Semantic Similarity is Multidimensional: Variation-Aware Evaluation for Medical Text

Akhil Juneja **Nils Feldhus** **Roland Roller**
German Research Center for Artificial Intelligence (DFKI)
Salzufer 15/16, 10587 Berlin, Germany
roland.roller@dfki.de

Abstract

Semantic similarity is commonly modeled as a single scalar score. However, two texts may appear highly similar while differing critically along specific semantic aspects such as factuality, entities, or temporal meaning. This limitation becomes particularly important in high-stakes domains and evaluation settings, where nuanced semantic differences can substantially affect downstream decisions. Within a medical use showcase, this work proposes a multidimensional semantic similarity framework that represents similarity across multiple meaning-relevant dimensions instead of collapsing it into a single value. We introduce a lightweight multi-head model that predicts dimension-specific similarity scores and evaluate the approach on a medical question answering use case. Our results show that modeling semantic similarity multidimensionally substantially improves the detection of meaning-altering variations compared to conventional similarity metrics and small open-source LLM judges, while remaining competitive with much larger LLM-based evaluators at significantly lower computational cost.

1 Introduction

Semantic similarity metrics are widely used to evaluate generated text in tasks such as summarization, question answering, dialogue systems, and retrieval-augmented generation. Existing approaches typically represent similarity between two texts as a single scalar score estimating overall semantic equivalence. Classical overlap-based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) primarily measure lexical similarity, while embedding-based and learned metrics such as BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), MoverScore (Zhao et al., 2019), PubMedBERT embeddings (Mezzetti, 2026), and

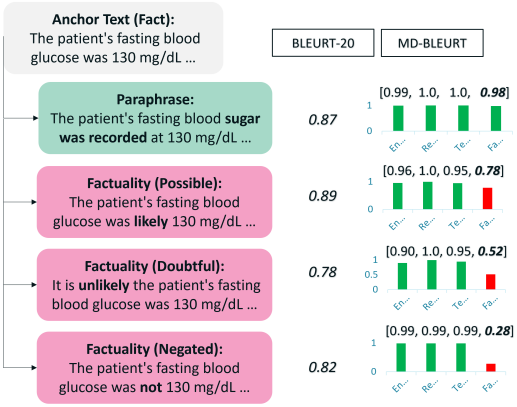

Figure 1: Comparison of multidimensional similarity scoring with conventional scalar similarity metrics for a factuality-related semantic divergence. Standard metrics assign high overall similarity despite contradictory meaning, whereas MD-BLEURT explicitly captures the divergence along the factuality dimension.

EmbeddingGemma (Vera et al., 2025) attempt to capture semantic similarity more directly.

Despite their success, these approaches share a common assumption: semantic similarity can be collapsed into a single dimension. However, semantic equivalence can refer to multiple aspects. Two texts may be highly similar overall while differing critically in specific semantic aspects such as factuality, entities, temporal meaning, or relations between concepts. Figure 1 illustrates a factuality-related semantic divergence, where a multidimensional similarity representation identifies the specific semantic dimension along which two sentences differ. Similar effects arise for other dimensions, such as temporal meaning: for example, '*Patient has cancer*' and '*Patient had cancer*' remain highly similar lexically, while conveying substantially different clinical implications.

Recent work has highlighted important failure modes in evaluation metrics. Semantic metrics often fail to penalize negation, polarity shifts, and other meaning-altering variations (He et al., 2023;

Cao, 2025; Ramprasad and Wallace, 2025). Similarly, work on contradiction detection and natural language inference shows that minimal edits preserving lexical overlap can still completely reverse meaning (Harabagiu et al., 2006; Nie et al., 2020). Behavioral testing and perturbation-based evaluation frameworks further demonstrate that models and metrics may achieve strong aggregate performance while systematically failing on targeted semantic phenomena (Ribeiro et al., 2020; Gehrmann et al., 2021; Deutsch et al., 2021). Together, these findings suggest that semantic variation is not uniform, and that changes in factuality, entities, relations, and temporal reasoning represent distinct semantic phenomena (Kalouli et al., 2022; Pavlick and Callison-Burch, 2016).

Recent work has explored the use of large language models (LLMs) as judges for text evaluation (Li et al., 2025; De la Iglesia et al., 2025; Zhang et al., 2025; Zhou et al., 2025). In parallel, researchers have increasingly recognized that similarity and evaluation should account for multiple semantic dimensions rather than a single scalar score. Chen et al. (2022) introduced a multilingual benchmark annotating news articles along seven dimensions of similarity, and Hatzel et al. (2023) argued that semantic similarity should be represented as interpretable dimensions, training a separate bi-encoder model per dimension for news retrieval. For text generation evaluation, Zhong et al. (2022) proposed UniEval, a unified evaluator scoring coherence, consistency, fluency, and relevance, while (Hashemi et al., 2024) extended this paradigm to calibrated LLM-based evaluation through LLM-RUBRIC.

Our work transfers this multidimensional perspective to the medical domain, where the relevant dimensions (entities, relations, temporality, factuality) differ from those used for news similarity and where evaluation errors can have direct clinical consequences. In contrast to training one model per dimension (Hatzel et al., 2023), MD-BLEURT learns all dimensions jointly within a single lightweight cross-encoder, and in contrast to LLM-based multidimensional evaluators (Zhong et al., 2022; Hashemi et al., 2024), it avoids the high computational cost and inference latency of LLM judges as well as the limited reproducibility and privacy concerns of proprietary systems – key requirements in large-scale, privacy-sensitive domains such as healthcare.

In this work, we build on this perspective that semantic similarity should not be represented as a single scalar quantity but as a multidimensional representation over distinct semantic dimensions. As a lightweight realization of this idea, we introduce MD-BLEURT[1], a multidimensional extension of BLEURT based on multiple regression heads predicting dimension-specific similarity scores. We evaluate our framework on a medical question answering use case, where subtle semantic differences can substantially alter clinical meaning and potentially lead to harmful downstream consequences (Finlayson et al., 2021; Kim et al., 2025).

Our experiments show that conventional similarity metrics frequently assign high similarity scores to meaning-altering variations, particularly when lexical overlap remains high. In contrast, MD-BLEURT explicitly captures these semantic divergences, substantially outperforming conventional metrics and smaller open-source LLM judges, while remaining computationally lightweight.

Type	Examples
Entity	disease, medication, dose, route, frequency, laterality
Relation	causal, conjunctive, comparative, event ordering
Temporal	past, present, future shifts
Factuality	fact, possible, doubtful, negated

Table 1: Semantic variation dimensions.

2 Dataset Construction

To study semantic similarity under controlled meaning variation, we construct two datasets: MedVariations, a synthetic fine-grained benchmark, and a modified version of the human-created MedNLP-CHAT dataset. Both datasets apply semantic modifications across the dimensions listed in Table 1, along with additional paraphrases. The datasets are released in two versions. Version 1 uses minimal edits to generate semantic variations, while Version 2 introduces more extensive paraphrastic rewrites through LLM prompting. Additional details and prompting templates are provided in Appendix A.

2.1 MedVariations

Following template-based and LLM-assisted dataset construction strategies (Warstadt et al., 2020; Ribeiro et al., 2020; Long et al., 2024), we generate MedVariations using a two-stage pipeline: (i) anchor generation and (ii) controlled variation

[1] https://github.com/akjuneja/SemEvalVar

generation: **First**, we prompt Gemini-2.5-Pro to produce short clinical statements (2–3 sentences) containing medical entities, relations, temporal references, and factual claims (Table 6 in Appendix A). These anchors resemble simplified clinical narratives while enabling controlled manipulation. **Second**, each anchor is transformed into semantically perturbed variants along four dimensions: entity, relation, temporal, and factuality, as depicted in Figure 1. A validation step ensures that each transformation isolates a single semantic dimension without introducing additional confounds, as shown in the following example:

Anchor: The patient's biopsy result was positive for prostate cancer. His Gleason score is 7.

Entity: The patient's biopsy result was positive for bladder cancer. His Gleason score is 7.

Similarity annotation We define absolute similarity scores via controlled perturbation strength. Paraphrases receive a score of 1, while other variations are penalized as $s = 1 - (\delta \cdot \mu)$, where μ controls penalty strength and δ encodes the semantic distance of the variation.

Entity and relation changes are treated as single-step deviations ($\mu = 0.5$, $\delta = 1$). Temporal and factuality changes are ordered ($\mu = 0.33$): temporal shifts range from past→present ($\delta = 1$) to past→future ($\delta = 2$), while factuality spans fact→possible/doubtful/negated ($\delta \in \{1, 2, 3\}$).

This construction yields structured, interpretable similarity signals aligned with specific meaning-altering phenomena. Further details in Appendix C.

2.2 MedNLP-CHAT

To evaluate generalization beyond synthetic settings, we use MedNLP-CHAT (Aramaki et al., 2025), a dataset of human-curated patient medical questions and corresponding expert answers. We treat those answers as anchors and generate corresponding perturbed versions using the same variation framework as above. This results in a more naturalistic question-answering setting where semantic differences arise from realistic clinical dialogue rather than templated generation. Additional details and examples are in Appendix A.

3 Multi-Dimensional BLEURT Model

For this work, we extend BLEURT (Sellam et al., 2020) into a multidimensional similarity model (MD-BLEURT) by introducing separate regression heads for each semantic dimension. A shared encoder produces a sentence-pair representation, which is passed to four heads corresponding to entity, relation, temporal, and factuality similarity. All heads are trained jointly. For a given training instance, only the head corresponding to its variation type receives the ground-truth score, while other heads are regularized toward high similarity (score = 1). This encourages consistent semantic alignment while enabling specialization per dimension. At inference time, MD-BLEURT outputs a vector of similarity scores, each reflecting a distinct semantic aspect rather than a single aggregated value.

We focus on four semantic dimensions – entity, relation, temporal, and factuality – as they capture complementary aspects of meaning that are critical for interpreting clinical text. Entity and relation similarity preserve the information being conveyed and the connections between clinical concepts, while temporal and factuality similarity capture when events occur and whether they are asserted, uncertain, or negated. Together, these dimensions encompass the most clinically significant semantic variations while remaining sufficiently compact to support interpretable evaluation and efficient model training.

For downstream tasks, the individual dimension scores provide fine-grained insight into semantic differences, whereas an aggregated score offers a single similarity measure for model comparison and benchmark reporting. Examining the individual dimensions enables users to identify whether discrepancies arise from entities, relations, temporal information, or factuality, making the evaluation both interpretable and actionable. In practice, a downstream system can threshold individual heads to flag clinically critical changes. For instance, rejecting candidate answers whose factuality score falls below a threshold while tolerating minor entity deviations or aggregate the heads into a single score for ranking and benchmarking, depending on the application's risk profile.

4 Experimental Setup

Data splits MedVariations is split into train/dev/test sets with non-overlapping anchors to prevent lexical memorization (Table 4 in Appendix A). The same process is applied to MedNLP-CHAT for evaluation (Table 5 in Appendix A).

Baselines We compare MD-BLEURT against: (i) surface metrics (BLEU, ROUGE, METEOR), (ii) semantic metrics (BERTScore, BLEURT), and (iii)

LLM-based evaluators including small open-source judges and larger proprietary models.

Training We fine-tune BLEURT-Base with four regression heads using a shared encoder. Models are selected based on average validation loss across heads, ensuring balanced performance across semantic dimensions.

Evaluation protocol All metrics are computed on identical test pairs, and we report average scores per variation category. Higher scores are desirable for paraphrase pairs, while lower scores indicate correct detection of meaning-altering variations. LLM judges return 1–5 ratings using a shared rubric prompt (Appendix B), linearly rescaled to $[0, 1]$ for comparability with the learned metrics.

5 Results and Discussion

Tables 2 and 3 compare conventional similarity metrics, LLM judges, and MD-BLEURT across controlled semantic variations. Complete per-model experimental results for all evaluated datasets are provided in Appendix E. We focus on cases where lexical overlap remains high while meaning changes along a specific semantic dimension.

How to read the tables. Each column reports the average similarity over all pairs of one variation category. The four MD-BLEURT rows correspond to the four output heads of a *single* jointly trained model, not to separate models. Desirable behavior is a high score in the Paraphrase column together with a low score in the column matching the head's dimension (yellow), while unrelated columns remain high – indicating that the head detects its

Metric	Version 1					Version 2				
	Paraphrase	Entity	Relation	Temporal	Factuality	Paraphrase	Entity	Relation	Temporal	Factuality
BLEU	0.79	0.88	0.88	0.87	0.85	0.40	0.30	0.31	0.36	0.37
ROUGE-L	0.89	0.95	0.95	0.94	0.94	0.66	0.60	0.61	0.63	0.65
METEOR	0.92	0.95	0.96	0.96	0.97	0.76	0.69	0.71	0.74	0.76
BERTScore_F1	0.98	0.99	0.99	0.99	0.99	0.95	0.94	0.94	0.94	0.94
BLEURT	0.89	0.85	0.89	0.91	0.84	0.81	0.76	0.78	0.80	0.76
NegBLEURT	0.59	0.45	0.50	0.75	**0.40**	0.45	0.32	0.35	0.49	**0.31**
Qwen3.5-4B	1.00	0.54	0.57	0.86	0.49	1.00	0.55	0.71	0.98	0.57
Qwen2.5-72B-Instruct-AWQ	1.00	0.64	0.61	0.79	0.55	0.99	0.66	0.69	0.91	0.59
Gemini-3.1-flash-lite	1.00	0.52	0.53	0.72	0.55	1.00	0.54	0.64	0.89	0.56
Gemini-3.1-pro	**1.00**	**0.51**	**0.46**	0.74	0.51	**1.00**	**0.51**	**0.52**	0.86	**0.50**
MD-BLEURT (Entity)	0.97	0.51	0.98	1.01	1.01	0.94	0.52	0.96	0.97	1.03
MD-BLEURT (Relation)	0.98	0.92	0.51	0.91	0.71	0.95	0.91	0.58	0.96	0.73
MD-BLEURT (Temporal)	0.99	1.02	0.98	**0.69**	0.99	0.84	0.92	0.88	**0.69**	0.88
MD-BLEURT (Factuality)	0.99	1.00	0.86	0.95	0.46	1.00	0.99	0.94	1.01	0.57

Table 2: Average similarity scores for the MedVariations dataset (test): The left block (Version 1) compares the similarity of text spans which have been only changed in one dimension. The right block (Version 2) compares the similarity according to the dimensions, including an LLM-based paraphrase step. Pink fields highlight models with a strong similarity drop while comparing semantically equivalent text spans; Yellow fields highlight the corresponding dimension of MD-BLEURT. The four MD-BLEURT rows correspond to the output heads of a single model (cf. reading guidance in §5). Higher scores are expected for paraphrase pairs, while lower scores are desirable for semantic variation categories, as they indicate that the metric correctly recognizes the introduced semantic differences.

Metric	Version 1					Version 2				
	Paraphrase	Entity	Relation	Temporal	Factuality	Paraphrase	Entity	Relation	Temporal	Factuality
BLEU	0.92	0.97	0.97	0.97	0.95	0.30	0.28	0.30	0.30	0.30
ROUGE-L	0.95	0.99	0.99	0.99	0.98	0.55	0.54	0.55	0.55	0.55
METEOR	0.96	0.99	0.99	0.99	0.99	0.63	0.61	0.63	0.63	0.63
BERTScore_F1	0.99	1.00	1.00	1.00	0.99	0.93	0.93	0.93	0.93	0.93
BLEURT	0.92	0.90	0.93	0.92	0.89	0.77	0.75	0.76	0.77	0.75
NegBLEURT	0.45	0.47	0.46	0.53	0.39	0.21	0.19	0.20	0.22	0.17
Qwen3.5-4B	1.00	0.59	0.77	0.95	0.58	1.00	0.79	0.94	1.00	0.81
Qwen2.5-72B-Instruct-AWQ	1.00	0.64	0.78	0.86	0.63	0.99	0.71	0.89	0.98	0.75
Gemini-3.1-flash-lite	1.00	0.57	0.74	0.84	0.57	1.00	0.61	0.86	0.99	0.66
Gemini-3.1-pro	**1.00**	**0.52**	**0.53**	**0.77**	**0.56**	**1.00**	**0.54**	0.76	0.94	**0.62**
MD-BLEURT (Entity)	0.99	0.70	0.99	0.99	0.98	0.93	0.68	0.92	0.93	0.94
MD-BLEURT (Relation)	0.94	0.94	0.74	0.92	0.78	0.83	0.84	**0.75**	0.84	0.73
MD-BLEURT (Temporal)	1.00	0.99	0.95	0.90	0.99	0.83	0.89	0.85	**0.81**	0.86
MD-BLEURT (Factuality)	0.96	1.00	0.92	0.95	0.70	0.95	0.97	0.96	0.95	0.88

Table 3: Average similarity scores on MedNLP-CHAT, following the same layout as Table 2: the MD-BLEURT rows are the heads of a single model, and yellow fields mark each head's matching variation column, where low scores indicate correct detection.

target variation without globally penalizing similarity.

Across both datasets in Version 1, overlap-based metrics such as BLEU, ROUGE, and METEOR struggle to distinguish paraphrases from meaning-altering variations. In Version 2, these metrics strongly penalize semantically equivalent paraphrases due to reduced lexical overlap, while still assigning relatively high scores to altered variants. Embedding-based and semantic similarity metrics including BERTScore and BLEURT improve paraphrase robustness, but still frequently assign high similarity to semantically contradictory text pairs, particularly for entity, temporal, and factuality changes. NegBLEURT is more sensitive to factuality-related changes, but often lowers similarity scores globally, including for paraphrases.

LLM-based judges demonstrate substantially stronger sensitivity to semantic changes. In particular, Gemini-3.1-Pro consistently produces the clearest separation between paraphrases and meaning-altering variations across both datasets, while larger open-source LLMs show competitive performance. However, these gains come at substantially higher inference cost and latency (Table 9 in Appendix D). Additional performance efficiency analyses using Pareto frontier curves are provided in Appendix F.

MD-BLEURT exhibits a fundamentally different behavior from scalar similarity metrics. Instead of collapsing similarity into a single value, the model decomposes similarity across semantic dimensions. On the synthetic MedVariations benchmark, each regression head responds primarily to its corresponding variation type: entity modifications mainly reduce the entity score, factuality changes strongly affect factuality similarity, and temporal changes selectively lower temporal similarity while leaving unrelated dimensions comparatively stable. This demonstrates that MD-BLEURT has a higher representational capacity in comparison to a single similarity score.

On the MedNLP-CHAT benchmark, the separation between dimensions becomes less pronounced, reflecting the increased linguistic diversity and complexity of naturally occurring text. Nevertheless, the overall trend remains consistent, suggesting that multidimensional similarity modeling generalizes beyond templated synthetic examples. The reduced separation also highlights an important limitation of the current work: the present model is trained on relatively controlled semantic transformations. Future work should therefore extend the framework using larger and more diverse human-curated datasets and broader semantic dimensions.

A closer inspection of MD-BLEURT's residual errors reveals where the current supervision breaks down. The relation head is the least dimension-specific: factuality changes also lower its scores substantially (0.71–0.78 across datasets and versions), plausibly because negation and modality shifts frequently co-occur with altered causal or contrastive markers in the training data. The temporal head shows the weakest separation on MedNLP-CHAT (0.90 and 0.81 in Versions 1 and 2), and its paraphrase scores drop under the more extensive rewrites of Version 2 (0.83–0.84), indicating sensitivity to reformulation. Finally, the factuality head degrades most when moving from synthetic to naturalistic text (0.57 on MedVariations vs. 0.88 on MedNLP-CHAT, Version 2). This suggests that factuality cues in expert answers are more diffuse than the single-sentence flips seen during training. These patterns indicate that cross-dimension leakage under domain shift is the main failure mode to address in future iterations.

6 Conclusion

This work highlights the limitations of single-score semantic similarity metrics: collapsing meaning into one scalar fails to capture clinically important differences in factuality, entities, relations, or temporal information. Existing metrics can assign similarly high scores to texts with substantially different meanings. With MD-BLEURT, we demonstrate that semantic similarity can instead be modeled in a fine-grained, multidimensional way, where dimension-specific regression heads enable more targeted and interpretable similarity judgments while remaining computationally efficient. Our findings suggest that moving beyond single-score similarity toward structured, dimension-aware evaluation is a promising direction, particularly for high-stakes domains such as medical text generation.

Limitations

This work provides a proof-of-concept for multidimensional semantic similarity, but it has several limitations. First, our synthetic MedVariations dataset is constructed using controlled transformations, which simplifies real-world semantic variability and may not fully capture the complexity of naturally occurring clinical text. As a result,

performance on this dataset may overestimate the separability of semantic dimensions. Second, although we include a more realistic evaluation using MedNLP-CHAT, the model is not explicitly trained on such diverse, human-generated data, which leads to reduced dimensional separation in that setting. This suggests limited robustness under domain shift and highlights the need for more diverse training data. Third, both datasets perturb exactly one dimension at a time, whereas real-world generation errors often entangle several dimensions at once (e.g., a confabulated entity introduced under hedged factuality); whether the heads remain calibrated under such composed perturbations is untested. Fourth, the set of semantic dimensions (entity, relation, temporal, factuality) is not exhaustive. Other clinically relevant aspects of meaning, such as uncertainty granularity or causal reasoning, are not explicitly modeled. Finally, while MD-BLEURT is computationally efficient compared to LLM-based judges, it still inherits biases from the underlying BLEURT encoder and depends on the quality of the synthetic supervision signal.

7 Acknowledgments

This work was supported by the Federal Ministry of Research, Technology and Space (BMFTR) through the project **VERANDA** (16KIS2048).

References

Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, Shohei Hisada, Tomohiro Nishiyama, Lenard Paulo Tamayo, Jingnan Xiao, Axalia Levenchaud, Pierre Zweigenbaum, Christoph Otto, Jerycho Pasniczek, Philippe Thomas, Nathan Pohl, Wiebke Duettmann, Lisa Raithel, and Roland Roller. 2025. NTCIR-18 MedNLP-CHAT Determining Medical, Ethical and Legal Risks in Patient-Doctor Conversations: Task Overview. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo, Japan. NII Institutional Repository.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Hongliu Cao. 2025. Semantic adapter for universal text embeddings: Diagnosing and mitigating negation blindness to enhance universality. In *ECAI 2025 - 28th European Conference on Artificial Intelligence*, pages 4137–4144. IOS Press.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 Task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

Iker De la Iglesia, Iakes Goenaga, Johanna Ramirez-Romero, Jose Maria Villa-Gonzalez, Josu Goikoetxea, and Ander Barrena. 2025. Ranking Over Scoring: Towards Reliable and Robust Automated Evaluation of LLM-Generated Medical Explanatory Arguments. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9456–9471, Abu Dhabi, UAE. Association for Computational Linguistics.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, and 37 others. 2021. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 96–120, Online. Association for Computational Linguistics.

Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. 2006. Negation, Contrast and Contradiction in Text Processing. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, page 755–762. AAAI Press.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Hans Ole Hatzel, Fynn Petersen-Frey, Tim Fischer, and Chris Biemann. 2023. Dimensions of Similarity: Towards Interpretable Dimension-Based Text Similarity. In *ECAI 2023 - 26th European Conference on Artificial Intelligence*, Frontiers in Artificial Intelligence and Applications, pages 1004–1011. IOS Press.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the Blind Spots of Model-Based Evaluation Metrics for Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Christin Beck, and Maribel Romero. 2022. Negation, Coordination, and Quantifiers in Contextualized Language Models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3074–3085, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, and 8 others. 2025. Medical Hallucinations in Foundation Models and Their Impact on Healthcare. *arXiv*, abs/2503.05777.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

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.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

David Mezzetti. 2026. PubMedBERT Embeddings: Domain-Specific Medical Text Representations for Semantic Search. `https://github.com/neuml/papers/blob/master/pubmedbert-embeddings/pubmedbert-embeddings.pdf`. Accessed: 2026-04-28.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

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.

Ellie Pavlick and Chris Callison-Burch. 2016. So-Called Non-Subsective Adjectives. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 114–119, Berlin, Germany. Association for Computational Linguistics.

Sanjana Ramprasad and Byron C Wallace. 2025. Do Automatic Factuality Metrics Measure Factuality? A Critical Evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, 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.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. EmbeddingGemma: Powerful and Lightweight Text Representations. *Preprint*, arXiv:2509.20354.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *Preprint*, arXiv:1904.09675.

Xiechi Zhang, Zetian Ouyang, Linlin Wang, Gerard De Melo, Zhu Cao, Xiaoling Wang, Ya Zhang, Yanfeng Wang, and Liang He. 2025. AutoMedEval: Harnessing Language Models for Automatic Medical Capability Evaluation. In *Proceedings of the*

63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6272–6285, Vienna, Austria. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, Zidu Xu, Yuen-Hei Chung, Yiyun Xing, Meng-Han Tsai, Emma Schaffer, Yucheng Shi, Ninghao Liu, Zirui Liu, and Rui Zhang. 2025. Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*, 9(1):34.

A Data Generation: Prompts & Workflow

Table 6 shows the prompt used to construct the anchor texts for MedVariations; due to space constraints, the remaining prompts are available in the repository[2]. The workflow for Version 1 of data generation is shown in the Figure 2. After completing Version 1, Version 2 was generated using a single prompt for all paraphrases and variations.

Tables 4 and 5 summarize the distribution of variation categories across dataset splits. The MedVariations dataset was divided into train, development, and test sets, while MedNLP-CHAT was used exclusively for evaluation. The statistics show a relatively balanced distribution of semantic perturbation types, with Temporal and Factuality variations constituting the largest portions due to their ordered divergence structure and multiple shift combinations.

The following example illustrates how controlled semantic variations are generated from an anchor sentence in the MedNLP-CHAT dataset across different perturbation categories, including entity, relation, temporal, and factuality modifications.

- **Anchor Text:** Persistent nausea, vomiting, and diarrhea can lead to complications in kidney transplant recipients. Get help, at the latest the day after tomorrow, if symptoms persist.
- **Entity Variation:** Persistent nausea, vomiting, and diarrhea can lead to complications in liver transplant recipients. Get help, at the latest the day after tomorrow, if symptoms persist.
- **Relation Variation:** Persistent nausea, vomiting, or diarrhea can lead to complications in kidney transplant recipients. Get help, at the latest the day after tomorrow, if symptoms persist.
- **Temporal Variations:**
  - Persistent nausea, vomiting, and diarrhea could lead to complications in kidney transplant recipients. Get help, at the latest the day after tomorrow, if symptoms persist.
  - Persistent nausea, vomiting, and diarrhea will lead to complications in kidney transplant recipients. Get help, at the latest the day after tomorrow, if symptoms persist.
- **Factuality Variations:**
  - Persistent nausea, vomiting, and diarrhea leads to complications in kidney transplant recipients. Get help, at the latest the day after tomorrow, if symptoms persist.
  - Persistent nausea, vomiting, and diarrhea is unlikely to lead to complications in kidney transplant recipients. Get help, at the latest the day after tomorrow, if symptoms persist.
  - Persistent nausea, vomiting, and diarrhea does not lead to complications in kidney transplant recipients. Get help, at the latest the day after tomorrow, if symptoms persist.

[2] https://github.com/akjuneja/SemEvalVar

Split	Anchor	Para.	Entity	Rel.	Temp.	Fact.
Train	1200	2400	1149	800	2346	3218
Dev	100	200	97	71	194	300
Test	200	400	186	127	388	597

Table 4: MedVariations dataset statistics.

Split	Anchor	Para.	Entity	Rel.	Temp.	Fact.
Test	212	424	196	210	422	633

Table 5: MedNLP-CHAT dataset statistics.

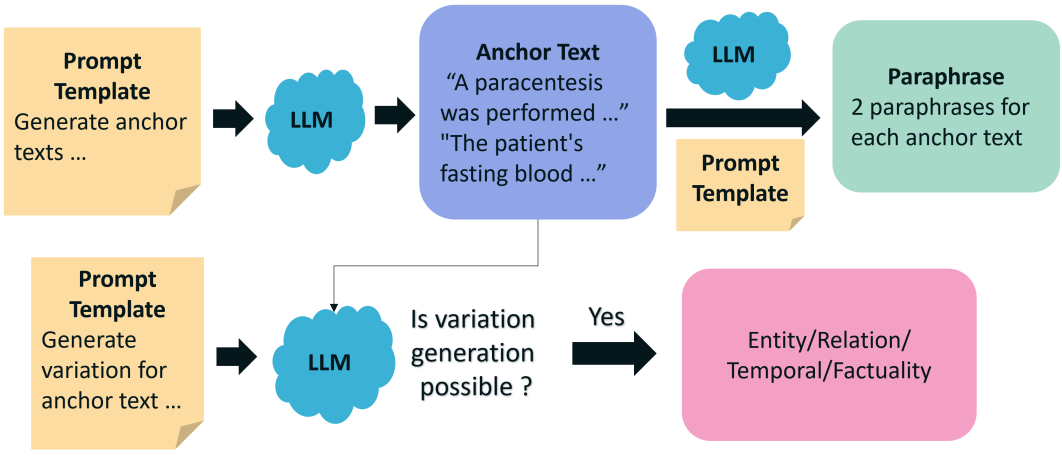

Figure 2: Overview of the synthetic data generation pipeline. Prompt templates are first used to generate anchor texts with an LLM. Each anchor text is then paraphrased to create semantically equivalent variants, while additional controlled variations are generated along entity, relation, temporal, and factual dimensions.

Your task is to generate short, plausible medical texts. Crucial Rules for each text:

1. Each text must be exactly 2 or 3 sentences long.
2. Each sentence should be simple and direct.
3. Each text must contain at least one “handle” that could be swapped to change the meaning.
4. Each example must be unique in wording and structure.

Part 1: A Rich Palette of Handles
The entire collection of texts you generate must contain a mix of the following "handles".
The examples provided for each category are a guide; you are encouraged to generate other related variations to ensure maximum diversity. Do NOT try to force all handles into every single example.
Examples:
Dosages (10mg, 2 pills), Frequencies (twice daily), Lab Values (A1C of 7.2%, WBC count of 15), measurements (lesion is 3cm), Ages (65-year-old male), Body Parts & Location (left arm, right kidney), Medications (Lisinopril, Metformin), Diagnoses (pneumonia, diabetes), Procedures (MRI, biopsy) Test Results (positive, negative), Patient State (stable, unstable), Progress (improving, worsening), Findings (normal, abnormal, clear). Specific Times (Monday, yesterday), Sequence (before, after), Duration (for 5 days, since last week, until Friday), Tense (was, is, will).

Part 2: Seeding for Structural Contradictions
In addition to the simple handles, ensure a good portion of the texts include sentences with clear relational structures. These "structural primitives" are essential for creating complex contradictions later.
Examples:
Cause-and-Effect: “...dehydration was caused by the flu." or “His knee pain is a result of the arthritis."
Purpose/Reason: “A CT scan was ordered to rule out a fracture." or “He takes Metformin for his diabetes."
Conditional Logic: “If symptoms worsen, the patient should return." or “Continue the medication unless a rash develops."

Output Requirements:
Generate a JSON list containing exactly {num_to_generate} unique examples based on these comprehensive rules.
Your output MUST be a single, valid JSON object which is a list of strings. Do not include explanations or any text outside the JSON.
Example format: ["text 1 ...", "text 2 ...", ...]

Table 6: Anchor generation prompt.

B LLM-as-a-Judge Evaluation

We used the prompt shown in Table 7 to instruct both open-source and proprietary LLMs to assign similarity scores to sentence pairs. Using a shared prompting framework ensured consistency across models during evaluation. The prompts incorporate scoring guidelines derived from the data annotation schema used for training MD-BLEURT, enabling fair and standardized comparisons across evaluation models.

You are an impartial evaluation metric for Medical Natural Language Generation.
You will be given two texts: Reference and Candidate.

Your task is to evaluate how well the Candidate aligns with the Reference using the rubric below.
Return ONLY a single integer score from 1 to 5. Do NOT include any explanation or additional text.

Evaluation Dimensions
Evaluate based on:

1. Clinical Entities

   (PROBLEM, MEDICATION, PROCEDURE, DOSE/VALUE, UNIT, ROUTE, FREQUENCY, LATERALITY, SITE, STAGE, EXPERIENCER)

2. Relations

   (COMPARATOR, CAUSAL, EVENT_ORDER, AND/OR)

3. Temporal Meaning

   PAST = already occurred
   PRESENT = ongoing/current
   FUTURE = not yet occurred
   Small Shift = PAST ↔ PRESENT, PRESENT ↔ FUTURE
   Large Shift = PAST ↔ FUTURE

4. Factuality

   - NEGATED = explicitly false
   - DOUBTFUL = uncertain/unlikely
   - POSSIBLE = plausible
   - FACT = asserted true
   - Factual distance:
   - Single level: NEGATED ↔ DOUBTFUL ↔ POSSIBLE ↔ FACT
   - Double level: NEGATED ↔ POSSIBLE, DOUBTFUL ↔ FACT
   - Triple level: NEGATED ↔ FACT

Scoring Rubric

5 = Perfect match

- Meaning identical (true paraphrase: wording may differ, meaning unchanged)
- All entities and attributes preserved
- Same clinical relations
- Same temporal meaning
- Same factuality

4 = Minor deviation

- Single-level factual shift (FACT ↔ POSSIBLE, POSSIBLE ↔ DOUBTFUL, DOUBTFUL ↔ NEGATED)
- OR small temporal shift (PAST ↔ PRESENT, PRESENT ↔ FUTURE)
- Entities and relations preserved or nearly identical

3 = Major contradiction

- Double-level factual shift (NEGATED ↔ POSSIBLE, DOUBTFUL ↔ FACT)
- OR large temporal shift (PAST ↔ FUTURE)
- OR incorrect or conflicting clinical entities
- OR incorrect relation that changes core meaning

2 = Major deviation

- Triple-level factual shift (FACT ↔ NEGATED)

1 = Unrelated or nonsensical

- No meaningful overlap
- OR completely incorrect clinical meaning

Table 7: Prompt template and scoring rubric used for LLM-as-a-Judge evaluation.

C Similarity Score Assignment

Sentence pairs used for MD-BLEURT training were annotated using a rule-based similarity scoring framework based on controlled semantic perturbations. The scoring scheme assigns penalties proportional to the degree of meaning divergence introduced by each variation type.

Entity and Relation perturbations are modeled as single-step deviations with a higher penalty strength ($\mu = 0.5$), since these modifications often introduce substantial semantic changes that directly alter the meaning of a sentence. In contrast, Temporal and Factuality perturbations involve ordered semantic transitions with varying levels of divergence. To model these gradual shifts, we use a lower penalty strength ($\mu = 0.33$) together with variable semantic distances (δ).

Table 8 summarizes the shift-divergence assignments used for Temporal and Factuality variations, covering all valid ordering transitions and their corresponding semantic distances.

Variation	**Anchor**	**Shift Divergence (δ)**
Temporal	Past	Past → Present (1)
	Past	Past → Future (2)
	Present	Present → Past (1)
	Present	Present → Future (1)
	Future	Future → Present (1)
	Future	Future → Past (2)
Factuality	Fact	Fact → Possible (1)
	Fact	Fact → Doubtful (2)
	Fact	Fact → Negated (3)
	Possible	Possible → Fact (1)
	Possible	Possible → Doubtful (1)
	Possible	Possible → Negated (2)
	Doubtful	Doubtful → Possible (1)
	Doubtful	Doubtful → Negated (1)
	Doubtful	Doubtful → Fact (2)
	Negated	Negated → Doubtful (1)
	Negated	Negated → Possible (2)
	Negated	Negated → Fact (3)

Table 8: The Temporal (Past, Present, Future) and Factuality (Fact, Possible, Doubtful, Negated) categories are ordered, and the scores are expected to shift from the current score in the ordered way.

Metric	**Time (sec)***
Gemma-4-E2B-it	0.24
Qwen3.5-4B	0.37
Llama-3.3-70B-Instruct-AWQ	0.24
Qwen2.5-72B-Instruct-AWQ	0.25
Gemini-3.1-flash-lite	1.41
Gemini-3.1-pro	3.60
MD-BLEURT	**0.01**

Table 9: Average inference time per sentence pair (*10 samples).

D Runtime Analysis

We conducted runtime experiments to compare the inference efficiency of MD-BLEURT against several high-performing LLMs. All open-source LLMs were served using the vLLM API on an NVIDIA H100 GPU. MD-BLEURT was also evaluated on the same hardware setup to ensure a fair comparison. Proprietary Gemini models were accessed through Vertex AI. Table 9 reports the average inference time per sentence pair.

E Complete Experimental Results

Tables 10 and 11 report the average test-set performance on the MedVariations dataset for Version 1 and Version 2, respectively. The evaluation includes a diverse set of baseline models spanning surface-level similarity metrics, sentence embedding approaches, semantic evaluation models, and both small and large language models.

Similarly, Tables 12 and 13 present the corresponding results on the MedNLP-CHAT test set for Version 1 and Version 2. These tables provide a comprehensive comparison across competitive evaluation models to assess their performance on medically grounded semantic variations.

Metric	Paraphrase	Entity	Relation	Temporal	Factuality
BLEU	0.79	0.88	0.88	0.87	0.85
ROUGE-L	0.89	0.95	0.95	0.94	0.94
METEOR	0.92	0.95	0.96	0.96	0.97
MoverScore	0.85	0.89	0.90	0.89	0.87
BERTScore_F1	0.98	0.99	0.99	0.99	0.99
BLEURT	0.89	0.85	0.89	0.91	0.84
NegBLEURT	0.59	0.45	0.50	0.75	0.40
All-MiniLM-L6-v2	0.97	0.92	0.99	0.99	0.98
PubMedBERT-Base-Embeddings	0.99	0.97	0.99	0.99	0.99
Qwen3-Embedding-0.6B	0.98	0.95	0.98	0.99	0.97
EmbeddingGemma-300m	0.95	0.96	0.98	0.97	0.97
Gemma-4-E2B-it	0.97	0.67	0.70	0.78	0.55
Qwen3.5-4B	1.00	0.54	0.57	0.86	0.49
Llama-3.3-70B-Instruct-AWQ	1.00	0.60	0.62	0.81	0.42
Qwen2.5-72B-Instruct-AWQ	1.00	0.64	0.61	0.79	0.55
Gemini-3.1-flash-lite	1.00	0.52	0.53	0.72	0.55
Gemini-3.1-pro	1.00	0.51	0.46	0.74	0.51
MD-BLEURT (Entity)	0.97	**0.51**	0.98	1.01	1.01
MD-BLEURT (Relation)	0.98	0.92	**0.51**	0.91	0.71
MD-BLEURT (Temporal)	0.99	1.02	0.98	**0.69**	0.99
MD-BLEURT (Factuality)	0.99	1.00	0.86	0.95	**0.46**

Table 10: Average similarity scores across semantic variation categories on the MedVariations test set.

Metric	Paraphrase	Entity	Relation	Temporal	Factuality
BLEU	0.40	0.30	0.31	0.36	0.37
ROUGE-L	0.66	0.60	0.61	0.63	0.65
METEOR	0.76	0.69	0.71	0.74	0.76
MoverScore	0.73	0.70	0.71	0.72	0.72
BERTScore_F1	0.95	0.94	0.94	0.94	0.94
BLEURT	0.81	0.76	0.78	0.80	0.76
NegBLEURT	0.45	0.32	0.35	0.49	0.31
All-MiniLM-L6-v2	0.94	0.87	0.94	0.94	0.94
PubMedBERT-Base-Embeddings	0.96	0.93	0.95	0.96	0.96
Qwen3-Embedding-0.6B	0.96	0.92	0.94	0.96	0.94
EmbeddingGemma-300m	0.89	0.86	0.88	0.88	0.88
Gemma-4-E2B-it	0.95	0.68	0.76	0.89	0.62
Qwen3.5-4B	1.00	0.55	0.71	0.98	0.57
Llama-3.3-70B-Instruct-AWQ	1.00	0.67	0.74	0.97	0.56
Qwen2.5-72B-Instruct-AWQ	0.99	0.66	0.69	0.91	0.59
Gemini-3.1-flash-lite	1.00	0.54	0.64	0.89	0.56
Gemini-3.1-pro	1.00	0.51	0.52	0.86	0.50
MD-BLEURT (Entity)	0.94	**0.52**	0.96	0.97	1.03
MD-BLEURT (Relation)	0.95	0.91	**0.58**	0.96	0.73
MD-BLEURT (Temporal)	0.84	0.92	0.88	**0.69**	0.88
MD-BLEURT (Factuality)	1.00	0.99	0.94	1.01	**0.57**

Table 11: Average similarity scores on the MedVariations Version 2 test set with additional paraphrastic rewrites applied to semantic variations.

Metric	Paraphrase	Entity	Relation	Temporal	Factuality
BLEU	0.92	0.97	0.97	0.97	0.95
ROUGE-L	0.95	0.99	0.99	0.99	0.98
METEOR	0.96	0.99	0.99	0.99	0.99
MoverScore	0.91	0.95	0.96	0.95	0.93
BERTScore_F1	0.99	1.00	1.00	1.00	0.99
BLEURT	0.92	0.90	0.93	0.92	0.89
NegBLEURT	0.45	0.47	0.46	0.53	0.39
All-MiniLM-L6-v2	0.99	0.94	1.00	1.00	0.99
PubMedBERT-Base-Embeddings	0.99	0.97	1.00	1.00	1.00
Qwen3-Embedding-0.6B	0.99	0.95	0.99	0.99	0.99
EmbeddingGemma-300m	0.97	0.98	0.99	0.99	0.98
Gemma-4-E2B-it	0.99	0.67	0.86	0.92	0.67
Qwen3.5-4B	1.00	0.59	0.77	0.95	0.58
Llama-3.3-70B-Instruct-AWQ	1.00	0.63	0.80	0.95	0.55
Qwen2.5-72B-Instruct-AWQ	1.00	0.64	0.78	0.86	0.63
Gemini-3.1-flash-lite	1.00	0.57	0.74	0.84	0.57
Gemini-3.1-pro	1.00	0.52	0.53	0.77	0.56
MD-BLEURT (Entity)	0.99	**0.70**	0.99	0.99	0.98
MD-BLEURT (Relation)	0.94	0.94	**0.74**	0.92	0.78
MD-BLEURT (Temporal)	1.00	0.99	0.95	**0.90**	0.99
MD-BLEURT (Factuality)	0.96	1.00	0.92	0.95	**0.70**

Table 12: Average similarity scores across semantic variation categories on the MedNLP-CHAT test set.

Metric	Paraphrase	Entity	Relation	Temporal	Factuality
BLEU	0.30	0.28	0.30	0.30	0.30
ROUGE-L	0.55	0.54	0.55	0.55	0.55
METEOR	0.63	0.61	0.63	0.63	0.63
MoverScore	0.69	0.69	0.69	0.69	0.69
BERTScore_F1	0.93	0.93	0.93	0.93	0.93
BLEURT	0.77	0.75	0.76	0.77	0.75
NegBLEURT	0.21	0.19	0.20	0.22	0.17
All-MiniLM-L6-v2	0.94	0.89	0.95	0.95	0.94
PubMedBERT-Base-Embeddings	0.96	0.93	0.96	0.96	0.96
Qwen3-Embedding-0.6B	0.93	0.90	0.94	0.94	0.93
EmbeddingGemma-300m	0.88	0.87	0.88	0.88	0.87
Gemma-4-E2B-it	0.99	0.79	0.94	0.98	0.83
Qwen3.5-4B	1.00	0.79	0.94	1.00	0.81
Llama-3.3-70B-Instruct-AWQ	1.00	0.78	0.93	1.00	0.79
Qwen2.5-72B-Instruct-AWQ	0.99	0.71	0.89	0.98	0.75
Gemini-3.1-flash-lite	1.00	0.61	0.86	0.99	0.66
Gemini-3.1-pro	1.00	0.54	0.76	0.94	0.62
MD-BLEURT (Entity)	0.93	**0.68**	0.92	0.93	0.94
MD-BLEURT (Relation)	0.83	0.84	**0.75**	0.84	0.73
MD-BLEURT (Temporal)	0.83	0.89	0.85	**0.81**	0.86
MD-BLEURT (Factuality)	0.95	0.97	0.96	0.95	**0.88**

Table 13: Average similarity scores on the MedNLP-CHAT test set (Version 2) with additional paraphrastic rewrites applied to semantic variations.

F Pareto Frontier Analysis

This section presents Pareto frontier analyses comparing average evaluation performance against average inference time across different metric families and model scales. The plots illustrate the trade-off between evaluation quality and computational efficiency, highlighting models that achieve strong semantic evaluation performance while maintaining lower inference costs.

Figure 3 presents the Pareto frontier results for the MedVariations dataset across both Version 1 and Version 2. The results show that MD-BLEURT consistently remains on or near the Pareto frontier across multiple variation categories, demonstrating strong evaluation performance while maintaining competitive inference efficiency compared to other LLM-based evaluation metrics.

A similar trend is observed for the MedNLP-CHAT dataset, as shown in Figure 4, where MD-BLEURT achieves favorable performance–efficiency trade-offs across both dataset versions.

Figure 3: Pareto frontier plots for the MedVariations dataset across Version 1 (top) and Version 2 (bottom), comparing average evaluation performance against average inference time for different evaluation models.

Figure 4: Pareto frontier plots for the MedNLP-CHAT dataset across Version 1 (top) and Version 2 (bottom), comparing average evaluation performance against average inference time for different evaluation models.