

From Words to Action: A National Initiative to Overcome Data Scarcity for the Slovene LLM

Špela Arhar Holdt (1), Špela Antloga (1, 2),
Tina Munda (1), Eva Pori (1),
Simon Krek (1, 3)

¹ University of Ljubljana, Slovenia

² University of Maribor, Slovenia

³ Institut "Jožef Stefan", Slovenia

1) The project **PoVeJMo—Adaptive Natural Language Processing with Large Language Models** aims to develop large language models for the **Slovene language**.

2) LLMs rely on vast and diverse datasets, which poses challenges for languages with limited resources, such as Slovene, due to its **smaller speaker base and restricted data availability**.

3) To meet this challenge, we have commenced a **national initiative for a large-scale collection of Slovene texts**.

4) To overcome legal constraints, we have taken a more direct approach by **actively seeking permission from copyright holders** to use their texts.

5) The text collection campaign operates along **two main strategies**:

- We engage large-scale text providers such as national libraries, publishers, media organizations and government ministries.
- We engage individuals, inviting them to donate their own texts.

6) The data is received either via network connections or on portable storage devices. We have implemented a comprehensive protocol for **secure data management** at all stages of its lifecycle.

Affiliations:



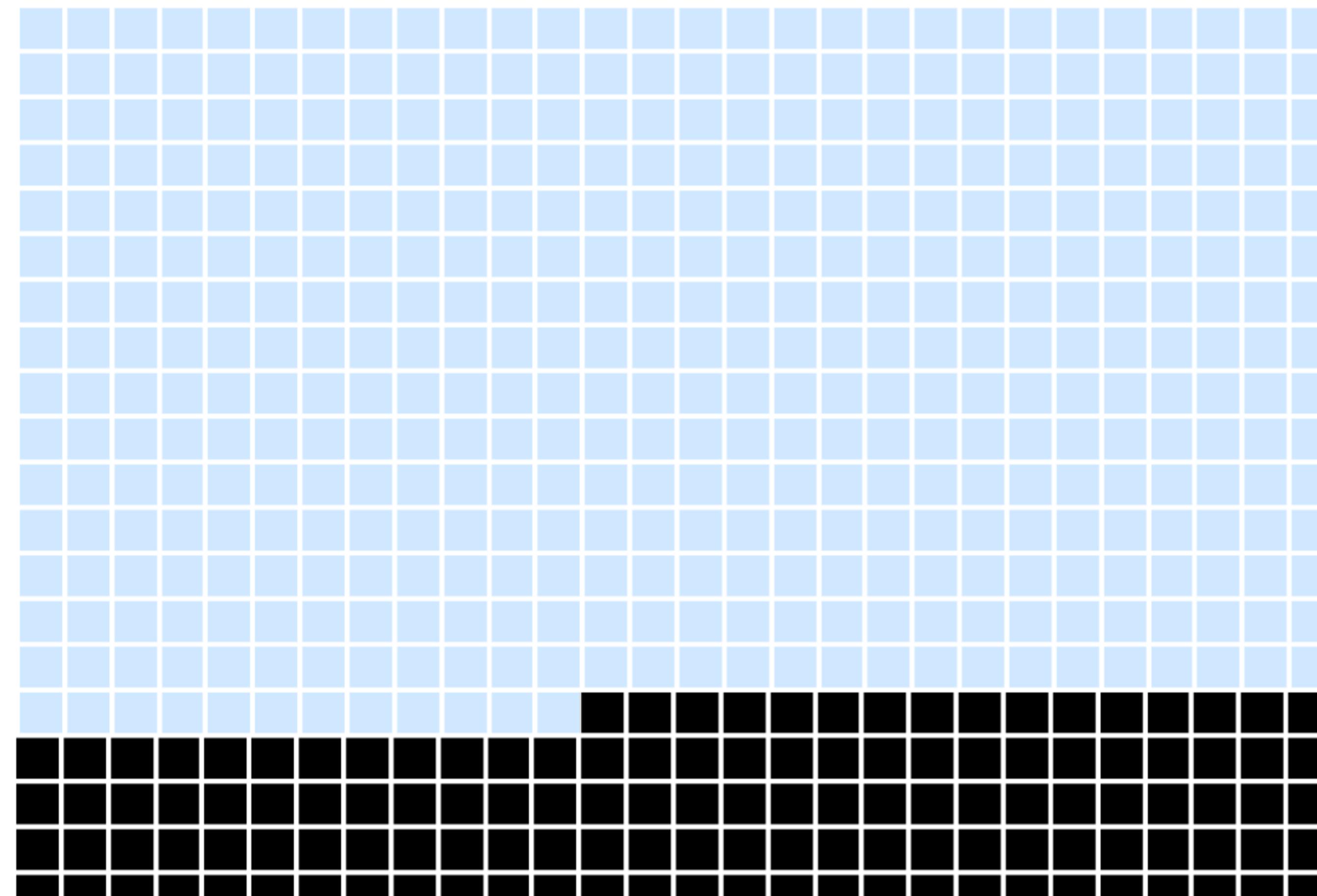
Institut
"Jožef Stefan"
Ljubljana, Slovenija

Univerza v Mariboru

Acknowledgements:

The research program Language Resources and Technologies for Slovene (P6-0411) and the projects Adaptive Natural Language Processing with Large Language Models and Large Language Models for Digital Humanities (GC-0002) are funded by the Slovenian Research and Innovation Agency.

40 billion words



Web portal for text submission available at: zbiranje.povejmo.si

CURRENT STAGE

9,2 BILLION WORDS

The screenshot shows a light blue-themed web page titled 'PoVeJMo' with the subtitle 'SLOVENSKI VELIKI JEZIKOVNI MODEL'. A 'Domov' link is in the top right. The main content area is titled 'Oddaja besedil' (Text Submission). It contains three numbered steps: 1. 'Vnesete osebne podatke' (Enter personal data), 2. 'Prejmete povezavo do obrazca za oddajo besedil' (Receive a link to the text submission form), and 3. 'Naložite datoteke' (Upload files). A 'ZAČNITE' button is at the bottom.

Figure 1: The section of the web portal where the interested participants can provide their texts (<https://zbiranje.povejmo.si/>).

The screenshot shows a dark-themed web page titled 'PoVeJMo' with the subtitle 'PREIZKUSITE TRENUTNO RAZLIČICO JEZIKOVNEGA MODELJA!'. It features a large text block about the GAMS-1B-Chat model, mentioning its size and how it was trained on Slovene news. Below this is a note about the existing model's limitations. At the bottom is a 'Postavi vprašanje' (Ask a question) button.

Figure 2: The section of the web portal where the interested participants can test the existing model (<https://povejmo.si/klepeta/>).

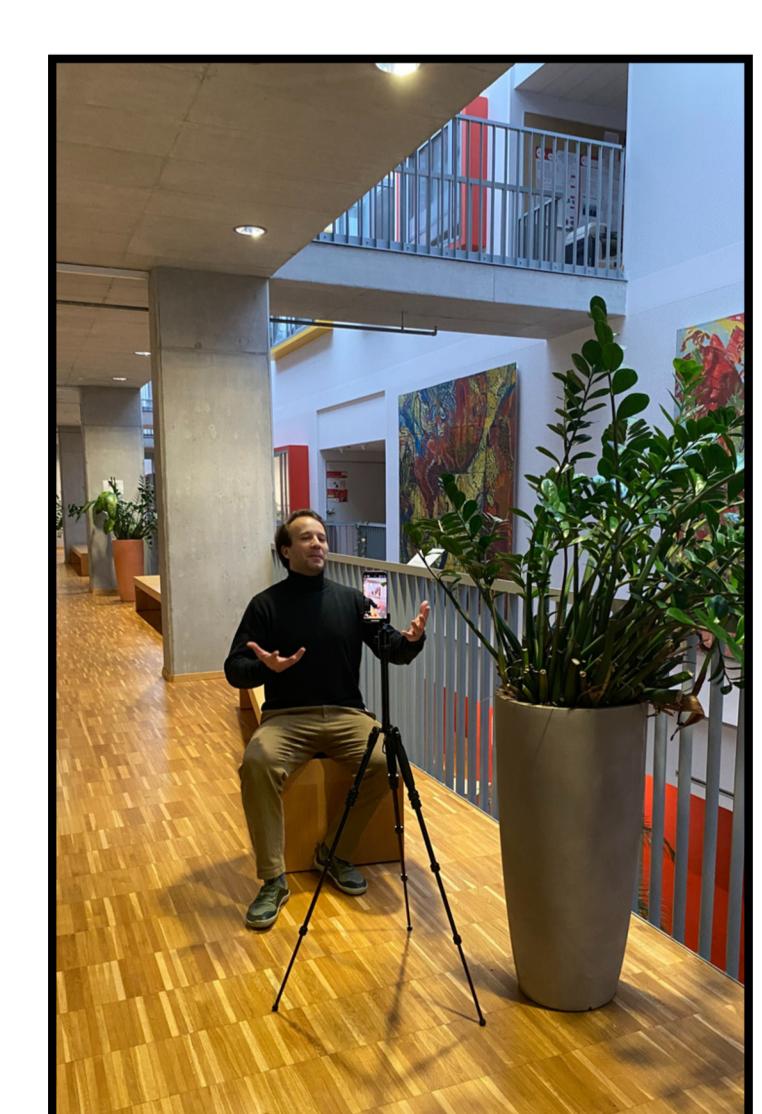
Engaging institutions:

- Slovene language must remain comparable and competitive with other similar languages in the digital age.
- A collective effort.
- Independence from foreign corporations.



Engaging individuals:

- Radio, TV and social media presence.
- Web portal **povejmo.si** for text submission.
- Interactive section where users can try the current version of the language model. Try it at <https://povejmo.si/klepeta/>



This poster was presented at the RESOURCEFUL-2025 Workshop at the NoDaLiDa/Baltic-HLT 2025 Conference (Tallinn, Estonia, 2. 3. 2025). Please refer to the corresponding short paper for further information, references, and contact details.