

Slovak language now has open AI models that run locally and match the quality of commercial APIs

9.7.2026 - | Kempelenov inštitút inteligentných technológií

Slovak researchers have developed SkMTEB, the first comprehensive benchmark for measuring how well artificial intelligence models understand and process Slovak text. The project also introduces two new openly available models for Slovak that deliver performance comparable to commercial embedding APIs while being several times smaller and more efficient. They can run locally on relatively affordable hardware, without the need to send data to third parties.

The work will be presented at ACL 2026, one of the world's leading conferences on natural language processing. It was developed by a team from the *Kempelen Institute of Intelligent Technologies (KInIT)*, the *Technical University of Košice (TUKE)*, and *Comenius University Bratislava (UK)*, together with other partners from the Slovak AI ecosystem.

Until now, AI systems have primarily been designed and evaluated for English and other major world languages. For Slovak, however, there has long been no reliable way to compare which technologies truly understand the language. SkMTEB changes that by providing a reliable way to evaluate which models perform best on Slovak.

At the core of the research are embeddings. This technology converts text into numerical representations, allowing a computer to recognize the meaning of text even when different words are used. Embeddings are the foundation of search engines, document-based chatbots, and tools for automatic text classification. They enable a system to find a relevant document even when a question is phrased differently, or recommend the right product even if the user does not know its exact name. If these tools do not understand Slovak well enough, the quality of services for both users and organizations is affected.

SkMTEB evaluates models across 31 datasets and seven types of tasks. These include document retrieval, reranking search results by relevance, measuring semantic similarity between texts, text classification, clustering, and retrieving matching sentences across different languages. Using the benchmark, the research team evaluated 31 existing models and introduced two new models tailored for Slovak: e5-sk-small with 45 million parameters and e5-sk-large with 365 million parameters. The smaller model is lightweight enough to run directly in a web browser. *"Its main advantage is that it can be used locally. Anyone can download it free of charge, run it on their own computer, and build Slovak-language applications without sending their data to anyone else,"* explains **Daniel Hládek, Associate Professor at the Department of Computer Networks at the Technical University of Košice**. While significantly larger instruction-tuned models achieved the highest overall benchmark scores, the Slovak models deliver results comparable to commercial APIs despite being only a fraction of their size.

"Our results show that high-quality Slovak language processing does not necessarily require the

*largest commercial models. Our smaller Slovak model achieved benchmark results comparable to commercial APIs on SkMTEB,” says **Andrej Ridzik, Research Engineer at the Kempelen Institute of Intelligent Technologies.***

The ability to run these models locally can be particularly valuable wherever sensitive data is involved, such as in public administration, healthcare, and the financial sector. The models are released under an open licence. Alongside them, the researchers have also published the datasets and source code, allowing others to build on their work. The use of individual datasets is governed by their respective licences.

*“We believe this will open the door to a new wave of applications wherever it is important for technology to truly understand written Slovak,” concludes **Marek Šuppa, an AI expert at the Department of Applied Informatics, Comenius University Bratislava.***

More information:

Research paper: <https://arxiv.org/abs/2606.13647>

Open-source code and models: <https://github.com/slovak-nlp/skmteb>

Benchmark overview: <https://slaih.sk/sk/benchmarks/skmteb>

<https://kinit.sk/slovak-language-now-has-open-ai-models-that-run-locally-and-match-the-quality-of-commercial-apis>