlation getting lower as the n-gram order increases. In contrast, chrF performs reliably across all configurations, often matching or exceeding learned metrics, suggesting character-level overlap captures semantic degradation more robustly than word-level n-grams.

Language-specific analysis shows substantially lower correlations for Ukrainian text in CUS-QA, partially due to tokenization issues. ROUGE-4 produces undefined values for Cyrillic entirely. English translations exhibit lower correlations with human judgment than the original languages, yet synthetic correlations remain stable, indicating robustness to translation-induced noise.

4.5 Prompt Robustness

To assess sensitivity to the precise wording of damage instructions, we tested five prompt variants per model on the Czech (original) CUS-QA dataset split. Table 2 reports mean and standard deviation of the meta-correlation. Standard deviations remain below 0.06 in all cases with Spearman correlation. This indicates that the Meta-Judge protocol is robust to the exact formulation of the damage descriptions.

5 Related Work

Synthetic data for NLG evaluation has been explored with different objectives. Sellam et al. (2020) used random perturbations (mask-filling, backtranslation, word dropout) to generate training data for BLEURT, focusing on metric *training* rather than *validation*. More closely related, Deviyani and Diaz (2025) introduced local metric accuracy using rule-based and LLM-based perturbations to analyze how metric performance varies

across evaluation contexts. Our Meta-Judge framework differs by validating whether synthetic degradation can serve as a *universal proxy* for human evaluation across tasks and languages.

The meta-correlation concept has been applied by Shen et al. (2023), who examined how LLM-metric agreement with human judgment degrades for higher-quality outputs. While they assess single-evaluator reliability across quality levels, our meta-correlation validates whether synthetic data preserves the relative ranking of *multiple metrics*, enabling metric validation without human labels.

6 Conclusions

We introduced LLM as a Meta-Judge, a protocol for validating NLG evaluation metrics using LLM-generated synthetic data with controlled semantic degradation, eliminating the need for expensive human annotation. Our meta-correlation analysis demonstrates that synthetic evaluation can serve as a reliable proxy for human judgment, particularly for question-answering tasks, where we achieve meta-correlations exceeding 0.9. The approach proves most effective in high-resource languages and shows promise for low-resource settings where human-annotated evaluation data is scarce or unavailable. The results vary across tasks, with stronger performance on QA compared to summarization and machine translation. The protocol provides a scalable alternative for metric validation when human evaluation is impractical.

Limitations

The reliability of synthetic data generation depends on the LLM's proficiency in the target language. For low-resource languages, the quality of semantic degradations may be inconsistent, as reflected in our lower meta-correlations for Hausa, Zulu, and Xhosa.

Our method requires specifying what types of errors the metrics should detect. The damage definitions in our prompts are task-specific, and applying the framework to new generation tasks requires designing appropriate degradation strategies based on domain knowledge.

A further concern is circularity, where the model used for data generation and the evaluation metric(s) share architecture and/or training data. We consider circularity to be a critical issue primarily when the LLM used in LLM-as-a-Judge matches the generator LLM, which is not the case in our setup.

Finally, validating the Meta-Judge approach requires datasets with human judgments to compute meta-correlation. For new tasks or languages without existing human-annotated evaluation data, limited pilot annotations may still be necessary to verify the method's reliability.

Acknowledgement

This research was supported by the Charles University project PRIMUS/23/SCI/023 and project CZ.02.01.01/00/23_020/0008518 of the Ministry of Education, Youth and Sports of the Czech Republic. Computational resources were provided by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

We used GitHub Copilot to assist in writing the source code for our experiments. We used Claude and Gemini for formatting tables, spell checking, and text shortening.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 2021. Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Xiang Cheng, Chengyan Pan, Minjun Zhao, Deyang Li, Fangchao Liu, Xinyu Zhang, Xiao Zhang, and Yong Liu. 2025. Revisiting Chain-of-thought Prompting: Zero-shot Can Be Stronger than Few-shot. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13533–13554, Suzhou, China. Association for Computational Linguistics.

Athiya Deviyani and Fernando Diaz. 2025. Contextual metric meta-evaluation by measuring local metric accuracy. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4906–4925, Albuquerque, New Mexico. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Jindrich Libovický, Jindrich Helcl, Andrei Manea, and Gianluca Vico. 2025. CUS-QA: Local-Knowledge-oriented Open-ended Question Answering Dataset. *CoRR*, abs/2507.22752.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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.

Chi-kiu Lo. 2019. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Meta AI. 2025. The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Llama Team. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 Technical Report. *CoRR*, abs/2505.09388.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020,*

Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Metric Parameters

We use four string overlap metrics: BLEU with smoothing and n-gram orders 1–4, chrF with character order $c \in \{4, 6\}$ and word order $w \in \{0, 2\}$, ROUGE-1, -2, -4, and -L, and METEOR with $\alpha \in \{0.2, 0.9\}$ and $\gamma \in \{0, 0.5\}$.

For embedding-based metrics, we compute BERTScore using Czech and English models with limited (5 layers) and full depth.

For model-based metrics, we use four COMET models (`Unbabel/wmt22-comet-da`, `eamt22-cometinho-da`, `Unbabel/wmt20-comet-da`, `Unbabel/wmt20-comet-qe-da`) and four BLEURT checkpoints (`bleurt-tiny-128`, `bleurt-base-512`, `bleurt-large-512`, `BLEURT-20-D12`).

B Prompts for Synthetic Data Generation

Tables 8–15 present the prompts used to generate synthetic data. We provide zero-shot and few-shot variants for each task: CUS-QA, MOCHA, RoSE, and machine translation. All prompts define six damage levels (0–5) with task-specific degradation strategies.

C Additional Results

Table 3 reports meta-correlation results using Kendall rank correlation, complementing the Spearman results in Table 1; patterns are consistent across both measures.

Tables 4–7 present detailed Spearman correlations between metric scores and either human judgments (column Hum) or synthetic damage levels for each Meta-Judge model under few-shot (F) and zero-shot (Z) prompting. These tables allow examination of individual metric behavior across datasets and languages.

Table 4 covers CUS-QA Czech, Table 5 covers CUS-QA Slovak and Ukrainian, Table 6 presents MOCHA, RoSE, and WMT 2021, and Table 7 reports WMT 2024 language pairs.

D Damage Specification Robustness

Tables 16–20 show the five damage specification variants used in the prompt robustness testing in Section 4.5. Apart from the damage specification, they are same as in Table 8 and Table 10

Model	Shot	CUS-QA (en)			CUS-QA (orig.)			MOCHA	RoSE	WMT 21		WMT 24		
		cs	sk	uk	cs	sk	uk			en-ha	xh-zu	cs-uk	en-cs	en-is
Llama 4 Scout	Few	.730	.661	.492	.751	.735	.533	.249	.365	.312	.265	.825	.148	.233
	Zero	.656	.624	.450	.688	.704	.573	.138	.365	.201	.318	.762	.318	.302
Llama 3.3 70B	Few	.619	.598	.524	.757	.730	.584	.487	.064	.333	.323	.656	.238	.286
	Zero	.524	.519	.355	.672	.656	.474	.265	.280	.386	.323	.656	.307	.360
Qwen 3 30B	Few	.651	.619	.429	.693	.709	.516	.571	.635	.032	-.064	.767	.153	.249
	Zero	.725	.513	.577	.804	.788	.681	.720	.513	-.021	.111	.810	.085	.201

Table 3: Meta-correlation between LLM as Meta-Judge and standard metric validation with human judgment measured by *Kendall correlation*. The values are black if they are significant at the 0.01 confidence, dark gray if they are only significant at 0.05 confidence, and light gray if they are not significantly different from zero.

Metric	Parameters	**CUS-QA cs (en)**							**CUS-QA cs (orig.)**						
		Hum	Llama 4		Llama 3.3		Qwen 3		GT	Llama 4		Llama 3.3		Qwen 3	
			F	Z	F	Z	F	Z		F	Z	F	Z	F	Z
BLEU	Order 1	.512	.132	-.035	.049	-.154	.048	.080	.478	.091	-.061	.099	-.103	.007	.134
	Order 2	.493	.112	-.088	.026	-.229	.022	.040	.451	.073	-.115	.095	-.183	-.006	.109
	Order 3	.467	.064	-.141	-.012	-.285	-.018	.018	.394	.003	-.193	.052	-.255	-.054	.060
	Order 4	.447	.025	-.173	-.044	-.318	-.048	.009	.343	-.052	-.238	.015	-.295	-.083	.029
chrF	c: 4, w: 0	.617	.487	.429	.420	.352	.425	.363	.655	.505	.433	.452	.418	.436	.488
	c: 4, w: 2	.626	.495	.427	.409	.338	.422	.351	.667	.515	.441	.453	.417	.446	.498
	c: 6, w: 0	.624	.510	.452	.444	.372	.451	.374	.665	.522	.452	.467	.437	.458	.502
	c: 6, w: 2	.630	.512	.445	.430	.358	.445	.364	.672	.527	.455	.466	.432	.462	.507
ROUGE	ROUGE-1	.615	.390	.338	.328	.241	.321	.258	.626	.385	.295	.334	.275	.280	.361
	ROUGE-2	.476	.405	.362	.336	.283	.352	.232	.580	.415	.348	.339	.318	.324	.360
	ROUGE-4	.259	.230	.204	.173	.120	.182	.070	.377	.267	.241	.215	.191	.211	.227
	ROUGE-L	.619	.385	.334	.319	.233	.319	.256	.629	.387	.294	.330	.270	.281	.361
METEOR	α: 0.2, γ: 0.0	.552	.247	.179	.186	.129	.196	.233	.553	.227	.146	.187	.139	.150	.270
	α: 0.2, γ: 0.5	.564	.303	.232	.237	.168	.246	.225	.576	.309	.225	.255	.200	.227	.314
	α: 0.9, γ: 0.0	.593	.462	.408	.394	.329	.429	.274	.621	.491	.423	.411	.383	.442	.464
	α: 0.9, γ: 0.5	.580	.459	.398	.393	.325	.423	.249	.616	.492	.423	.418	.391	.443	.453
BERTSc.	cs, Not Limited	.590	.342	.243	.271	.127	.270	.273	.609	.354	.236	.315	.180	.257	.342
	cs, Limited (L5)	.509	.172	.030	.130	-.070	.110	.156	.520	.166	.020	.174	-.028	.081	.191
	en, Not Limited	.598	.481	.422	.439	.368	.425	.311	.613	.398	.324	.360	.289	.319	.384
	en, Limited (L5)	.547	.338	.298	.275	.215	.294	.226	.532	.264	.207	.243	.144	.195	.281
COMET	wmt20-da	.481	.077	-.005	.026	-.117	.045	.198	.534	.150	.059	.167	.047	.108	.230
	wmt20-qe-da	.013	.089	.184	.182	.374	.133	.018	.102	.170	.275	.148	.372	.207	.125
	wmt22-da	.530	.115	.035	.087	-.063	.092	.221	.537	.195	.129	.203	.129	.153	.271
	cometinho-da	.539	.138	.053	.087	-.047	.107	.189	.526	.165	.080	.180	.060	.110	.229
BLEURT	20-D12	.595	.365	.272	.319	.183	.287	.339	.626	.370	.269	.356	.270	.310	.387
	base-512	.474	.119	.027	.056	-.062	.053	.081	.358	-.044	-.102	-.045	-.150	-.072	.098
	large-512	.536	.292	.180	.277	.152	.214	.114	.442	.057	-.049	.084	-.061	.015	.131
	tiny-128	.608	.436	.317	.376	.205	.350	.251	.615	.407	.260	.386	.187	.328	.373

Table 4: Spearman Correlations across different metrics and model variations for CUS-QA cs (en) and CUS-QA cs (orig.).

Metric	Parameters	CUS-QA sk (en) Hum	Llama 4 F	Llama 4 Z	Llama 3.3 F	Llama 3.3 Z	Qwen 3 F	Qwen 3 Z	CUS-QA sk (orig.) GT	Llama 4 F	Llama 4 Z	Llama 3.3 F	Llama 3.3 Z	Qwen 3 F	Qwen 3 Z
BLEU	Order 1	.459	.123	-.020	.028	-.134	.019	.037	.407	.071	-.026	.052	-.151	.004	.126
	Order 2	.430	.098	-.083	.010	-.208	.004	.026	.378	.053	-.079	.046	-.219	-.017	.093
	Order 3	.408	.037	-.150	-.036	-.269	-.045	.004	.331	-.021	-.158	-.004	-.287	-.074	.037
	Order 4	.390	-.009	-.193	-.076	-.300	-.077	-.010	.290	-.077	-.207	-.044	-.325	-.110	.005
chrF	c: 4, w: 0	.565	.476	.419	.417	.355	.446	.395	.575	.494	.453	.415	.380	.474	.487
	c: 4, w: 2	.583	.494	.430	.422	.352	.449	.368	.602	.517	.472	.425	.377	.491	.505
	c: 6, w: 0	.577	.494	.438	.437	.374	.468	.398	.591	.510	.470	.426	.391	.491	.500
	c: 6, w: 2	.588	.506	.443	.438	.368	.468	.379	.608	.525	.483	.434	.387	.503	.514
ROUGE	ROUGE-1	.596	.402	.357	.301	.220	.312	.205	.540	.372	.330	.288	.207	.315	.360
	ROUGE-2	.454	.414	.364	.345	.284	.356	.216	.502	.410	.379	.313	.268	.350	.353
	ROUGE-4	.184	.178	.159	.135	.112	.152	.061	.253	.213	.203	.150	.124	.191	.197
	ROUGE-L	.597	.399	.353	.294	.211	.309	.205	.543	.372	.330	.286	.200	.315	.357
METEOR	α: 0.2, γ: 0.0	.536	.244	.173	.161	.108	.159	.140	.468	.210	.157	.153	.053	.143	.253
	α: 0.2, γ: 0.5	.546	.320	.248	.240	.162	.241	.152	.496	.319	.265	.234	.137	.256	.309
	α: 0.9, γ: 0.0	.585	.511	.466	.433	.367	.488	.239	.547	.508	.469	.389	.333	.517	.501
	α: 0.9, γ: 0.5	.572	.500	.448	.431	.356	.468	.217	.541	.502	.462	.392	.335	.500	.466
BERTSc.	cs, Not Limited	.548	.322	.219	.250	.128	.249	.232	.523	.324	.262	.273	.126	.284	.332
	cs, Limited (L5)	.460	.147	.020	.103	-.056	.078	.125	.422	.140	.051	.133	-.080	.086	.177
	en, Not Limited	.582	.479	.435	.450	.398	.470	.349	.561	.410	.368	.355	.274	.382	.426
	en, Limited (L5)	.541	.336	.305	.291	.238	.328	.235	.461	.241	.199	.191	.077	.227	.298
COMET	wmt20-da	.438	.060	-.038	.027	-.104	.051	.194	.432	.132	.092	.159	.018	.124	.208
	wmt20-qe-da	.040	.087	.197	.198	.331	.138	.047	.100	.172	.238	.182	.365	.213	.137
	wmt22-da	.486	.096	-.002	.077	-.066	.080	.219	.441	.174	.135	.180	.078	.169	.255
	cometinho-da	.521	.116	.033	.083	-.045	.101	.190	.444	.153	.116	.159	.038	.124	.219
BLEURT	20-D12	.569	.335	.229	.305	.193	.271	.329	.538	.294	.240	.303	.174	.286	.331
	base-512	.379	.078	-.036	.036	-.053	.029	.034	.250	-.060	-.075	-.082	-.173	-.087	.081
	large-512	.477	.193	.094	.218	.108	.173	.074	.342	.014	-.025	.045	-.085	.006	.122
	tiny-128	.601	.460	.352	.384	.232	.386	.237	.559	.404	.329	.346	.160	.347	.372

Metric	Parameters	CUS-QA uk (en) Hum	Llama 4 F	Llama 4 Z	Llama 3.3 F	Llama 3.3 Z	Qwen 3 F	Qwen 3 Z	CUS-QA uk (orig.) GT	Llama 4 F	Llama 4 Z	Llama 3.3 F	Llama 3.3 Z	Qwen 3 F	Qwen 3 Z
BLEU	Order 1	.426	.093	-.034	.115	-.145	.008	.041	.412	.006	-.019	.011	-.134	-.035	.085
	Order 2	.433	.066	-.086	.110	-.199	-.009	.013	.393	-.044	-.083	-.025	-.203	-.072	.041
	Order 3	.432	.024	-.133	.084	-.238	-.045	-.005	.365	-.105	-.137	-.064	-.260	-.124	-.011
	Order 4	.431	-.006	-.158	.059	-.260	-.074	-.014	.342	-.145	-.169	-.095	-.297	-.162	-.048
chrF	c: 4, w: 0	.514	.425	.372	.381	.295	.374	.302	.591	.429	.392	.364	.331	.385	.426
	c: 4, w: 2	.519	.427	.373	.376	.281	.374	.287	.601	.444	.406	.362	.333	.393	.440
	c: 6, w: 0	.527	.448	.392	.400	.313	.402	.312	.601	.449	.407	.376	.345	.403	.443
	c: 6, w: 2	.526	.445	.388	.393	.299	.396	.299	.606	.456	.415	.372	.345	.406	.450
ROUGE	ROUGE-1	.460	.332	.283	.278	.168	.245	.198	.138	.169	.142	.114	.104	.109	.106
	ROUGE-2	.221	.276	.251	.208	.152	.246	.119	.128	.096	.081	.066	.061	.066	.060
	ROUGE-4*	.057	.100	.086	.039	.003	.084	.030	nan	nan	nan	nan	nan	nan	nan
	ROUGE-L	.477	.323	.272	.271	.159	.241	.199	.138	.169	.142	.114	.104	.109	.106
METEOR	α: 0.2, γ: 0.0	.399	.228	.185	.202	.101	.154	.182	.465	.197	.191	.149	.091	.135	.232
	α: 0.2, γ: 0.5	.398	.264	.214	.231	.134	.200	.159	.469	.252	.237	.185	.135	.194	.263
	α: 0.9, γ: 0.0	.465	.441	.389	.379	.285	.408	.253	.518	.415	.377	.317	.292	.396	.417
	α: 0.9, γ: 0.5	.435	.411	.356	.357	.273	.391	.204	.500	.407	.367	.304	.281	.389	.391
BERTSc.	cs, Not Limited	.521	.330	.256	.317	.164	.267	.250	.578	.305	.282	.270	.181	.232	.300
	cs, Limited (L5)	.453	.187	.086	.203	.002	.121	.164	.486	.085	.073	.109	-.026	.038	.143
	en, Not Limited	.499	.428	.371	.413	.319	.385	.252	.431	.201	.215	.170	.135	.154	.231
	en, Limited (L5)	.416	.279	.241	.256	.167	.243	.159	.456	.179	.188	.159	.103	.127	.207
COMET	wmt20-da	.348	.091	.035	.121	-.048	.070	.179	.490	.129	.127	.136	.039	.104	.171
	wmt20-qe-da	.083	.086	.168	.085	.280	.140	.039	-.024	.223	.210	.189	.327	.247	.185
	wmt22-da	.386	.138	.079	.178	-.007	.104	.197	.481	.181	.178	.184	.111	.153	.223
	cometinho-da	.420	.167	.115	.177	.040	.145	.164	.507	.114	.110	.131	.030	.086	.157
BLEURT	20-D12	.507	.353	.295	.376	.243	.318	.314	.584	.308	.290	.315	.226	.261	.307
	base-512	.332	.062	.022	.122	-.020	.036	.058	.217	-.145	-.114	-.142	-.218	-.164	-.033
	large-512	.394	.193	.147	.284	.139	.190	.093	.291	-.121	-.111	-.102	-.190	-.140	-.027
	tiny-128	.507	.384	.294	.377	.220	.337	.212	.347	.053	.035	.051	-.047	.047	.085

Table 5: Spearman Correlations across different metrics and model variations for CUS-QA sk (en) and CUS-QA sk (orig.) above and CUS-QA uk (en) and CUS-QA uk (orig.) below. The row marked by * contains nan values for the Ukrainian text because the default ROUGE implementation uses a tokenizer that skips most Cyrillic characters.

Metric	Parameters	MOCHA Validation							RoSE CNNDM Test						
		Hum	Llama 4		Llama 3.3		Qwen 3		GT	Llama 4		Llama 3.3		Qwen 3	
			F	Z	F	Z	F	Z		F	Z	F	Z	F	Z
BLEU	Order 1	.362	.319	.387	.255	.333	.218	.411	.472	.409	.331	.257	.396	.556	.658
	Order 2	.345	.338	.408	.266	.350	.208	.393	.534	.429	.354	.237	.413	.595	.709
	Order 3	.312	.334	.402	.254	.349	.195	.386	.529	.429	.358	.222	.410	.599	.719
	Order 4	.277	.320	.389	.238	.345	.183	.381	.514	.426	.358	.209	.404	.593	.718
chrF	c: 4, w: 0	.527	.300	.359	.246	.286	.313	.402	.691	.511	.421	.300	.420	.619	.683
	c: 4, w: 2	.530	.308	.366	.259	.300	.319	.408	.705	.501	.419	.280	.426	.628	.710
	c: 6, w: 0	.543	.309	.367	.259	.296	.324	.404	.695	.516	.432	.289	.433	.641	.719
	c: 6, w: 2	.541	.312	.369	.267	.305	.326	.409	.702	.507	.428	.278	.433	.640	.727
ROUGE	ROUGE-1	.539	.293	.341	.235	.279	.299	.428	.600	.480	.396	.305	.434	.613	.698
	ROUGE-2	.375	.194	.228	.139	.156	.196	.247	.577	.449	.388	.243	.425	.628	.733
	ROUGE-4	.218	.058	.087	.011	.020	.062	.082	.466	.331	.303	.136	.352	.522	.659
	ROUGE-L	.540	.295	.342	.238	.282	.299	.429	.558	.458	.383	.273	.435	.633	.731
METEOR	α: 0.2, γ: 0.0	.542	.286	.337	.230	.285	.267	.423	.398	.437	.349	.295	.404	.562	.652
	α: 0.2, γ: 0.5	.527	.284	.334	.231	.278	.270	.398	.474	.443	.349	.264	.408	.588	.686
	α: 0.9, γ: 0.0	.533	.317	.362	.270	.294	.357	.419	.712	.499	.435	.299	.451	.601	.715
	α: 0.9, γ: 0.5	.516	.299	.340	.256	.275	.329	.375	.689	.483	.409	.263	.438	.617	.729
BERTSc.	cs, Not Limited	.528	.315	.372	.271	.330	.276	.424	.632	.457	.396	.266	.445	.618	.733
	cs, Limited (L5)	.440	.282	.346	.253	.318	.215	.403	.610	.406	.354	.239	.436	.589	.721
	en, Not Limited	.538	.338	.382	.292	.345	.295	.403	.634	.496	.425	.292	.452	.628	.739
	en, Limited (L5)	.423	.277	.333	.245	.318	.223	.392	.626	.516	.425	.393	.499	.617	.729
COMET	wmt20-da	.620	.311	.350	.310	.356	.318	.471	.475	.438	.410	.455	.507	.520	.628
	wmt20-qe-da	.015	-.132	-.129	-.036	-.048	-.048	-.229	.235	.105	.116	.217	.211	.065	.025
	wmt22-da	.595	.318	.359	.333	.367	.313	.455	.503	.474	.444	.475	.526	.519	.643
	cometinho-da	.570	.306	.374	.292	.370	.286	.448	.459	.403	.379	.362	.459	.529	.653
BLEURT	20-D12	.529	.313	.339	.311	.346	.328	.454	.633	.582	.467	.421	.431	.601	.660
	base-512	.582	.310	.344	.326	.372	.295	.474	.423	.536	.431	.501	.445	.491	.572
	large-512	.624	.393	.393	.439	.449	.412	.525	.548	.633	.525	.610	.545	.570	.636
	tiny-128	.576	.375	.430	.340	.424	.344	.482	.419	.582	.487	.410	.424	.540	.624

Metric	Parameters	WMT 21 en-ha							WMT 21 xh-zu						
		Hum	Llama 4		Llama 3.3		Qwen 3		GT	Llama 4		Llama 3.3		Qwen 3	
			F	Z	F	Z	F	Z		F	Z	F	Z	F	Z
BLEU	Order 1	.174	.373	.285	.179	.200	.285	.078	.212	.349	.232	.193	.220	.349	.152
	Order 2	.175	.357	.275	.162	.185	.289	.081	.217	.331	.224	.187	.211	.355	.156
	Order 3	.166	.345	.267	.154	.177	.288	.080	.215	.314	.213	.175	.199	.355	.157
	Order 4	.160	.335	.259	.148	.170	.287	.080	.208	.298	.200	.161	.186	.355	.157
chrF	c: 4, w: 0	.192	.392	.292	.171	.181	.298	.091	.268	.371	.245	.200	.216	.365	.163
	c: 4, w: 2	.193	.376	.280	.158	.173	.289	.085	.266	.357	.238	.193	.215	.356	.156
	c: 6, w: 0	.196	.379	.285	.163	.174	.295	.090	.276	.363	.244	.198	.214	.361	.161
	c: 6, w: 2	.195	.372	.280	.157	.171	.289	.085	.273	.356	.240	.195	.214	.356	.157
ROUGE	ROUGE-1	.173	.371	.294	.184	.191	.293	.098	.201	.356	.259	.207	.215	.359	.156
	ROUGE-2	.159	.347	.275	.161	.175	.297	.097	.193	.302	.213	.151	.174	.355	.158
	ROUGE-4	.103	.286	.221	.105	.118	.286	.092	.135	.206	.134	.055	.091	.346	.151
	ROUGE-L	.180	.388	.320	.210	.249	.305	.106	.215	.368	.277	.224	.243	.366	.172
METEOR	α: 0.2, γ: 0.0	.153	.377	.323	.228	.235	.286	.087	.195	.361	.275	.233	.237	.363	.167
	α: 0.2, γ: 0.5	.158	.372	.316	.220	.231	.281	.089	.206	.359	.274	.231	.240	.359	.165
	α: 0.9, γ: 0.0	.168	.364	.264	.148	.156	.269	.074	.187	.340	.221	.178	.200	.340	.140
	α: 0.9, γ: 0.5	.169	.355	.261	.148	.163	.267	.078	.201	.337	.223	.182	.207	.337	.140
BERTSc.	cs, Not Limited	.191	.384	.299	.193	.220	.273	.076	.257	.370	.250	.212	.224	.355	.160
	cs, Limited (L5)	.195	.374	.295	.195	.228	.273	.074	.226	.357	.236	.210	.223	.353	.154
	en, Not Limited	.178	.375	.296	.187	.201	.270	.074	.251	.371	.260	.221	.233	.363	.159
	en, Limited (L5)	.181	.368	.289	.181	.208	.269	.073	.277	.367	.251	.221	.229	.359	.159
COMET	wmt20-da	.254	.468	.402	.346	.355	.308	.096	.276	.408	.302	.284	.296	.370	.182
	wmt20-qe-da	.228	.267	.211	.463	.213	.283	.039	.276	.280	.293	.342	.318	.256	.125
	wmt22-da	.269	.463	.394	.362	.369	.304	.101	.277	.381	.283	.278	.295	.360	.177
	cometinho-da	.200	.382	.300	.224	.238	.295	.091	.237	.377	.253	.225	.226	.363	.162
BLEURT	20-D12	.192	.384	.306	.209	.233	.284	.082	.185	.363	.247	.214	.241	.361	.162
	base-512	.186	.393	.312	.238	.266	.299	.096	.239	.368	.252	.230	.244	.351	.164
	large-512	.185	.392	.303	.211	.245	.291	.090	.229	.358	.228	.202	.223	.355	.160
	tiny-128	.179	.390	.302	.207	.213	.284	.114	.215	.354	.232	.199	.206	.324	.145

Table 6: Spearman Correlations across different metrics and model variations for MOCHA Validation and RoSE CNNDM Test (above) and WMT 21 en-ha and WMT 21 xh-zu (below).

Metric	Parameters	WMT 24 cs-uk							WMT 24 en-cs						
		Hum	Llama 4		Llama 3.3		Qwen 3		GT	Llama 4		Llama 3.3		Qwen 3	
			F	Z	F	Z	F	Z		F	Z	F	Z	F	Z
BLEU	Order 1	.273	.469	.409	.316	.323	.423	.277	.232	.393	.339	.227	.245	.295	.209
	Order 2	.223	.434	.383	.305	.311	.421	.272	.246	.379	.324	.205	.223	.291	.202
	Order 3	.174	.415	.362	.302	.307	.419	.260	.250	.367	.312	.192	.209	.290	.198
	Order 4	.151	.394	.342	.294	.297	.416	.248	.253	.354	.299	.177	.195	.290	.194
chrF	c: 4, w: 0	.323	.517	.459	.333	.318	.481	.341	.229	.417	.357	.227	.229	.333	.242
	c: 4, w: 2	.316	.494	.433	.307	.304	.446	.310	.232	.405	.347	.214	.221	.319	.229
	c: 6, w: 0	.318	.508	.453	.323	.313	.474	.333	.233	.407	.350	.217	.222	.327	.235
	c: 6, w: 2	.315	.495	.436	.307	.306	.450	.313	.234	.402	.345	.211	.219	.320	.228
ROUGE	ROUGE-1	.020	.136	.120	.100	.091	.110	.066	.222	.416	.362	.245	.248	.332	.251
	ROUGE-2	.003	.095	.085	.067	.064	.079	.056	.211	.376	.323	.203	.206	.314	.228
	ROUGE-4	-.004	.067	.055	.034	.039	.050	.039	.122	.319	.273	.154	.154	.296	.212
	ROUGE-L	.020	.136	.120	.101	.092	.110	.067	.235	.413	.363	.237	.264	.327	.252
METEOR	α: 0.2, γ: 0.0	.284	.414	.369	.274	.263	.380	.271	.214	.399	.357	.262	.270	.319	.241
	α: 0.2, γ: 0.5	.236	.399	.349	.266	.261	.353	.242	.215	.388	.346	.247	.258	.307	.227
	α: 0.9, γ: 0.0	.291	.423	.349	.228	.237	.360	.240	.230	.407	.343	.213	.225	.313	.224
	α: 0.9, γ: 0.5	.242	.406	.338	.238	.246	.340	.222	.226	.393	.332	.203	.219	.303	.212
BERTSc.	cs, Not Limited	.330	.547	.485	.372	.369	.484	.349	.236	.417	.370	.246	.255	.327	.238
	cs, Limited (L5)	.326	.545	.487	.387	.402	.475	.337	.260	.405	.359	.239	.263	.310	.222
	en, Not Limited	.311	.514	.476	.357	.352	.477	.339	.200	.415	.364	.250	.252	.325	.242
	en, Limited (L5)	.321	.499	.464	.349	.343	.472	.336	.220	.409	.354	.232	.245	.311	.227
COMET	wmt20-da	.368	.632	.596	.536	.510	.583	.437	.350	.570	.536	.474	.466	.479	.397
	wmt20-qe-da	.263	.293	.339	.295	.345	.317	.177	.333	.355	.379	.430	.417	.378	.248
	wmt22-da	.388	.619	.589	.539	.518	.573	.428	.392	.565	.536	.499	.498	.484	.400
	cometinho-da	.361	.603	.560	.481	.459	.555	.401	.351	.513	.476	.387	.389	.412	.329
BLEURT	20-D12	.352	.597	.563	.494	.455	.556	.417	.375	.496	.457	.364	.377	.418	.332
	base-512	.292	.498	.462	.373	.346	.455	.312	.250	.408	.373	.281	.287	.324	.244
	large-512	.281	.485	.451	.356	.365	.446	.294	.238	.406	.367	.258	.277	.312	.228
	tiny-128	.221	.445	.413	.327	.339	.412	.216	.221	.443	.404	.279	.283	.351	.275

Metric	Parameters	WMT 24 en-is						
		Hum	Llama 4		Llama 3.3		Qwen 3	
			F	Z	F	Z	F	Z
BLEU	Order 1	.337	.393	.318	.237	.245	.301	.201
	Order 2	.368	.387	.314	.220	.228	.299	.196
	Order 3	.372	.378	.308	.208	.214	.299	.194
	Order 4	.373	.367	.299	.192	.198	.300	.192
chrF	c: 4, w: 0	.319	.419	.343	.231	.222	.337	.235
	c: 4, w: 2	.335	.409	.333	.221	.215	.318	.219
	c: 6, w: 0	.333	.413	.339	.225	.218	.331	.229
	c: 6, w: 2	.341	.407	.333	.219	.214	.320	.220
ROUGE	ROUGE-1	.324	.415	.347	.244	.242	.326	.232
	ROUGE-2	.291	.395	.326	.224	.218	.315	.220
	ROUGE-4	.146	.342	.279	.172	.163	.299	.203
	ROUGE-L	.347	.420	.358	.255	.273	.328	.236
METEOR	α: 0.2, γ: 0.0	.335	.406	.349	.268	.269	.314	.221
	α: 0.2, γ: 0.5	.330	.398	.341	.258	.261	.309	.211
	α: 0.9, γ: 0.0	.331	.397	.312	.204	.203	.299	.200
	α: 0.9, γ: 0.5	.329	.389	.308	.204	.207	.296	.193
BERTSc.	cs, Not Limited	.359	.425	.352	.246	.248	.325	.229
	cs, Limited (L5)	.382	.413	.343	.243	.258	.317	.217
	en, Not Limited	.308	.427	.358	.250	.248	.340	.243
	en, Limited (L5)	.370	.423	.353	.241	.252	.328	.232
COMET	wmt20-da	.443	.544	.509	.424	.416	.432	.337
	wmt20-qe-da	.379	.389	.369	.405	.386	.368	.239
	wmt22-da	.474	.535	.500	.429	.421	.413	.312
	cometinho-da	.418	.482	.428	.349	.346	.385	.288
BLEURT	20-D12	.416	.450	.394	.307	.313	.367	.265
	base-512	.335	.422	.359	.289	.291	.338	.243
	large-512	.342	.407	.337	.240	.268	.325	.221
	tiny-128	.331	.442	.380	.278	.269	.353	.271

Table 7: Spearman Correlations across different metrics and model variations for WMT 24 cs-uk and WMT 24 en-cs (above) and WMT 24 en-is (below).

```
You are a Semantic Corruption Engine for NLP evaluation.
Your task is to generate a single "synthetic hypothesis" string by modifying the provided '
    input_answer' based on the requested 'damage_level'.

### GROUND TRUTH PROTOCOL
1. **Facts:** Treat the 'input_answer' as the absolute factual truth for this task.
   - At Level 0, you must agree with the 'input_answer'.
   - At Level 5, you must contradict the 'input_answer'.
2. **Context:** Use the provided 'question' to understand the topic, gender, and grammatical context
    required for the answer.

### DAMAGE SPECIFICATIONS
Level 0 (Paraphrase): Rewrite the 'input_answer' using different words or grammar, but strictly
    preserve the original meaning.
Level 1 (Surface Noise): Keep the meaning true. You may remove minor adjectives, generalize numbers,
    or simplify phrasing.
Level 2 (Omission): Remove a specific detail (like a name, date, or location). Make the answer
    vaguely true but less informative.
Level 3 (Minor Semantic Error): Keep the topic, but alter a specific entity to a plausible but
    incorrect one (e.g., swap a city for a nearby town, change a date by a few years).
Level 4 (Major Semantic Error): Significantly alter the meaning. Swap the main Subject or Object to
    something clearly wrong but related (e.g., change the actor to a different actor).
Level 5 (Hallucination): Produce a fluent, confident answer to the 'question' that is completely
    factually wrong compared to the 'input_answer'.

### CONSTRAINTS
1. OUTPUT LANGUAGE: The output must be in the SAME LANGUAGE and script as the 'input_answer' (e.g.,
    if input is Czech, output must be Czech).
2. FORMAT: Output ONLY the resulting text string. Do not include labels like "Output:" or
    explanations.
```

Table 8: Prompt for zero-shot synthetic data generation for question answering on the CUS-QA dataset.

```
You are a Semantic Corruption Engine for Reading Comprehension.
Your task is to generate a single "synthetic text" string by modifying the provided 'input_answer'
    based on the requested 'damage_level'.

### GROUND TRUTH PROTOCOL
1. **Source of Truth:** The 'passage' is the absolute factual truth. Any deviation from the passage
    counts as damage.
2. **Relevance:** The output must still attempt to answer the 'question', even if the facts are
    modified (at higher levels).

### DAMAGE SPECIFICATIONS
Level 0 (Paraphrase): Rewrite the 'input_answer' using different words or syntax. You MUST preserve
    the exact meaning supported by the 'passage'.
Level 1 (Surface Noise): Keep the meaning true. You may remove minor adjectives, generalize numbers
    slightly, or simplify phrasing.
Level 2 (Loss of Precision): Omit a secondary detail or make the answer slightly less specific than
    the 'input_answer'.
Level 3 (Minor Semantic Error): Alter a specific entity or relationship. Swap a name, date, or
    location with a plausible but incorrect one not supported by the 'passage'.
Level 4 (Major Semantic Error): Significantly alter the core meaning. Swap the main Subject/Object or
     negate the main verb.
Level 5 (Total Hallucination): Produce a fluent, confident answer that is completely unsupported by
    the 'passage' or explicitly contradicts it.

### CONSTRAINTS
1. OUTPUT LANGUAGE: English (unless the input is in another language).
2. FORMAT: Output ONLY the resulting text string. Do not include labels, explanations, or quotes.
```

Table 9: Prompt for zero-shot synthetic data generation for question answering on the MOCHA dataset.

```
You are a Semantic Corruption Engine for NLP evaluation.
Your task is to generate a single "synthetic hypothesis" string by modifying the provided '
    input_answer' based on the requested 'damage_level'.

### GROUND TRUTH PROTOCOL
1. **Facts:** Treat the 'input_answer' as the absolute factual truth for this task.
   - At Level 0, you must agree with the 'input_answer'.
   - At Level 5, you must contradict the 'input_answer'.
2. **Context:** Use the provided 'question' to understand the topic, gender, and grammatical context
    required for the answer.

### DAMAGE SPECIFICATIONS
Level 0 (Paraphrase): Rewrite the 'input_answer' using different words or grammar, but strictly
    preserve the original meaning.
Level 1 (Surface Noise): Keep the meaning true. You may remove minor adjectives, generalize numbers,
    or simplify phrasing.
Level 2 (Omission): Remove a specific detail (like a name, date, or location). Make the answer
    vaguely true but less informative.
Level 3 (Minor Semantic Error): Keep the topic, but alter a specific entity to a plausible but
    incorrect one (e.g., swap a city for a nearby town, change a date by a few years).
Level 4 (Major Semantic Error): Significantly alter the meaning. Swap the main Subject or Object to
    something clearly wrong but related (e.g., change the actor to a different actor).
Level 5 (Hallucination): Produce a fluent, confident answer to the 'question' that is completely
    factually wrong compared to the 'input_answer'.

### EXAMPLES

User:
question: Who directed the movie 'Titanic'?
input_answer: James Cameron directed it.
damage_level: 0

Assistant:
The film was directed by James Cameron.

User:
question: What is the capital of Slovakia?
input_answer: Bratislava.
damage_level: 3

Assistant:
Košice.

User:
question: Jaké je hlavní město České republiky?
input_answer: Hlavním městem je Praha.
damage_level: 5

Assistant:
Hlavním městem je Ostrava, známá svými plážemi.

### CONSTRAINTS
1. OUTPUT LANGUAGE: The output must be in the SAME LANGUAGE and script as the 'input_answer'.
2. FORMAT: Output ONLY the resulting text string. Do not include labels like "Output:" or
    explanations.
```

Table 10: Prompt for few-shot synthetic data generation for question answering on the CUS-QA dataset.

```
You are a Semantic Corruption Engine for Reading Comprehension.
Your task is to generate a single "synthetic text" string by modifying the provided 'input_answer'
    based on the requested 'damage_level'.

### GROUND TRUTH PROTOCOL
1. **Source of Truth:** The 'passage' is the absolute factual truth. Any deviation from the passage
    counts as damage.
2. **Relevance:** The output must still attempt to answer the 'question', even if the facts are
    modified (at higher levels).

### DAMAGE SPECIFICATIONS
Level 0 (Paraphrase): Rewrite the 'input_answer' using different words or syntax. You MUST preserve
    the exact meaning supported by the 'passage'.
Level 1 (Surface Noise): Keep the meaning true. You may remove minor adjectives, generalize numbers
    slightly, or simplify phrasing.
Level 2 (Loss of Precision): Omit a secondary detail or make the answer slightly less specific than
    the 'input_answer'.
Level 3 (Minor Semantic Error): Alter a specific entity or relationship. Swap a name, date, or
    location with a plausible but incorrect one not supported by the 'passage'.
Level 4 (Major Semantic Error): Significantly alter the core meaning. Swap the main Subject/Object or
     negate the main verb.
Level 5 (Total Hallucination): Produce a fluent, confident answer that is completely unsupported by
    the 'passage' or explicitly contradicts it.

### CONSTRAINTS
1. OUTPUT LANGUAGE: English (unless the input is in another language).
2. FORMAT: Output ONLY the resulting text string. Do not include labels, explanations, or quotes.

### EXAMPLES

User:
passage: The Apollo 11 mission landed humans on the Moon in July 1969.
question: When did the landing occur?
input_answer: It happened in 1969.
damage_level: 0

Assistant:
The landing took place in the year 1969.

User:
passage: Photosynthesis takes place inside the chloroplasts, which contain chlorophyll.
question: Where does photosynthesis happen?
input_answer: It occurs in the chloroplasts.
damage_level: 3

Assistant:
It occurs in the mitochondria.

User:
passage: The blue whale is the largest animal known to have ever lived.
question: What is the largest animal?
input_answer: The blue whale.
damage_level: 5

Assistant:
The largest animal is the African Elephant.
```

Table 11: Prompt for few-shot synthetic data generation for question answering on the MOCHA dataset.

```
You are an Atomic Fact Corruption Engine.
Your task is to generate a "synthetic text" by modifying a 'reference_summary' based on a '
    damage_level', specifically targeting Atomic Content Units (ACUs).

### THE ACU PROTOCOL
Summaries are evaluated by breaking them down into "Atomic Content Units" (fine-grained, independent
    facts) and checking their recall.
- **Goal:** As Damage Level increases, the number of ACUs from the 'reference_summary' preserved in
    your output must DECREASE.
- **Constraint:** You must maintain the **fluency** and **length** of the original text. Do not
    simply delete sentences; replace facts with non-facts or plausible hallucinations.

### DAMAGE SPECIFICATIONS (ACU RECALL)
Level 0 (100% ACU Recall): Paraphrase the text but preserve **every single atomic fact** (names,
    dates, relations, quantities).
Level 1 (80% ACU Recall): Preserve the main story but blur specific details. (e.g., Change "David
    Ospina" to "the goalkeeper", or "16th minute" to "early on").
Level 2 (60% ACU Recall): Remove minor ACUs. Replace specific facts with generic filler text that
    sounds relevant but conveys no specific information from the source.
Level 3 (40% ACU Recall): Entity Swap. Keep the sentence structure but swap key entities (Subject/
    Object) so the ACUs become factually false (e.g., "Chelsea won" -> "Arsenal won").
Level 4 (20% ACU Recall): Major Contradiction. Rewrite the summary to describe a different outcome or
     event involving the same entities, falsifying nearly all original facts.
Level 5 (0% ACU Recall): Total Hallucination. Generate a fluent summary of the same length that
    contains **ZERO** facts from the reference. It can be about the same topic but must be factually
     disjoint.

### CONSTRAINTS
1. LENGTH: The output must be within ±10% word count of the 'reference_summary'.
2. FLUENCY: The text must be grammatically perfect.
3. FORMAT: Output ONLY the resulting summary string. No labels.
```

Table 12: Prompt for zero-shot synthetic data generation for summarization on the RoSE dataset.

```
You are a Translation Corruption Engine.
Your goal is to take a perfect 'reference_translation' and degrade it according to the specific '
    damage_level' requested.

### GROUND TRUTH PROTOCOL
1. **Source of Truth:** The 'source_sentence' and 'reference_translation' define the correct meaning.
2. **Strict adherence:** You must NOT improve the text. You must damage it.

### DAMAGE SPECIFICATIONS
Level 0 (Paraphrase): Rewrite the 'reference_translation' using different synonyms or sentence
    structures. It MUST remain a valid, high-quality translation of the 'source_sentence' with
    perfect grammar.
Level 1 (Surface/Mechanical Noise): Keep the words mostly identical to the 'reference_translation',
    but **inject a visible technical error**. You MUST include a spelling mistake, a capitalization
    error, missing punctuation, or a blatant subject-verb agreement error (e.g., "he go" instead of
    "he goes"). The meaning must remain perfect, but the fluency must be damaged.
Level 2 (Omission/Under-translation): Remove a specific detail or nuance found in the '
    source_sentence' (e.g., drop an adjective or adverb). The translation is understandable but
    clearly incomplete compared to the reference.
Level 3 (Word-Level Semantic Error): Mistranslate a specific content word (noun/verb) to a plausible
    but incorrect alternative (e.g., "car" -> "truck", "walked" -> "ran"). This must be a specific,
    local error.
Level 4 (Major Semantic Error): Significantly alter the meaning of the whole sentence. Swap the
    Subject and Object, negate the main verb, or change the tense dramatically (past -> future) if
    it contradicts the source.
Level 5 (Hallucination/Catastrophic Failure): Produce a fluent sentence in the target language that
    has NOTHING to do with the 'source_sentence', or is a translation of a completely different
    input.

### CONSTRAINTS
1. OUTPUT LANGUAGE: The output must be in the SAME LANGUAGE as the 'reference_translation'.
2. NO EXACT MATCHES: For Damage Level 1 and above, the output **MUST NOT** be identical to the '
    reference_translation'.
3. FORMAT: Output ONLY the resulting translation string. No labels, no explanations.
```

Table 13: Prompt for zero-shot synthetic data generation for machine translation.

```
You are an Atomic Fact Corruption Engine.
Your task is to generate a "synthetic text" by modifying a 'reference_summary' based on a '
    damage_level', specifically targeting Atomic Content Units (ACUs).

### THE ACU PROTOCOL
Summaries are evaluated by breaking them down into "Atomic Content Units" (fine-grained, independent
    facts) and checking their recall.
- **Goal:** As Damage Level increases, the number of ACUs from the 'reference_summary' preserved in
    your output must DECREASE.
- **Constraint:** You must maintain the **fluency** and **length** of the original text. Do not
    simply delete sentences; replace facts with non-facts or plausible hallucinations.

### DAMAGE SPECIFICATIONS (ACU RECALL)
Level 0 (100% ACU Recall): Paraphrase the text but preserve **every single atomic fact** (names,
    dates, relations, quantities).
Level 1 (80% ACU Recall): Preserve the main story but blur specific details. (e.g., Change "David
    Ospina" to "the goalkeeper", or "16th minute" to "early on").
Level 2 (60% ACU Recall): Remove minor ACUs. Replace specific facts with generic filler text that
    sounds relevant but conveys no specific information from the source.
Level 3 (40% ACU Recall): Entity Swap. Keep the sentence structure but swap key entities (Subject/
    Object) so the ACUs become factually false (e.g., "Chelsea won" -> "Arsenal won").
Level 4 (20% ACU Recall): Major Contradiction. Rewrite the summary to describe a different outcome or
     event involving the same entities, falsifying nearly all original facts.
Level 5 (0% ACU Recall): Total Hallucination. Generate a fluent summary of the same length that
    contains **ZERO** facts from the reference. It can be about the same topic but must be factually
     disjoint.

### CONSTRAINTS
1. LENGTH: The output must be within ±10% word count of the 'reference_summary'.
2. FLUENCY: The text must be grammatically perfect.
3. FORMAT: Output ONLY the resulting summary string. No labels.

### EXAMPLES

User:
source_text: (Context about the Chelsea vs Arsenal match)
reference_summary: Chelsea weren't awarded a penalty for David Ospina's clash with Oscar. Arsenal
    goalkeeper clattered Oscar inside the box.
damage_level: 0

Assistant:
David Ospina collided with Oscar in the box, but Chelsea did not receive a penalty for the incident
    involving the Arsenal goalkeeper.

User:
source_text: (Context about the Chelsea vs Arsenal match)
reference_summary: Chelsea weren't awarded a penalty for David Ospina's clash with Oscar. Arsenal
    goalkeeper clattered Oscar inside the box.
damage_level: 3

Assistant:
Arsenal weren't awarded a penalty for Petr Cech's clash with Willian. The Chelsea goalkeeper
    clattered Willian outside the box.

User:
source_text: (Context about SpaceX Starship)
reference_summary: SpaceX's Starship rocket exploded four minutes after launch on Thursday, with no
    injuries reported.
damage_level: 5

Assistant:
NASA's Falcon 9 booster landed successfully on the drone ship on Friday, marking another milestone
    for the agency.
```

Table 14: Prompt for few-shot synthetic data generation for summarization on the RoSE dataset.

```
You are a Translation Corruption Engine.
Your goal is to take a perfect 'reference_translation' and degrade it according to the specific '
    damage_level' requested.

### GROUND TRUTH PROTOCOL
1. **Source of Truth:** The 'source_sentence' and 'reference_translation' define the correct meaning.
2. **Strict adherence:** You must NOT improve the text. You must damage it.

### DAMAGE SPECIFICATIONS
Level 0 (Paraphrase): Rewrite the 'reference_translation' using different synonyms or sentence
    structures. It MUST remain a valid, high-quality translation of the 'source_sentence' with
    perfect grammar.
Level 1 (Surface/Mechanical Noise): Keep the words mostly identical to the 'reference_translation',
    but **inject a visible technical error**. You MUST include a spelling mistake, a capitalization
    error, missing punctuation, or a blatant subject-verb agreement error (e.g., "he go" instead of
    "he goes"). The meaning must remain perfect, but the fluency must be damaged.
Level 2 (Omission/Under-translation): Remove a specific detail or nuance found in the '
    source_sentence' (e.g., drop an adjective or adverb). The translation is understandable but
    clearly incomplete compared to the reference.
Level 3 (Word-Level Semantic Error): Mistranslate a specific content word (noun/verb) to a plausible
    but incorrect alternative (e.g., "car" -> "truck", "walked" -> "ran"). This must be a specific,
    local error.
Level 4 (Major Semantic Error): Significantly alter the meaning of the whole sentence. Swap the
    Subject and Object, negate the main verb, or change the tense dramatically (past -> future) if
    it contradicts the source.
Level 5 (Hallucination/Catastrophic Failure): Produce a fluent sentence in the target language that
    has NOTHING to do with the 'source_sentence', or is a translation of a completely different
    input.

### CONSTRAINTS
1. OUTPUT LANGUAGE: The output must be in the SAME LANGUAGE as the 'reference_translation'.
2. NO EXACT MATCHES: For Damage Level 1 and above, the output **MUST NOT** be identical to the '
    reference_translation'.
3. FORMAT: Output ONLY the resulting translation string. No labels, no explanations.

### EXAMPLES

User:
source_sentence: The cat sat on the mat.
reference_translation: Le chat s'est assis sur le tapis.
damage_level: 0

Assistant:
Le chat était assis sur le tapis.

User:
source_sentence: She bought a red car yesterday.
reference_translation: Sie hat gestern ein rotes Auto gekauft.
damage_level: 3

Assistant:
Sie hat gestern ein blaues Fahrrad gekauft.

User:
source_sentence: Technology is evolving rapidly.
reference_translation: La technologie évolue rapidement.
damage_level: 5

Assistant:
J'aime manger des pommes au petit déjeuner.
```

Table 15: Prompt for few-shot synthetic data generation for machine translation.

```
### DAMAGE SPECIFICATIONS
Level 0 (Semantic Equivalence): Reformulate the input text using novel
  vocabulary or syntax while retaining 100% of the original information
  and meaning.
Level 1 (Lossy Compression): Maintain factual accuracy but degrade
  precision. Approximate numerical values, strip non-essential
  modifiers, or use simpler diction.
Level 2 (Information Redaction): Omit key identifiers (such as specific
  names, dates, or locations). The statement should remain truthful but
  lack specificity.
Level 3 (Low-Level Distortion): Retain the general context but
  substitute a single specific entity with a plausible but incorrect
  alternative (e.g., shifting a date slightly or swapping a city for
  a neighbor).
Level 4 (High-Level Distortion): Fundamentally invalidate the core
  meaning. Replace the primary Subject or Object with an entity that
  is contextually related but demonstrably incorrect.
Level 5 (Complete Fabrication): Generate a hallucinated response that
  answers the question with high confidence and fluency, but is
  diametrically opposed to the facts in the input answer.
```

Table 16: Damage specifications for CUS-QA sensitivity Variant 1. The surrounding system role, ground-truth protocol, and output constraints are identical to Table 8.

```
### DAMAGE SPECIFICATIONS
Level 0 (Restate): Say exactly the same thing as the input answer,
  but say it differently. Do not change the facts.
Level 1 (Simplify): Keep the answer correct, but make it rougher.
  Remove specific adjectives or round off numbers to be less precise.
Level 2 (Delete Details): Remove specific proper nouns or data points.
  The sentence should still be technically true, just vague.
Level 3 (Tweak Entity): Keep the general topic correct, but swap one
  specific detail for a close semantic neighbor
  (e.g., Tuesday -> Wednesday, Paris -> Lyon).
Level 4 (Break Relationship): Change the main actor or target of the
  sentence. Make the statement clearly wrong by swapping the key
  subject/object.
Level 5 (Lie): Ignore the truth. Write a smooth, well-phrased answer
  to the question that is completely false compared to the input answer.
```

Table 17: Damage specifications for CUS-QA sensitivity Variant 2 (colloquial action-label style).

```
### DAMAGE SPECIFICATIONS
Level 0 (True - Rephrased): The output must be factually identical to
  the input, differing only in style or phrasing.
Level 1 (True - Noisy): The output remains factually compatible with
  the input, but loses fine-grained detail or precision.
Level 2 (True - Incomplete): The output is factually compatible but
  insufficient; critical specifics like names or dates are missing.
Level 3 (False - Plausible): The output contains a small factual
  error. It sounds correct but includes a near-miss entity
  substitution.
Level 4 (False - Major): The output contains a significant factual
  error regarding the main subject or object, rendering the core
  statement false.
Level 5 (False - Total): The output is a complete fabrication. It
  presents a confident, fictional reality that contradicts the input
  entirely.
```

Table 18: Damage specifications for CUS-QA sensitivity Variant 3. The surrounding system role, ground-truth protocol, and output constraints are identical to Table 8.

```
### DAMAGE SPECIFICATIONS
Level 0 (Synonym Swap): Edit the sentence structure and vocabulary
  without altering the underlying logic or facts.
Level 1 (Generalization): Edit the text to be less specific. Convert
  exact figures to ranges or estimates; remove descriptive flair.
Level 2 (Redaction): Edit the text to remove proper nouns
  (Who/Where/When). Leave the "What" intact but vague.
Level 3 (Minor Glitch): Edit a single entity. Change a specific detail
  to something that looks similar but is factually wrong.
Level 4 (Major Swap): Edit the key players. Change the Subject or
  Object to a different entity, breaking the factual link.
Level 5 (Creative Writing): Discard the facts. Write a convincing but
  entirely invented answer to the prompt.
```

Table 19: Damage specifications for CUS-QA sensitivity Variant 4. The surrounding system role, ground-truth protocol, and output constraints are identical to Table 8.

```
### DAMAGE SPECIFICATIONS
Level 0: Paraphrase. Keep meaning exact.
Level 1: Generalize. Remove adjectives, simplify numbers.
        Keep meaning true.
Level 2: Omit. Remove names/dates/locations. Answer becomes vague.
Level 3: Minor Error. Swap one entity for a plausible incorrect one.
Level 4: Major Error. Swap the main Subject or Object.
        Meaning is now false.
Level 5: Hallucinate. Generate a confident, fluent, but totally
        false answer.
```

Table 20: Damage specifications for CUS-QA sensitivity Variant 5. The surrounding system role, ground-truth protocol, and output constraints are identical to Table 8.