Electronics and Information Technologies, 2026, 34, 51–64
http://publications.lnu.edu.ua/collections/index.php/electronics/index

UDC: 004.89

SEMANTIC TEXT SIMILARITY ESTIMATION ACROSS DIVERSE LARGE LANGUAGE MODELS

Vitalii Parubochyi**, Oleksandr Chmykhalo, Oleh Husak, Roman Shuvar
System Design Department,
Ivan Franko National University of Lviv,
50 Drahomanova Str., Lviv, 79005, Ukraine

Parubochyi, V., Chmykhalo, O., Husak, O., & Shuvar, R. (2026). Semantic Text Similarity Estimation Across Diverse Large Language Models. *Electronics and Information Technologies*, 34, 51–64. https://doi.org/10.30970/eli.34.3

ABSTRACT

Background. Semantic Text Similarity (STS) estimation is one of the useful proxy-based approaches for analyzing semantic differences between responses generated by various large language models (LLMs) used in AI chats and AI agents. Various studies in recent years have attempted to estimate STS for AI-generated responses, but they typically focus on ground-truth estimation by comparing AI-generated responses with predefined answers that, in most cases, differ structurally from the generated texts. However, cross-model STS estimation can provide valuable insights into model similarity, independent of how well they align with the predefined answers.

Materials and Methods. In this paper, we estimate both types of STS, but primarily focus on the STS between responses generated by several selected LLMs from different providers. We build our estimation framework using a subset of Frequently Asked Questions (FAQs) in English from the MFAQ dataset, as it offers a wide range of topics and a clear, easy-to-use structure.

Results and Discussion. The results of our experiments are organized into two parts. In the first part, we estimate ground-truth STS scores using multiple metrics (Cosine Similarity, BERTScore, and Latent Semantic Analysis (LSA)) to analyze how closely the generated answers from each model match the ground-truth answer in the MFAQ dataset. In the second part, we estimated cross-model STS scores using the same metrics to analyze how closely models align with each other. Additionally, we measure average request latency and use publicly available information on model context size and input and output prices to explore possible relationships between these parameters and STS-based quality indicators.

Conclusion. Our results indicate relatively high semantic similarity between models from the same provider in the selected experimental setup and demonstrate that high ground-truth STS scores do not necessarily imply high inter-model semantic similarity within the selected benchmark and metrics, as seen in the analysis of cross-model STS scores. Additionally, we show that not all metrics are equally suitable for cross-model STS estimation, and that metrics like Cosine Similarity or LSA scores can provide better sensitivity than BERTScore or similar metrics.

Keywords: semantic text similarity (STS), large language model (LLM), MFAQ dataset, cosine similarity, BERTS core, latent semantic analysis (LSA)

INTRODUCTION

Extremely rapid growth and evolution of large language models (LLMs) in the context of AI chats and AI agents make the problem of selecting an appropriate LLM for everyday tasks increasingly important. In most cases, different research and surveys compare cost, response speed (or request latency, depending on how you look at it), and answer correctness to address this question. When cost and response speed can be estimated objectively, the estimation of answer correctness is usually less reliable. It strongly depends on the chosen dataset and the additional settings provided to the LLM, often in the form of an additional prompt. Additionally, the criteria for correctness can vary, as well as the application area.

Much research focuses on ground-truth estimation [1-5], comparing model responses with predetermined answers that are usually not generated and structurally don't correspond to LLM responses. In this case, the correctness of the responses is usually measured by text similarity between the ground truth and generated responses. In other cases, the correctness can be measured as the consistency [3-4], coherence [3], or comprehensiveness [3] of the LLM-generated responses.

When we look at the application areas, most research focuses on specific domains, such as medicine [1, 5], education [3], market research [6-8], etc. In contrast, only a few focus on general knowledge that covers a wide range of topics.

In our research, we aim to examine LLM response estimation more broadly and focus on general knowledge, including information from fields such as medicine, retail, and technology. Therefore, we decided to use STS as a proxy indicator of semantic proximity to reference answers, while recognizing that STS does not directly measure factual correctness, completeness, or safety. That reduces sensitivity to structural differences between the reference and generated responses, but also to estimate cross-model STS, which enables us to indirectly analyze similarities and differences between LLMs. For the estimation dataset, we chose the Multilingual FAQ Dataset (MFAQ) [9], which contains more than 6M question-answer pairs across 21 languages. Though we used only question-answer pairs in English for our estimation framework and restricted the MFAQ subset to 1000 pairs from the validation dataset due to constraints on available LLM resources, we still estimated STS on a sample of question-answer pairs drawn from 883 web domains.

The most complex part of our estimation framework was selecting LLMs. We aimed not only to choose models from different LLM providers that, in theory, should yield different results, but also to estimate the different model classes and service tiers, such as Fast, Thinking and Pro, that model providers typically use to accommodate different LLM choices. In addition, we should have accounted for constraints across different LLMs when choosing models to fully estimate our test set. Therefore, we analyzed the available LLM providers and the models they offer, including not only the latest models, which are usually more expensive and have stricter constraints, but also earlier model generations (see **Table 1**). Based on it and available LLM resources, we chose two LLM providers and their corresponding models for the different model classes and service tiers:

- OpenAI: GPT-5-mini, GPT-4.1;
- Anthropic: Claude Haiku 4.5, Claude Sonnet 4.6, Claude Opus 4.6.

Despite the growing number of benchmarks for LLMs, there remains a methodological gap between evaluating reference-based responses and benchmarking model behavior. Existing approaches typically evaluate how close a generated response is to a predefined reference, but they do not sufficiently explain how semantically similar responses are between different models, nor do they support model selection given latency and cost constraints. This poses a scientific challenge: to demonstrate an integrated framework for assessing both response adequacy and semantic consistency between models. Therefore, the scientific objective of this study is to develop and

Table 1. **Overview of LLM providers and their models by operational mode**

LLM Provider	Fast / Small	Thinking / Medium	Pro / Large
2026 – Present			
OpenAI (GPT-5 Era)	GPT-5.4-mini	GPT-5.4	GPT-5.4-pro
Google	Gemini 3 Flash	Gemini 3 Thinking	Gemini 3.1 Pro
Anthropic	Claude Haiku 4.5	Claude Sonnet 4.6	Claude Opus 4.6
Meta AI	Llama 4 Scout	Llama 4 Maverick	Llama 4 Behemoth
DeepSeek	DeepSeek-V3.2	DeepSeek-R2	DeepSeek-V3.2-Speciale
Mistral AI	Mistral Small 4	Mistral Medium 3.1	Mistral Large 3
xAI	Grok-4.1 Fast	Grok-4 Heavy	Grok-4 o3
2023 – 2025			
OpenAI (GPT-4 Era)	GPT-4o mini	GPT-4o / Turbo	OpenAI o1-preview
Google	Gemini 1.5 Flash	Gemini 1.5 Pro	Gemini 1.0 Ultra
Anthropic	Claude 3 Haiku	Claude 3.5 Sonnet	Claude 3 Opus
Meta AI	Llama 3 8B	N/A	Llama 3 70B
Mistral AI	Mistral 8x7B	N/A	Mistral Large
DeepSeek	DeepSeek-R1-Lite	N/A	DeepSeek-V2

experimentally validate an STS-based framework that combines reference-based comparison, inter-model comparison, and operational indicators for a set of LLMs evaluated on a general-domain FAQ dataset.

In the next sections, we provide a detailed overview of related works, analyze the advantages of semantic text similarity, and discuss the metrics that can be used to estimate it. In the last two sections, we provide a detailed description of our estimation framework, how we obtained our test set from the MFAQ subset, analyzed the results, and draw insights from them.

RELATED WORK

Since the release of ChatGPT in November 2022 [10], the massive adoption of LLMs in the form of AI chats and agents has affected many people's daily lives in many ways. And this is of not only practical interest, but also scientific interest, since the application of LLMs significantly changes both individual aspects of activity and entire industries.

In recent years, there has been a lot of research that has tried to estimate LLMs [1-8] in different ways, but all of them face the same issues:

- Extremely rapid evolution makes any research quickly outdated.
- Choosing a good estimation dataset is extremely hard, since we don’t know, what data was used to train these models. And most of the proposed datasets have become biased, since the next generation of models will very likely be fine-tuned on them.

- Choosing the right estimation criteria and metrics that represent them is also challenging. That's why most researchers choose different metrics or calculate multiple metrics for the same experiments.
- Model constraints limit or make independent research very costly. As a result, researchers tend to use limited datasets [2-5] or prefer open-weighted local or older free models [1-2, 4-5], which often cannot compete with cloud-based counterparts.

In addition to these issues, most research focuses on specific domains, such as medicine [1, 5], education [3], market research [6-8], etc. So, their results and datasets cannot be directly generalized to broad-domain LLM estimation. And even papers that don't focus on specific domains [2, 4] use specific data representations, such as role-play dialogues [2] or multiple responses and user ratings [4].

Moreover, most research focuses on ground-truth estimation [1-5], comparing model responses with predetermined answers that are usually not generated and structurally don't correspond to LLM responses. Such a comparison requires more sophisticated preprocessing and/or postprocessing. Preprocessing may include prompting the model to limit its answers and make them more structurally similar to the expected answers. Postprocessing requires more complex embedding methods to improve precision and reduce sensitivity to structural differences between the generated and expected answers.

We also conduct ground-truth estimation in our work, but our main assumption is based on cross-model STS. We hypothesize that models with high inter-model semantic similarity may produce responses that are semantically close under the selected benchmark and evaluation metrics but may differ in request latency, context size, input price, and output price. So, identifying groups of semantically similar models may support selecting a more efficient model on these metrics while maintaining relatively similar responses.

Thus, the reviewed literature suggests that existing studies rarely combine reference-based semantic adequacy, inter-model semantic similarity, and operational indicators within a single comparative framework. This gap defines the scientific task of this study: to develop and test a comparative framework for assessing the STS between different LLMs and their operating modes, using a common FAQ-based benchmark, multiple STS metrics, and operational indicators such as latency and token cost.

MATERIALS AND METHODS

Understanding how close two texts are to each other is a classical problem in natural language processing (NLP). However, even now, choosing the best metric remains an open question, and in many cases, different researchers can propose their own metrics to estimate text similarity [4, 16-22]. But different metrics may use different approaches, so it is important to understand exactly how each metric is calculated.

Two main approaches to measuring text similarity are lexical and semantic. The lexical similarity measures the degree to which the word sets of two given languages or texts are similar. The semantic similarity measures the degree to which the meanings of two text fragments are similar, rather than just matching words [11-12]. While lexical similarity is easier to calculate and interpret, it focuses solely on a text's lexical structure, ignoring its meaning. Therefore, lexical similarity metrics usually perform poorly when the structures of the texts being compared differ.

In contrast, semantic similarity is less sensitive to text structure and focuses on overall text meaning, allowing comparisons of texts with different structures and lengths that cover the same topic. But at the same time, semantic similarity is often confused with semantic relatedness [12], which measures the relationship between two words or concepts based on the closeness of their meanings [13]. For this reason, semantic similarity is more specific and can be considered as a particular case of semantic relatedness. However, semantic similarity estimation is more complex and requires additional text processing to extract

meaning. This makes semantic similarity estimation more challenging and requires more sophisticated methods.

There are a lot of different metrics that can be used for semantic estimation [12, 14-15], but in general, they can be divided into three categories:

- Corpus-based [12] or structure-based [14] metrics;
- Knowledge-based [12] or feature-based [14] metrics;
- Hybrid [14] metrics.

Though the boundaries between groups aren't always clear, and some research [15] doesn't distinguish between hybrid and feature-based metrics, each metric has its own advantages. After analyzing prior studies [4, 13-22], we selected three metrics that are suitable for our experimental purpose:

- Latent Semantic Analysis (LSA) [23] using the TF-IDF vectorization [24].
- Cosine similarity based on Sentence-BERT (SBERT) [18] embeddings;
- BERTScore [19].

Latent Semantic Analysis (LSA) [23] is a traditional natural language processing (NLP) technique that uses singular value decomposition (SVD) [25-27] to identify latent patterns and relationships between a set of documents and the terms and concepts they contain. As input to SVD, we provide text vectors generated using the TF-IDF (term frequency–inverse document frequency) vectorizer. For our implementation, we used truncated SVD [26] with 10 components.

Cosine similarity [12] is one of the most widely used methods for measuring similarity. For this, we convert text into high-dimensional vectors that capture context using Sentence-BERT (SBERT) [18]. Then we calculate the cosine of the angle between vectors, which measures similarity regardless of text length:

$$similarity = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

BERTScore [19] computes similarity using the contextualized token embeddings from the BERT model. It captures deep semantic nuances by matching tokens in the candidate and reference sentences rather than relying on surface-level overlaps.

In our estimation framework, we estimate two types of STS:

- Ground-truth STS score that measures the similarity between the ground-truth answer from the MFAQ subset and the answer generated by some LLM.
- Cross-model STS score that measures the similarity between answers generated by different LLMs.

While the ground-truth STS score provides a baseline based on the expected answer and allows evaluating how close the generated response is to the commonly provided answer, the cross-model STS score shows how close the answers of different LLMs are.

Combining these two perspectives allows us to compare models both by semantic proximity to reference responses and by semantic similarity between model outputs, together with descriptive latency and cost indicators.

RESULTS AND DISCUSSION

As mentioned earlier, we selected the 1000-pair English subset from the validation dataset of the Multilingual FAQ Dataset (MFAQ) [9] as the estimation dataset. The original English validation dataset contains 151825 question-answer pairs across 8178 unique web domains. To form our subset, we downloaded the corresponding Parquet file for the English validation dataset from Hugging Face (see the MFAQ official repository for more details) and randomly chose 1000 question-answer pairs. For random generation, we read the Parquet file as a pandas DataFrame in Python and used the DataFrame's built-in *sample*

method with *random_state* = 42 to select 1000 records. As a result, our estimation subset contained 1000 question-answer pairs from 883 unique web domains.

For ground-truth STS metrics, we calculate the LSA (Latent Semantic Analysis) score, the Cosine Similarity score, the BERTScore and the latency for each model, comparing its answers to the ground-truth answers from the MFAQ dataset. Cross-model STS metrics are estimated as the LSA score, the Cosine Similarity score, and the BERTScore for each pair of models. Additionally, to compare model costs and available context, we obtained the context size and the price in USD for model input and output per 1M tokens from the Price Per Token site [28] (see **Table 2**).

Table 2. **Context size and model input/output prices per 1M tokens for the researched models from the Price Per Token site [28]**

LLM	Context size	Input price, USD	Output price, USD
GPT-5-mini	400K	0.25	2.00
GPT-4.1	1.0M	2.00	8.00
Claude Haiku 4.5	200K	1.00	5.00
Claude Sonnet 4.6	1.0M	3.00	15.00
Claude Opus 4.6	1.0M	5.00	25.00

After running our estimation framework for each selected model, we grouped our results for ground-truth STS metrics by the average request latency (see **Table 3**), context size (see **Table 4**), input price (see **Table 5**), and output price (see **Table 6**).

Table 3. **The ground-truth STS metrics for the researched models, depending on the average request latency (in seconds)**

Model	Latency, s	Cosine Similarity Score	BERTScore	LSA Score
GPT-5-mini	25.677	0.5934	0.8267	0.6259
GPT-4.1	10.233	0.6060	0.8451	0.6439
Claude Haiku 4.5	7.359	0.4197	0.8145	0.4570
Claude Sonnet 4.6	7.863	0.4810	0.8280	0.5027
Claude Opus 4.6	8.310	0.4505	0.8252	0.4658

From the results obtained, we observe that some models from different providers achieve comparable reference-based STS scores. This may indicate partially similar semantic response patterns on the selected benchmark. However, no conclusions can be drawn about internal optimization priorities from these results alone. Interestingly, BERTScore places GPT-5-mini between Claude Sonnet 4.6 and Claude Opus 4.6 and reports lower STS than in GPT-4.1 (see **Table 3**).

Also, as shown in **Table 3**, the average request latency for GPT-5-mini is much higher than for other models, especially GPT-4.1, the larger model from the same provider. It may seem contradictory, since GPT-5-mini has much less context size and, intuitively, should

Table 4. **The ground-truth STS metrics for the researched models, depending on the context size in thousands**

Model	Context size, K	Cosine Similarity Score	BERTScore	LSA Score
GPT-5-mini	400	0.5934	0.8267	0.6259
GPT-4.1	1000	0.6060	0.8451	0.6439
Claude Haiku 4.5	200	0.4197	0.8145	0.4570
Claude Sonnet 4.6	1000	0.4810	0.8280	0.5027
Claude Opus 4.6	1000	0.4505	0.8252	0.4658

Table 5. **The ground-truth STS metrics for the researched models, depending on the input price per 1M tokens in USD**

Model	Input price, USD	Cosine Similarity Score	BERTScore	LSA Score
GPT-5-mini	0.25	0.5934	0.8267	0.6259
GPT-4.1	2.00	0.6060	0.8451	0.6439
Claude Haiku 4.5	1.00	0.4197	0.8145	0.4570
Claude Sonnet 4.6	3.00	0.4810	0.8280	0.5027
Claude Opus 4.6	5.00	0.4505	0.8252	0.4658

Table 6. **The ground-truth STS metrics for the researched models, depending on the output price per 1M tokens in USD**

Model	Output price, USD	Cosine Similarity Score	BERTScore	LSA Score
GPT-5-mini	2.00	0.5934	0.8267	0.6259
GPT-4.1	8.00	0.6060	0.8451	0.6439
Claude Haiku 4.5	5.00	0.4197	0.8145	0.4570
Claude Sonnet 4.6	15.00	0.4810	0.8280	0.5027
Claude Opus 4.6	25.00	0.4505	0.8252	0.4658

be faster. We manually investigated these two models and compared the results. First of all, we compared additional model properties using the Price Per Token site [28] and found that GPT-5-mini has much higher latency in generating the first output token, as measured

by Time to First Token (TTFT). For GPT-4.1, TTFT is only 0.61 s; for GPT-5-mini, it is 14.39 s. Additionally, the output speed for token generation in GPT-4.1 is higher than in GPT-5-mini, at 109 tokens/second, compared to 77 tokens/second in GPT-5-mini.

After manual investigation of the results, we also found that GPT-5-mini tends to generate longer responses, which, in combination with higher TTFT and lower token generation speed, results in approximately 2.51 times the latency of GPT-4.1.

Another interesting insight is that the more advanced and costly Claude Opus 4.6 provides semantically less similar answers compared to the cheaper Claude Sonnet 4.6. Although the reason for this requires further investigation, one possible explanation is that longer or more elaborate responses may receive lower similarity scores than shorter FAQ-style reference answers; however, this assumption requires additional verification.

Moreover, the comparison with the context size (see **Table 4**) clearly indicates that it does not reveal a clear relationship between the STS estimation and models with the same context size can have very different STS scores, and even some models with smaller context windows achieved higher STS scores than certain models with larger context windows in our experiment.

The input and output price comparison (see **Table 5-6**) shows that a high cost for model tokens doesn't guarantee high STS scores, and that using relatively less expensive models, such as GPT-4.1 and GPT-5-mini, can yield better STS than more expensive models.

The cross-model STS metrics are displayed as heat maps for each metric to visualise similarities across models (see **Fig. 1-3**).

The cross-model STS analysis yields several informative observations, especially regarding the metrics that can be used to estimate it. Our experiments show that, in the selected setup, BERTScore has limited discriminative power for cross-model comparison, since it produces very similar values for many model pairs (see **Fig. 2**). These results make it hard to use BERTScore to estimate cross-model STS, since these differences aren't practically meaningful or statistically significant and can be misinterpreted.

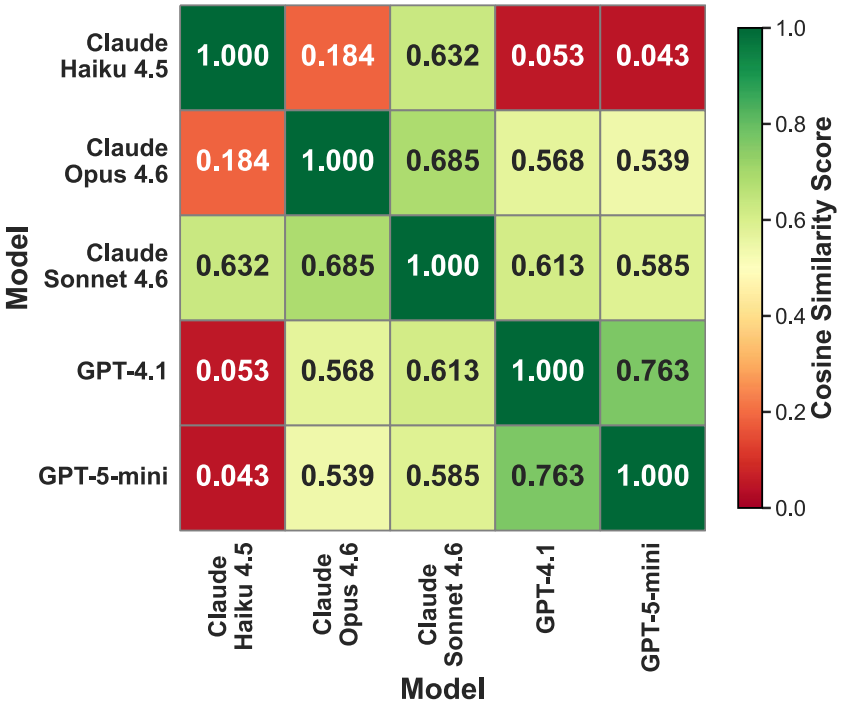

Fig. 1. The cross-model STS estimation using the Cosine Similarity Score

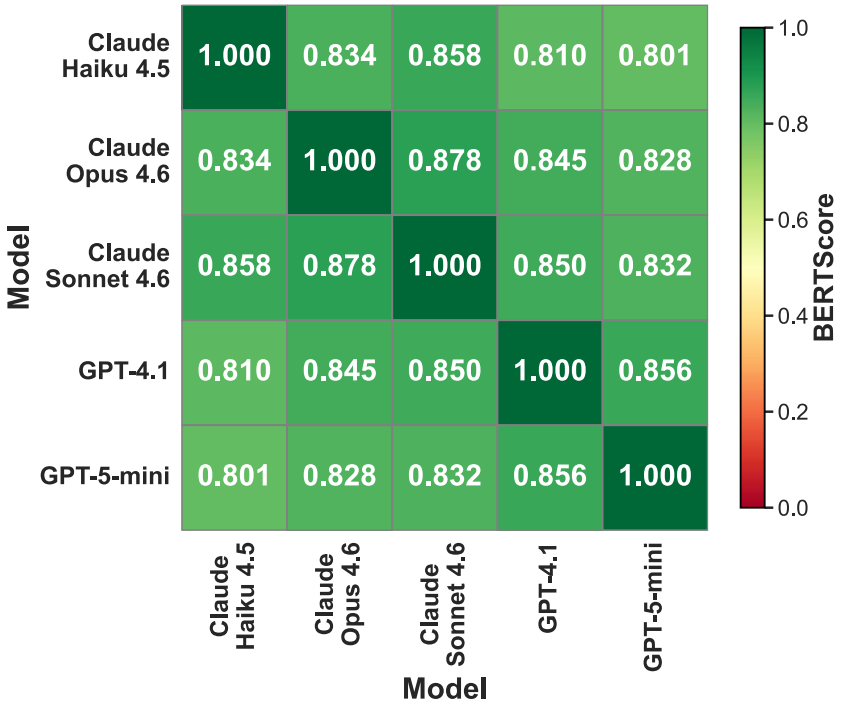

Fig. 2. The cross-model STS estimation using the BERTScore

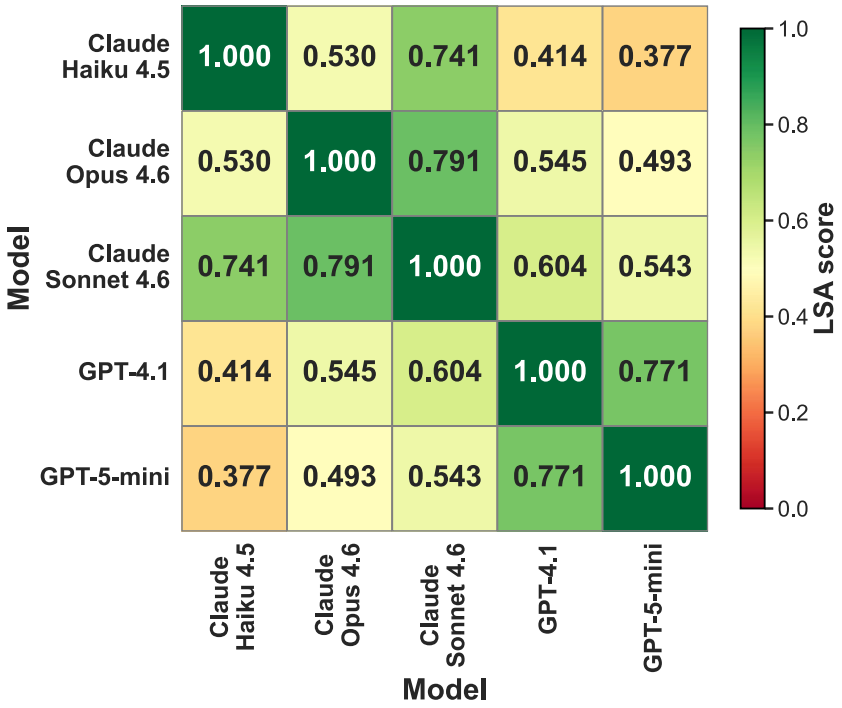

Fig. 3. The cross-model STS estimation using the LSA Score

In contrast, cosine similarity provides a wider spread of pairwise estimates and more clearly separates model pairs with relatively high and low similarity (see **Fig. 1**). At the

same time, the LSA score provides an intermediate level of separation between model pairs, clearly showing the high and low similarity of the models (see **Fig. 3**).

If we analyse the results indicate high semantic similarity between the two OpenAI models under the selected metrics (see **Fig. 1-3**). Anthropic models also show high similarity among themselves, but we can clearly see that Claude Sonnet 4.6 occupies an intermediate position between Claude Haiku 4.5 and Claude Opus 4.6 and is highly similar to both. However, the comparison between Claude Haiku 4.5 and Claude Opus 4.6 indicates lower semantic similarity than for other within-provider model pairs.

CONCLUSION

STS is an important metric for estimating the similarity of two texts, especially when we consider AI-generated texts and the variety of LLMs that can generate them. In this work, we presented the results of STS estimation for 5 LLMs from two providers (Anthropic and OpenAI). For our research, we developed the estimation framework based on the 1000-pair English subset from the MFAQ validation dataset.

In contrast to many reference-based studies, we estimated not only reference-based STS scores but also cross-model STS scores by comparing models. This extends the reviewed approaches by adding cross-model STS analysis, which provides additional insight into inter-model response similarity. However, as shown in the work, it strongly depends on the metric used, since different metrics yield different sensitivities for cross-model STS estimation. In our experimental setup, the LSA score and cosine similarity were the most informative metrics for cross-model comparison, whereas BERTScore yielded less sensitive and interpretable results.

Also, our results suggest that, in this experimental setting, models from the same provider are more semantically similar to each other than to models from another provider. We also showed that similar ground-truth estimates don’t necessarily imply high STS between the models, as evidenced by the cross-model STS scores.

Additionally, our experimental results showed that more expensive models did not consistently achieve the highest STS scores on the selected benchmark. Therefore, for the benchmark considered, some medium-cost or relatively inexpensive models may offer a reasonable balance between STS-based indicators, latency, and cost.

Despite the results obtained, the STS of AI-generated texts still requires further research using larger datasets and more models to provide a more complete picture. Also, in this research, we minimized any additional prompting from our side, but it can significantly affect the responses generated by models.

ACKNOWLEDGMENTS AND FUNDING SOURCES

The authors received no financial support for the research, writing, and/or publication of this paper.

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, [V.P., O.C., O.H., R.S.]; methodology, [V.P., O.C., O.H.]; software, [V.P., O.C.]; validation, [V.P., O.C., O.H.]; formal analysis, [V.P., O.C., O.H.].; investigation, [V.P., O.C., O.H.]; resources, [V.P., O.C., O.H.]; data curation, [V.P., O.C., O.H.]; writing –

original draft preparation, [V.P., O.C.]; writing – review and editing, [V.P., O.C., O.H., R.S.]; visualization, [V.P., O.C.] supervision, [V.P., R.S.]; project administration, [V.P., R.S.].

All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Sblendorio, E., Lo Cascio, A., Napolitano, D., Germini, F., Dentamaro, V., Piredda, M., & Cicolini, G. (2025). Assessing and Comparing Free Large Language Models' Responses to a Clinical Case: Accuracy, Safety, and Reliability. *Computational Intelligence Methods for Bioinformatics and Biostatistics*, CIBB 2024. Lecture Notes in Computer Science, 15276, 134-149. https://doi.org/10.1007/978-3-031-89704-7_11

[2] Lu, D., Jeuring, J., & Gatt, A. (2025). Evaluating LLM-Generated Versus Human-Authored Responses in Role-Play Dialogues. In *Proceedings of the 18th International Natural Language Generation Conference*, Association for Computational Linguistics, Hanoi, Vietnam, 20–40. https://doi.org/10.48550/arXiv.2509.17694

[3] Ding, X. et al. (2024). Evaluation of LLMs and Other Machine Learning Methods in the Analysis of Qualitative Survey Responses for Accessible Engineering Education Research. *2024 ASEE Annual Conference & Exposition*, Portland, Oregon, USA, 1-24. https://doi.org/10.18260/1-2--47360

[4] Wu, X., Lin, W., Akgul, O., & Bauer, L. (2025). Estimating LLM Consistency: A User Baseline vs Surrogate Metrics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Suzhou, China, 30530-30544. https://doi.org/10.18653/v1/2025.emnlp-main.1554

[5] Lee, P. C., Sharma, S. K., Motaganahalli, S., & Huang, A. (2023). Evaluating the Clinical Decision-Making Ability of Large Language Models Using MKSAP-19 Cardiology Questions. *JACC: Advances*, 2(9). https://doi.org/10.1016/j.jacadv.2023.100658

[6] Zhang, L., Wang, J., Wu, J., & Zhang, Z. (2026). RetailBench: Evaluating Long-Horizon Autonomous Decision-Making and Strategy Stability of LLM Agents in Realistic Retail Environments. *ArXiv preprint*. https://doi.org/10.48550/arXiv.2603.16453

[7] Brand, J., Israeli, A., & Ngwe, D. (2023). Using LLMs for Market Research. *Harvard Business School Working Paper*, No. 23-062, April 2023 (Revised October 2025). https://doi.org/10.2139/ssrn.4395751

[8] Wang, M., Zhang, D., & Zhang, H. (2024). Large Language Models for Market Research: A Data-augmentation Approach. *Social Science Research Network* (SSRN). https://doi.org/10.2139/ssrn.5057769

[9] De Bruyn, M., Lotfi, E., Buhmann, J., & Daelemans, W. (2021). MFAQ: a Multilingual FAQ Dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering, Association for Computational Linguistics*, Punta Cana, Dominican Republic, 1-13. https://doi.org/10.18653/v1/2021.mrqa-1.1.

[10] OpenAI. Introducing ChatGPT. *OpenAI*, November 30, 2022. Retrieved from: https://openai.com/index/chatgpt/

[11] Resnik, P. (1999). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11, 95-130. https://doi.org/10.1613/jair.514

[12] Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2015). *Semantic Similarity from Natural Language and Ontology Analysis. Synthesis Lectures on Human Language*

Technologies Series, 8(1). Springer Cham. ISBN: 978-3-031-01028-6. https://doi.org/10.2200/S00639ED1V01Y201504HLT027

[13] Feng, Y., Bagheri, E., Ensan, F., & Jovanovic, J. (2017). The state of the art in semantic relatedness: a framework for comparison. *Knowledge Engineering Review*, 32(10), 1-30. https://doi.org/10.1017/S0269888917000029

[14] Slimani, T. (2013). Description and Evaluation of Semantic Similarity Measures Approaches. *International Journal of Computer Applications*, 80, 25-33. https://doi.org/10.5120/13897-1851

[15] Haque, S., Eberhart, Z., Bansal, A., & McMillan, C. (2022). Semantic similarity metrics for evaluating source code summarization. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension (ICPC '22)*, Association for Computing Machinery, New York, NY, USA, 36–47. https://doi.org/10.1145/3524610.3527909

[16] Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. St., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics*, Brussels, Belgium, 169-174. https://doi.org/10.18653/v1/D18-2029

[17] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2018). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *ArXiv preprint*. https://doi.org/10.48550/arXiv.1705.02364

[18] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*, 3973-3983. https://doi.org/10.18653/v1/D19-1410

[19] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. *International Conference on Learning Representations*, 1-43. https://doi.org/10.48550/arXiv.1904.09675

[20] Fu, J., Ng, S.-K., Jiang, Z., & Liu, P. (2023). GPTScore: Evaluate as You Desire. *ArXiv preprint*. https://doi.org/10.48550/arXiv.2302.04166

[21] Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 9004-9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

[22] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 2511-2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

[23] Dumais, S. T. (2004). Latent Semantic Analysis. *Annual Review of Information Science and Technology*, 38, 188–230. https://doi.org/10.1002/aris.1440380105

[24] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, 24(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

[25] Golub G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5), 403-420. https://doi.org/10.1007/BF02163027
[26] Hansen, P. C. (1987). The truncated SVD as a method for regularization. *BIT Numerical Mathematics*, 27(4), 534-553. https://doi.org/10.1007/BF01937276
[27] Wall, M. E., Rechtsteiner, A., & Rocha, L. (2003). Singular Value Decomposition and Principal Component Analysis. In *A Practical Approach to Microarray Data Analysis*, 5, 91-109. https://doi.org/10.1007/0-306-47815-3_5
[28] Ventures, LLC (2026). LLM API Pricing 2026 – Compare 300+ AI Model Costs. Price Per Token. https://pricepertoken.com/

ОЦІНКА СЕМАНТИЧНОЇ ПОДІБНОСТІ ТЕКСТУ РІЗНИХ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Віталій Парубочий*  , ***Олександр Чмихало, Олег Гусак, Роман Шувар***
Кафедра системного проектування,
Львівський національний університет імені Івана Франка,
вул. М. Драгоманова, 50, Львів, 79005, Україна

АНОТАЦІЯ

Вступ. Оцінка семантичної подібності текстів (СПТ) є одним із корисних підходів на основі проксі-методів для аналізу семантичних відмінностей між відповідями, згенерованими різними великими мовними моделями (ВММ), що використовуються в чатах штучного інтелекту та агентами штучного інтелекту. Різні дослідження останніх років намагалися оцінити СПТ для відповідей, згенерованих штучним інтелектом, але вони зазвичай зосереджені на оцінці істинності шляхом порівняння відповідей, згенерованих штучним інтелектом, із заздалегідь визначеними відповідями, які в більшості випадків структурно відрізняються від згенерованих текстів. Однак, міжмодельна оцінка СПТ може надати цінну інформацію про подібність моделі, незалежно від того, наскільки добре вони узгоджуються із заздалегідь визначеними відповідями.

Матеріали та методи. У цій статті ми оцінюємо обидва типи СПТ, але, в першу чергу, зосереджуємося на СПТ між відповідями, згенерованими кількома вибраними ВММ від різних постачальників. Ми будуємо нашу систему оцінювання, використовуючи підмножину часто задаваних питань (ЧЗП, англ. FAQ – Frequently Asked Questions) англійською мовою з набору даних MFAQ, оскільки вона пропонує широкий спектр тем та чітку, просту у використанні структуру.

Результати. Результати наших експериментів складаються з двох частин. У першій частині ми оцінюємо базові показники СПТ, використовуючи кілька метрик (косинусна подібність, BERTScore та латентно-семантичний аналіз (ЛСА)), щоб проаналізувати, наскільки близько відповіді, згенеровані кожною моделлю, співвідносяться з базовими відповідями набору даних MFAQ. У другій частині ми оцінюємо міжмодельні показники СПТ, використовуючи ті самі метрики, щоб проаналізувати, наскільки близько моделі узгоджуються одна з одною. Крім того, ми вимірюємо середню затримку запиту та використовуємо загальнодоступну інформацію про розмір контексту моделі та ціни вхідних і вихідних даних, щоб дослідити можливі зв'язки між цими параметрами та показниками якості на основі СПТ.

Висновки. Наші результати вказують на відносно високу семантичну подібність між моделями від одного постачальника у вибраній експериментальній установці та

демонструють, що високі показники СПТ, що відповідають реальним даним, не обов'язково означають високу міжмодельну семантичну подібність у межах вибраного тесту оцінки та метрик, як видно з аналізу міжмодельних показників СПТ. Крім того, ми показуємо, що не всі метрики однаково придатні для міжмодельної оцінки СПТ, і що такі метрики, як косинусної подібності або ЛСА, можуть забезпечити кращу чутливість, ніж BERTScore або подібні метрики.

Ключові слова: семантична подібність тексту (СПТ), велика мовна модель (ВММ), набір даних MFAQ, косинусна подібність, BERTScore, латентно-семантичний аналіз (ЛСА)

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
06 April, 2026	16 May, 2026	18 May, 2026	29 June, 2026