International Journal
Information Technology Applications (ITA)

Volume 14, Number 1, July 2025

# BRIDGING THE LANGUAGE GAP: EVALUATING AND ENHANCING SLOVAK LANGUAGE SUPPORT IN LARGE LANGUAGE MODELS

**Patrik Skovajsa**

***Abstract:***

*This study investigates the current level of Slovak-language support in large language models (LLMs) and proposed practical pathways toward high-quality, resource-efficient deployment. I benchmarked several state-of-the-art open-source and commercial LLMs on a newly created set of 100 Slovak questions covering grammar, semantics, style, slang, translation, and complex constructions. I evaluated the answers automatically with OpenAI GPT-4o-mini. Results show that Google Gemma 3 27 B achieves near parity with GPT-4o while running on a single high-end GPU, outperforming LLaMA 3.1 70 B by 27 percentage points in overall quality and cutting latency by a factor of four. My findings highlight Gemma 3 27 B as the best current trade-off for Slovak, while underscoring the strategic need for a dedicated Slovak LLM built on open resources.*

***Keywords:***

*Slovak language, large language models, language evaluation, Gemma 3, natural language processing.*

## Introduction

Slovak, like many typologically related languages, presents unique challenges when deploying large language models (LLMs) without task or language-specific fine-tuning. Current open-source models such as **Meta LLaMA 3.1** with 70 billion parameters or **Mistral-Large** demand substantial computational resources in both training and inference. A recent leap in model quality most noticeably in **Google's Gemma 3** family, and especially the 27 billion-parameter varian suggests that these obstacles can be mitigated. Gemma 3 supports more than 140 languages and was explicitly designed for efficient single-machine deployment.

Multilingual encoder–decoder architectures like **mT5** already cover up to 101 languages, including Slovak. Without any additional adaptation they perform competitively in downstream tasks such as text classification, structured prediction, and open-domain question answering [1].

For the Slavic language group there exist models that have been tuned explicitly, for example **XLM-R**, which achieves state-of-the-art results in named-entity recognition, normalization, and entity linking across Czech, Polish, and Russian. With $F_1$ scores reaching 0.914 for Czech, these results indicate a strong likelihood of comparable success on Slovak data sets [2].

A fully Slovak-centric approach is represented by **SlovakBERT**, a RoBERTa-style transformer trained on a large Slovak web corpus. SlovakBERT attains excellent performance in morpho-syntactic tagging, sentiment analysis, document classification, and semantic textual similarity, and therefore constitutes a valuable building block for the Slovak NLP community [3].

Broader multilingual models such as **mGPT**, trained on Wikipedia and the C4 corpus in 61 languages, have demonstrated credible zero-shot performance in low-resource settings.

Although Slovak was not a primary training target, the model's cross-lingual capabilities make it attractive for general NLP tasks that do not warrant Slovak-specific fine-tuning [4].

Among the models directly optimised for Slovak, **mistral-sk-7** a fine-tuned derivative of Mistral-7B trained on the Araneum Slovacum VII Maximum corpu provides a solid foundation for further customisation [5]. Iteratively fine-tuning such medium-sized backbones, including Mistral and Gemma, was emerged as a pragmatic route toward high-quality Slovak LLMs.

Progress on evaluation resources has kept pace. **SK-QuAD**, the first manually curated Slovak question-answering data set, contains over 91 000 factoid questions aligned with the SQuAD v2.0 format. By providing unanswerable questions and "plausible distractor" answers in addition to positive examples, SK-QuAD significantly improves zero-shot accuracy of multilingual models on Slovak QA tasks [6].

Complementary language-specific optimisations are equally important. The Slovak Morphological Tokeniser (**SKMT**), a byte-pair-encoding (BPE) variant that preserves stem integrity, can markedly boost downstream performance [7].

Like many non-dominant language communities, Slovakia faces systematic barriers to adopting state-of-the-art AI. Model training data rarely include sufficient Slovak text, and the few closed models that do support Slovak are unavailable for security-sensitive or confidential deployments. As a result, public institutions and companies struggle to exploit LLMs fully.

The European Union has begun to address this gap by creating funding frameworks for national-language models. Yet, Slovak development is complicated by the prevalence of Czech in shared corpora; close linguistic proximity often leads LLMs to conflate the two languages unless explicitly disambiguated. This phenomenon is common across closely related languages and underscores the need for models that can make finer distinctions.

Even where open-source models nominally support Slovak, practical use may be impossible: LLaMA 3.1-70B, for example, requires more than 200 GB of GPU memory merely to run inference.

Finally, conventional automatic metrics such as **BLEU** and **ROUGE** designed around English—poorly capture quality in morphologically rich languages like Slovak. Current research therefore calls for language-sensitive evaluation methods [8]. In this work I adopt OpenAI GPT-4o-mini, a proprietary model with high Slovak proficiency, as an automatic judge to benchmark various open-source LLMs.

The following section first outlines the experimental design used to evaluate Slovak-language competence across selected LLMs, including the newly created 100-question benchmark and the six-component evaluation scale (grammar, semantics, style & context, slang and regional expressions, translation quality, and complex constructions) together with an automated scoring pipeline built on GPT-4o-mini.

The **Evaluating LLM Performance in Slovak** block summarizes practical insights from the test runs, detailing the custom Python helper that ensured consistent, reproducible scoring. Finally, **Results and Model Comparison** synthesizes the quantitative findings, highlights the qualitative leap delivered by Google's Gemma3:27B, and discusses the trade-offs between cloud-based and on-premise deployment, pointing to Gemma3:27B as the current best balance of Slovak accuracy and hardware efficiency.

## 1 Evaluating LLM Performance in Slovak

As shown in (Fig.1), generating a single answer with the LLaMA-3.1-70B model can take nearly five minutes. In a full run of 100 diverse Slovak questions, the evaluation was never completed without errors; one attempt lasted more than six hours. This makes the model impractical for large-scale evaluations.

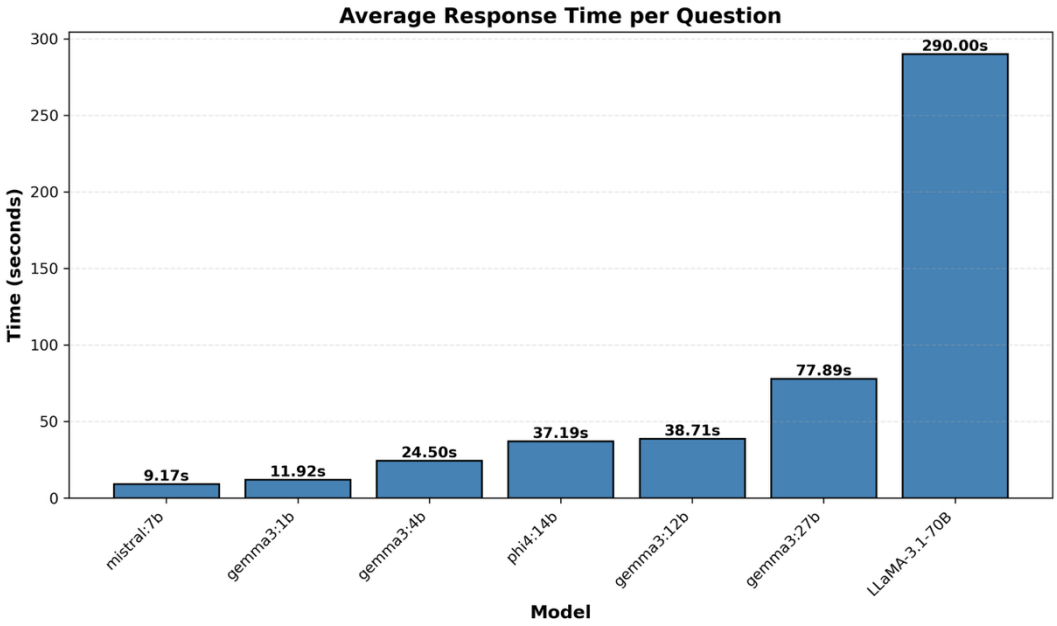


Fig.1. Average response time per question for each evaluated model.

For an in-depth analysis, I posed 100 diverse Slovak-language questions to several language models, assessing each response against six clearly defined criteria:

1. **Grammar**—evaluating the grammatical correctness of the generated sentences.
2. **Semantics**—assessing the accuracy of the conveyed meaning.
3. **Style and context**—determining the model's ability to maintain contextual relevance and appropriately answer follow-up queries.
4. **Slang and regional expressions**—measuring the understanding and accurate usage of Slovak slang and regional dialects.
5. **Translation**—examining the translation quality between Slovak and other languages.
6. **Complex constructions**—testing the models' handling of advanced grammatical structures and nuanced linguistic expressions.

To automate and streamline the evaluation process, an orchestration workflow was implemented (Fig.2). This system systematically submits test questions, collects model responses, and uses the GPT-4o-mini judging service to score them across six criteria, ensuring accuracy and consistency in the evaluation.

## 2 Results and Model Comparison

The evaluation results are presented in (Tab.1). They reveal a significant qualitative leap in the latest language models, particularly in Gemma 3:27B. Google's Gemma 3 series, available in 1B, 4B, 12B, and 27B parameter versions, features multimodal capabilities and context windows reaching up to 128k tokens. These models offer extensive language support exceeding 140 languages and are specifically engineered for efficient deployment on resource-limited systems, making genuine edge computing scenarios feasible.

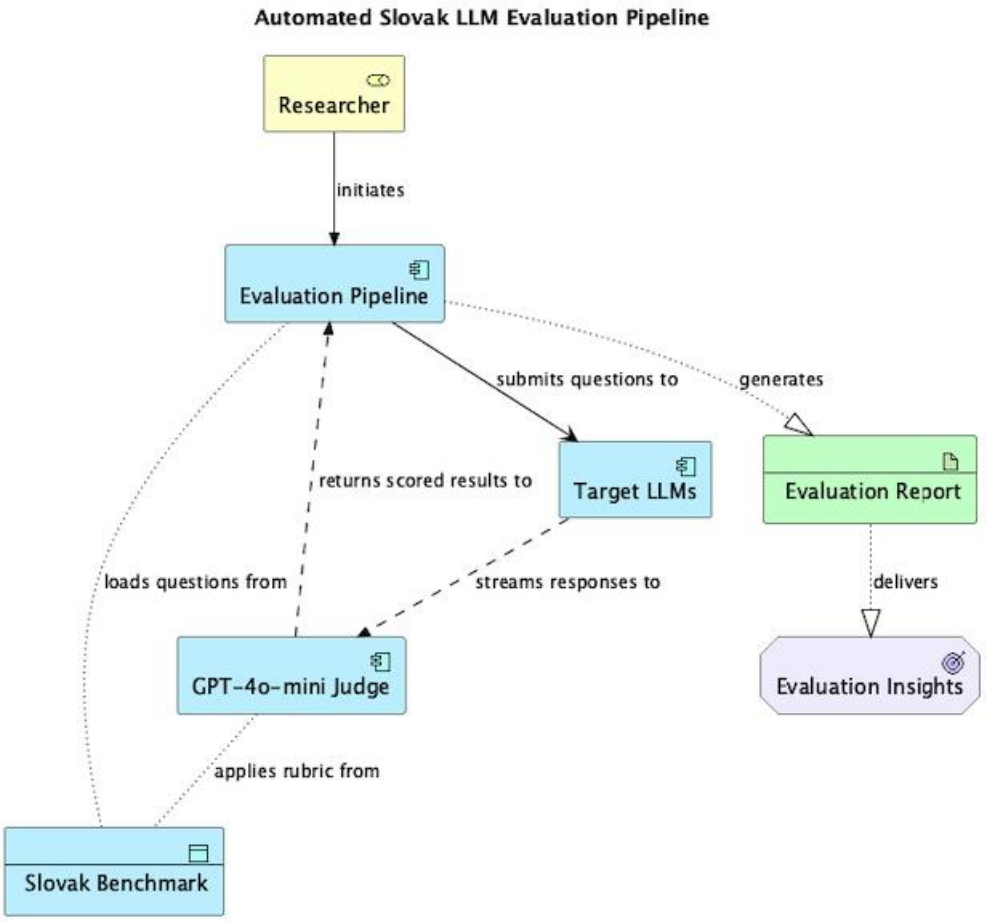


Fig.2. Automated Slovak LLM evaluation workflow.

Table 1. Evaluation results and response times of selected large-language models.

| Large Language Model | Response time (s) | Grammatical accuracy | Semantic precision | Context retention | Understanding slang and regional expressions | Translation quality | Handling complex language constructions |
|---|---|---|---|---|---|---|---|
| gemma3:27b | 77.89 | 9.1 | 9.2 | 9.9 | 8.2 | 8.1 | 8.8 |
| gemma3:12b | 38.71 | 8.9 | 9.0 | 9.8 | 8.0 | 7.5 | 8.5 |
| gemma3:4b | 24.5 | 8.0 | 8.2 | 9.0 | 7.0 | 6.2 | 7.6 |
| deepseek-r1:32b | 81.75 | 7.4 | 8.2 | 8.4 | 6.4 | 5.5 | 7.5 |
| phi4:14b | 37.19 | 7.0 | 7.6 | 8.0 | 5.8 | 5.8 | 6.8 |
| deepseek-r1:14b | 37.35 | 6.0 | 6.8 | 6.8 | 4.8 | 3.8 | 6.0 |
| gemma3:1b | 11.92 | 4.7 | 4.9 | 5.8 | 3.6 | 3.4 | 4.6 |
| granite3.2:8b | 24.22 | 4.7 | 5.6 | 6.0 | 4.2 | 3.5 | 5.0 |
| mistral:7b | 9.17 | 4.4 | 5.1 | 5.6 | 4.0 | 3.1 | 4.5 |

Identifying an appropriate language model for Slovak remains challenging due to limited native-language support, significantly hindering regional AI progress. Additionally, operational expenses, especially the stark contrast between cloud-based and on-premise deployments, present substantial hurdles.

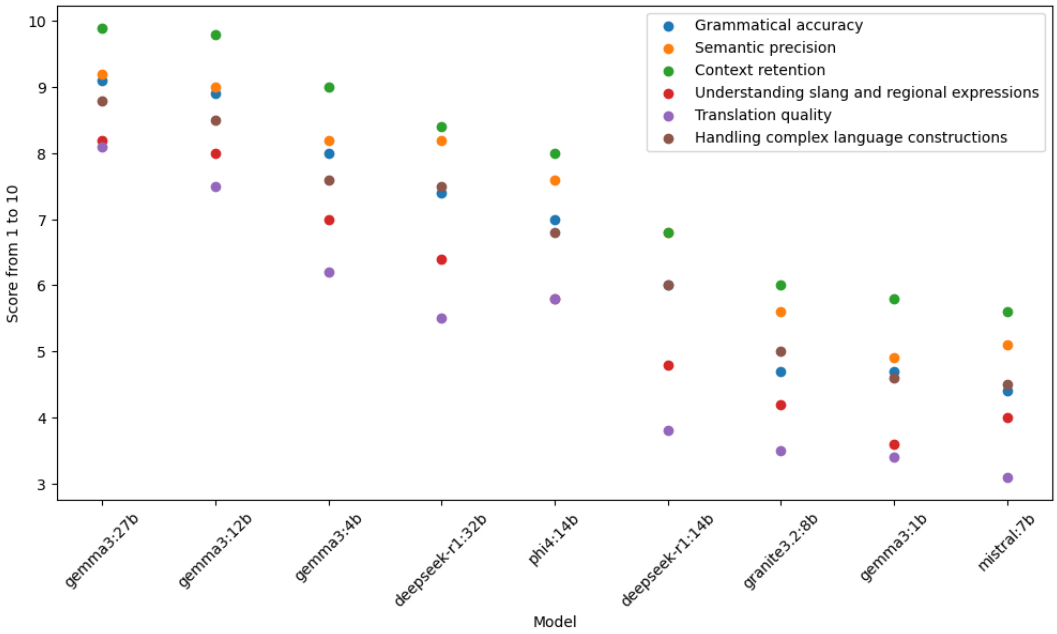


Fig.3. Performance comparison of large language models by efficiency and accuracy.

Several factors influence the choice between cloud and on-premise deployments. Cloud solutions offer scalability, reduced initial investment, and flexible cost structures. Conversely, on-premise deployments necessitate substantial initial expenditures on hardware but offer superior control over data privacy and security. Organizations must strategically evaluate these factors based on their specific operational requirements and resources.

Initially, models from OpenAI such as GPT-3 and subsequently GPT-4o established a robust benchmark for Slovak-language support. Google's Gemini later enhanced this landscape. Nevertheless, there remained a notable gap for a robust, open-source solution suitable for efficient on-premise utilization. Their relative efficiency and accuracy are visualized in (Fig.3).

Experiments with open-source alternatives, notably Meta's LLaMA 3.1-70B, yielded underwhelming results. Although Slovak was nominally supported, the quality of responses was inadequate, coupled with impractical hardware demands. Subsequent generations marked substantial improvements, notably the Mistral-Large, which significantly enhanced Slovak handling and achieved commendable performance.

The real breakthrough emerged with Gemma 3:27B, whose Slovak language proficiency closely rivals that of GPT-4o while remaining practically viable for on-premise deployment scenarios. Effective native-language support extends beyond mere convenience, becoming a critical component of national security, given its profound impact on the capability to directly analyze and manage local data, thereby reinforcing cyber-resilience.

## 3 Related Work

Ahuja et al. [9] introduced MEGAVERSE, a multilingual benchmark covering 22 datasets and 83 languages, many of them low-resource [9]. While MEGAVERSE provides broad cross-lingual comparisons, it does not address whether mid-sized open models can reach near state-of-the-art performance in specific languages. The evaluation reported here shows that, for Slovak, the Gemma 3:27B model achieves results comparable to proprietary models, suggesting that efficient alternatives exist even in constrained settings.

Skadiņa et al. [10] constructed the first Latvian benchmark by translating tasks such as COPA and MMLU, and demonstrated that manual post-editing significantly improves model accuracy. The study reported here avoids translation artifacts by employing a manually crafted Slovak dataset, highlighting the importance of high-quality, native-language resources in morphologically rich contexts.

Ojo et al. [11] proposed AfroBench, a benchmark evaluating 64 African languages across 15 tasks, and reported substantial performance gaps between proprietary and open-source models [11]. A similar disparity was confirmed for Slovak through the experimental results presented in this work. However, the evaluation also demonstrates that optimized tokenization and targeted data curation can substantially reduce this performance gap.

Arnett & Bergen [12] investigated the underperformance of LLMs in morphologically complex languages and attributed the issue primarily to reduced effective training data rather than linguistic structure itself. This interpretation is supported by the results presented here, which indicate that increasing the availability of Slovak-language data allows models like Gemma3:27B to nearly match English-language performance.

Azime et al. [13] in the ProverbEval benchmark, highlighted that culturally grounded tasks such as proverb understanding tend to increase performance variability across languages. The study reported here responds to that challenge by incorporating “slang and regional expressions” as one of six evaluation categories, exposing limitations even in otherwise strong models.

## Conclusion

The evaluation of large language models (LLMs) for Slovak has highlighted significant advancements, particularly with Google’s Gemma 3:27B model. This model demonstrates a remarkable balance between performance and efficiency, making it a viable option for on-premise applications. Its ability to handle complex linguistic structures and provide accurate translations underscores its potential in various domains.

However, relying solely on foreign-developed models poses challenges. While models like Gemma 3:27B offer impressive capabilities, they may not fully capture the nuances of the Slovak language or cater to specific national needs. Moreover, dependence on external models raises concerns about data sovereignty, security, and long-term sustainability.

Although (Tab.1) shows that Gemma 3:27B scores above 9 in grammatical accuracy and semantic precision—and maintains excellent context retention—a closer look at its output still reveals several micro-level deficiencies. The most common are orthographic mistakes (missing diacritics or wrong dash characters), morphological slips (incorrect adjective gradation or case forms), lexical/terminological inaccuracies (e.g., using *“časový základ - time base”* instead of the standard grammatical term *“časovanie – timing”*), inconsistent capitalisation of names, minor typographic issues, and occasional stylistic redundancy caused by overloaded bullet lists.

These errors stem from limited exposure to expertly proof-read Slovak data in the training corpus and from the language’s highly inflected morphology, which introduces many exceptions the model only partly “guesses” correctly.

For most everyday use cases the text remains fluent and perfectly comprehensible—especially when contrasted with older open-source LLMs such as mistral-7b, whose grammatical and semantic scores hover around 4–5 in (Tab.1). However, for official publications or legal documents a human proof-reader is still indispensable to eliminate the systematic fine-grained errors current models cannot yet fully capture.

To address these issues, it is imperative for Slovakia to invest in developing its own LLMs. Such an initiative would not only enhance the country's technological autonomy but also ensure that the unique characteristics of the Slovak language are adequately represented and preserved. By building a dedicated Slovak LLM, tailored to the nation's linguistic and cultural context, Slovakia can foster innovation, support local industries, and strengthen its position in the global AI landscape.

In conclusion, while leveraging existing models like Gemma 3:27B is beneficial in the short term, the strategic development of a national Slovak LLM is essential for long-term growth, resilience, and cultural preservation. This endeavor will empower Slovakia to harness the full potential of AI technologies while safeguarding its linguistic heritage.

## References


[1] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel C., (2021). Google Research, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[2] Viksna, R., Skadin, I., Deksne, D., Rozis, R., (2023). Large Language Models for Multilingual Slavic Named Entity Linking, Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023, Association for Computational Linguistics.

[3] Pikuliak, M., Grivalský, Š., Konôpka, M., Blšták, M., Tamajka, M., Bachratý V., Šimko, M., (2022). SlovakBERT: Slovak Masked Language Model. In Findings of the Association for Computational Linguistics: EMNLP 2022 (proceedings). Abu Dhabi, United Arab Emirates.

[4] Shliazhko, O., Fenogenova, A., Tikhonova, M., Kozlova, A., Mikhailov, V., Shavrina, T., (2025). mGPT: Few-Shot Learners Go Multilingual. Institute of Linguistics RAS, Russia.

[5] Bednár P., Dobeš, M., Garabík R., (2023). Mistral-sk-7b. Hugging Face. Štúr Institute of Linguistics, Slovak Academy of Sciences, supported by DiusAI a. s..

[6] Hládek, D., Staš, J., Juhár J., Koctúr O., (2023). Slovak Dataset for Multilingual Question Answering.

[7] Držík, D., Forgáč, F., (2024). Slovak morphological tokenizer using the Byte-Pair Encoding algorithm.

[8] Dobeš, M., (2025). Evaluation of quality of Slovak language use in LLMS, Acta Electrotechnica et Informatica, Vol. 25, No. 1, 2025.

[9] Ahuja, S., Aggarwal, D., Gumma, V., Watts, I., Sathe, A., Ochieng, M., Sitaram, S. (2024). MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. Proceedings of NAACL-HLT 2024. Mexico City, Mexico.

[10] Skadiņa, I., Bakanovs, B., Darģis, R. (2025). First Steps in Benchmarking Latvian in Large Language Models. Proceedings of RESOURCEFUL-2025. Tallinn, Estonia.

[11] Ojo, J., Ogundepo, O., Oladipo, A., et al. (2023). AfroBench: How Good Are Large Language Models on African Languages.

[12] Arnett, C., Bergen, B. (2025). Why Do Language Models Perform Worse for Morphologically Complex Languages. Proceedings of COLING 2025.

[13] Azime, I. A., Tonja, A. L., Belay, T. D., et al. (2025). ProverbEval: Exploring LLM Evaluation Challenges for Low-resource Language Understanding. In Findings of the Association for Computational Linguistics: NAACL 2025 (proceedings). Albuquerque, New Mexico.

## Authors

**Bc. Patrik Skovajsa**
Pan-European University,
Faculty of Informatics, Bratislava, Slovakia
patrik.skovajsa@gmail.com
Secure AI solutions, risk management, LLMs, IT Expert, build the future with AI, one innovation at a time—ready to lead or collaborate on groundbreaking projects.