

# ChatPULI: Enhancement to the first Hungarian conversational model

Zijian Győző Yang, Ágnes Bánfi, Réka Dodé,  
Gergő Ferenczi, Flóra Földesi, Péter Hatvani, Enikő Héja,  
Mariann Lengyel, Gábor Madarász, Mátyás Osváth,  
Bence Sárossy, Kristóf Varga, Tamás Váradi,  
Gábor Prószéky, Noémi Ligeti-Nagy

ELTE Research Centre for Linguistics

[[yang.zijian.gyozo](mailto:yang.zijian.gyozo), [banfi.agnes](mailto:banfi.agnes), [dode.reka](mailto:dode.reka), [ferenczi.gergo](mailto:ferenczi.gergo), [foldesi.flora](mailto:foldesi.flora), [hatvani.peter](mailto:hatvani.peter),  
[heja.eniko](mailto:heja.eniko), [lengyel.mariann](mailto:lengyel.mariann), [madarasz.gabor](mailto:madarasz.gabor), [osvath.matyas](mailto:osvath.matyas), [sarossy.bence](mailto:sarossy.bence), [varga.kristof](mailto:varga.kristof),  
[varadi.tamas](mailto:varadi.tamas), [proszeky.gabor](mailto:proszeky.gabor), [ligeti-nagy.noemi@nytud.elte.hu](mailto:ligeti-nagy.noemi@nytud.elte.hu)]

**Abstract.** This paper presents the development and evaluation of PULI-LlumiX-Llama-3.1 Chat and PULI Trio Q Chat, the first Hungarian-focused conversational large language models based on the Llama 3.1 and Qwen 2.5 architectures. Extending previous work on Hungarian instruction-following models, we applied continual pre-training on multilingual and Hungarian corpora, followed by supervised fine-tuning on an expanded instruction dataset including Hungarian, English, and Chinese prompts. Our models demonstrate significant performance improvements on Hungarian language understanding benchmarks, as well as on few-shot and zero-shot tasks, compared to earlier PULI models. Additionally, they show enhanced capabilities in machine translation and multi-turn dialogue handling. These results highlight the effectiveness of continual pre-training and fine-tuning strategies for adapting large language models to low-resource languages like Hungarian, and provide a foundation for future research in conversational AI for under-represented languages.

**Keywords:** PULI models, Llama, Qwen, large language model, conversational language model

**AMS Subject Classification:** 68T07, 68T50, 68U15

# 1. Background

In recent years, large language models (LLMs) have gained significant attention, as companies compete to develop models capable of solving a wide range of natural language processing tasks through extensive data training. The release of ChatGPT<sup>1</sup> by OpenAI<sup>2</sup> showcased unprecedented capabilities, achieved through a multi-stage fine-tuning process. Nevertheless, while ChatGPT and similar large language models have been extensively trained for major languages, no conversational model has been developed specifically for Hungarian. Companies such as OpenAI typically include minimal Hungarian content in their training corpora, as is the case with most small, low-resource language with limited amount of available datasets.

For Hungarian, the first instruction-following model was introduced by Yang et al. [34], based on the Llama 2 architecture [29]. While this model demonstrated significantly improved performance compared to previous Hungarian LLMs on local benchmarks, it still lacks key functionality, most notably, the ability to engage in natural, conversational interaction.

Conversational fine-tuning datasets require substantial human resources due to their multi-turn interaction structure, and, as mentioned earlier, such datasets are currently lacking in the Hungarian language. In our recent research, we leveraged the advantages of transfer learning and, following the approach of Yang et al. [34], we applied continual pre-training to the Llama 3.1 8B Instruct [10] and Qwen2.5 7B Instruct [31] models to adapt their conversational capabilities to the Hungarian language. Our newly continual pre-trained base models are named **PULI-LlumiX-Llama-3.1** and **PULI Trio Q**. The name “PULI” refers to the eponymous Hungarian dog breed. This dog is not only a cultural symbol in Hungary, but is also known for its agility, energy, and intelligence despite its small size. By choosing this name, we emphasize similar qualities in our small but capable model. Our base models are available at our Hugging Face repository<sup>3</sup>.

# 2. Related work

The global prominence for decoder-only transformer-based large language chat models was brought forth by OpenAI’s GPT model family [21] which was primarily trained on English data. In the past years, an array of new models have appeared at a rapid pace. Besides OpenAI’s GPT series, the most notable closed-source families include Claude [2] by Anthropic and Gemini [7] by Google. More recently, open-source Deepseek models [8] have emerged as one of the most notable developments within the LLM landscape. They introduced techniques such as supervised fine-tuning (SFT) and Group Relative Policy Optimization (GRPO),

---

<sup>1</sup><https://chatgpt.com>

<sup>2</sup><https://openai.com>

<sup>3</sup><https://huggingface.co/NYTK>

which helped improving training efficiency and lowering computational costs, while maintaining competitive performance.

Other open-source model families include Alibaba’s Qwen series and the subsequent Qwen2 models [31], which comprise a range of base models as well as instruction-tuned LLMs, spanning from 0.5B to 72B parameters. The series also includes a Mixture-of-Experts (MoE) model, further broadening the available architectures and training strategies within open-source development.

In some instances makers of closed-source model families also release open-source series such as is the case with Google’s Gemma model family. Their latest model, Gemma 3 [28] is a multimodal and multilingual model, available in various parameter sizes ranging from 1B to 27B, with a 128k-token context window.

Meta’s Llama series [29, 30] has also played a central role in improving access to high-performance models. Their multilingual pre-training, range of model sizes, and general-purpose capabilities have made them widely adopted, especially in contexts of smaller language model development, where training data and infrastructural capacities are limited.

Llama-2-7b was adapted to Hungarian by SambaLingo for its pre-trained bilingual base model, Sambalingo-Hungarian-Base [6]. The model, that also has its chat version, was trained on 59 billion tokens sourced from the Hungarian split of the Cultura-X corpus [19]. While it represents a significant step for Hungarian LLM-development, the monolithic nature of the training data places limitations on its ability to recognize the nuances and genre diversity of the language.

Finally, another model family with notable Hungarian representation is the EuroLLM series [18]. As of the time of writing, their flagship model, EuroLLM-9B, was trained on 4 trillion tokens drawn from all EU official languages. However, during pretraining – prior to instruction tuning and alignment – the dataset remained predominantly English, with the language comprising 32.5–50% of the overall token mix. This raises concerns about the effective representation of mid- and low-resource languages in the model, including Hungarian.

Due to limited computational resources, we opted for an approach involving continual pre-training. This method allows us to incrementally update the model with new data without retraining from scratch, making it a more resource-efficient alternative.

### 3. Datasets and PULI models

In our continual pre-training experiment, we followed the methodology of [34]. In Table 1, we present the main characteristics of the continual pre-training corpora. For the PULI long and English parts, we used the same corpora (containing 8,077,464,716 words) as employed in the work of Yang et al. [34], but we additionally incorporated the Hungarian Wikipedia to enhance the model’s knowledge of the Hungarian language. For the Chinese part, we used the Wu Dao 2.0 corpus [37], which consists of long documents (over 15,000 characters) drawn from its publicly available portion. During the pre-training experiments, we observed that

the model generated more grammaticality errors in its outputs. To address this issue, we concluded the training with a Hungarian-only dataset. In this final phase, our goal was not only to enhance the model’s proficiency in Hungarian, but also to provide it with up-to-date information. Accordingly, the dataset was composed of carefully selected Hungarian articles published over the past six years. The corpora used in this stage were: Hungarian Wikipedia (2019–2025), containing 174,945,601 words, and Hungarian news articles (2019–2025), containing 622,275,769 words. The characteristics of the final-stage corpus are presented in Table 1.

**Table 1.** Corpus characteristics for continual pre-training.

|                      | Documents | Words                         | Average document length<br>average / median (word) |
|----------------------|-----------|-------------------------------|----------------------------------------------------|
| PULI long (hu)       | 763 704   | 7 902 519 115                 | 10 823.38 / 7 149                                  |
| Long Context QA (en) | 88 957    | 1 009 562 704                 | 11 348.88 / 11 274                                 |
| BookSum (en)         | 9 600     | 42 339 698                    | 4 410.39 / 3 266                                   |
| Wu Dao 2.0 long (zh) | 174 118   | 2 855 217 266<br>(Characters) | 16 398.17 / 14 767<br>(Characters)                 |
| Final stage (hu)     | 1 411 979 | 797 221 370                   | 443.65 / 217                                       |

The continual pre-training phase employed different multilingual mixtures for each model. The PULI-LlumiX-Llama-3.1 (bilingual) model was trained on a combination of the PULI long and English components. In contrast, the PULI Trio Q (trilingual) model utilized a mixture of the PULI long, English, and Chinese components. The training process for both models concluded with a final stage using exclusively Hungarian data.

For fine-tuning, we used the same basic corpus as in the work of Yang et al. [34], but we significantly expanded the instruction dataset. To better align the model with conversational use, we also collected a large number of dialogue-style prompts and replaced the Stanford Alpaca instruction template with the Llama 3 and Qwen2 chat templates to ensure consistency with the base models’ instruction style. Our contributions to the fine-tuning dataset include:

- HuLU benchmark [15] prompts: Increased to 7,344 segments;
- MILQA benchmark [20] prompts: Increased to 2,191 segments;
- Graduation tasks: Increased to 5,589 segments;
- Title, keyword generation, and summarization tasks [3, 33]: Increased to 3,000 segments;
- Conversational and instructional prompts [14, 17, 26, 38]: Hungarian-specific content, totaling 10,884 prompts;
- Psychiatric content [9]: 315 interviews;

- Text simplification [24]: 2,274 prompts;
- Named Entity Recognition prompts: 2,300 prompts generated from NYTK-NerKor [25];
- University exam: 166 publicly available questions covering tourism, business, geography, and economics.
- User questions: Increased to 935 segments. These prompts are generated by questions that users ask from our demo models. In these cases, the answers are manually written by annotators.
- Poem prompts: 11,800 prompts collected from Hugging Face<sup>4</sup>.
- Miscellaneous: Increased to 3,358 segments. Prompts from different customer datasets that can be made public.

In total, our Hungarian fine-tuning dataset consists of 44,626 segments. In our supervised fine-tuning, inspired by Yang et al. [34], we mixed English and Chinese prompts to our Hungarian prompts:

- Bilingual prompt dataset: In the case of Llumix model, we added 85,775 English prompts: Stanford Alpaca [27], Dolly [5] and llama-instruct<sup>5</sup>.
- Trilingual prompt dataset: In the case of Trio model, beside the English prompts, we added 50,130 Chinese prompts: neo\_sft\_phase2<sup>6</sup> Chinese part.

To evaluate the performance of our models, we compared them with the previous PULI models:

- **PULI 3SX** is a GPT-NeoX-based [4] model with 6.85 billion parameters, pre-trained from scratch on a monolingual Hungarian corpus of 36.3 billion words.
- **PULI Trio**: Another GPT-NeoX-based model, with 7.67 billion parameters, trained from scratch as a Hungarian-English-Chinese trilingual model. The Hungarian subset of the training data contains 41.5 billion words.
- **PULI Llumix**: A continued pre-training adaptation of Llama 2 [30]. The original Llama 2 model was trained on approximately 600 million Hungarian tokens. This version was further trained on an additional 7.9 billion Hungarian words, emphasizing long-form documents exceeding 5,000 words. Additionally, the context window has been expanded to 32,768 tokens.

<sup>4</sup><https://huggingface.co/datasets/leinadsened/hungarian-poems-with-instructions>

<sup>5</sup><https://huggingface.co/datasets/togethercomputer/llama-instruct>

<sup>6</sup>[https://huggingface.co/datasets/m-a-p/neo\\_sft\\_phase2](https://huggingface.co/datasets/m-a-p/neo_sft_phase2)

## 4. Methods and experiments

In our current research, we adapted two multilingual LLMs to Hungarian. Our goal was to balance computational efficiency and performance; therefore, inspired by [34], we selected the Llama 3.1 8B Instruct [10] and Qwen2.5 7B Instruct for continued pre-training to adapt it to Hungarian. According to prior research, transfer learning plays a crucial role in language adaptation, and instruction-tuned models contain valuable structured knowledge compared to base models. However, we found no previous research that applies continued pre-training on an instruct model for language adaptation.

For the training, we used the LLaMA-Factory [40] framework with the following hyperparameters: global batch size = 128; learning rate = 1e-5; precision = bf16 (Llama) / fp32 (Qwen); context length = 16,384 (Llama) and 32,768 (Qwen); epochs = 1.

We trained two base models from the Llama and Qwen Instruct models, and two conversational or chat models from it:

- **PULI-LlumiX-Llama-3.1:** The Llama 3.1 8B Instruct model were continual pre-trained on bilingual (Hungarian-English) dataset, then completed on monolingual Hungarian dataset.
- **PULI-LlumiX-Llama-3.1 Chat:** The PULI-LlumiX-Llama-3.1 model were supervised fine-tuned on our bilingual (Hungarian-English) fine-tuning dataset, with 3 epochs.
- **PULI Trio Q:** The Qwen2.5 7B Instruct model were continual pre-trained on trilingual (Hungarian-English-Chinese) dataset, then completed on monolingual Hungarian dataset.
- **PULI Trio Q Chat:** The PULI Trio Q model were supervised fine-tuned on our trilingual (Hungarian-English-Chinese) fine-tuning dataset, with 3 epochs.

For the conversational supervised fine-tuning, we also used the LLaMA-Factory framework with the following hyperparameters: global batch size = 64; learning rate = 1e-5; precision = bf16 (Llama) / fp32 (Qwen); context length = 32,768; epochs = 3.

For the fine-tuning evaluation on the HuLU benchmarks, we used the Hugging Face implementation<sup>7</sup>.

All training was conducted on four NVIDIA A100 GPUs (80 GB each), utilizing the Accelerate library [11], Fully Sharded Data Parallel [39] and Liger Kernel [12] in all cases.

For the few-shot and zero-shot evaluation tasks, to ensure a fair comparison, we applied the same settings, evaluation metrics, prompts, and methodologies as described in the study by Yang et al. [35].

<sup>7</sup><https://github.com/huggingface/transformers/tree/main/examples/pytorch>

## 5. Results and evaluation

First, we examined the vocabularies and the Hungarian training data subsets of the openly available Hungarian LLMs. In Table 2, the Hungarian dataset/subset (in words), vocabulary sizes, and fertility values [1] are observed. The fertility values were calculated from the largest PULI corpus [36], which consists of 41 billion words. Table 2 highlights the impact of training corpus composition on vocabulary and fertility values. The earlier PULI models (3SX and Trio), trained exclusively or primarily on Hungarian data, used vocabularies optimized for Hungarian morphology and word formation, resulting in lower fertility values (1.58–1.63). This reflects a tighter, language-specific tokenization. In contrast, the newer LlumiX and Trio Q models – pre-trained on multilingual corpora with larger token vocabularies – exhibit higher fertility (around 3.0), indicating a shift toward tokenization schemes influenced by multilingual instruction models. This change reflects a balance between leveraging multilingual model architectures and retaining adequate representation for Hungarian.

**Table 2.** Hungarian datasets, vocabularies and fertilities of PULI LLMs.

|                       | Dataset | Vocab   | Fertility |
|-----------------------|---------|---------|-----------|
| PULI 3SX              | 36.3B   | 50,048  | 1.634     |
| PULI Trio             | 41.5B   | 150,016 | 1.585     |
| PULI LlumiX           | 7.9B    | 32,000  | 3.058     |
| PULI-LlumiX-Llama-3.1 | 9.7B    | 128,256 | 3.062     |
| PULI Trio Q           | 9.7B    | 152,064 | 3.126     |

**Table 3.** Few-shot and zero-shot results.

|                            | Few-shot     |              |              |
|----------------------------|--------------|--------------|--------------|
|                            | HuCOLA       | HuRTE        | HuSST        |
| PULI 3SX                   | 51.51        | 52.40        | 54.57        |
| PULI Trio                  | 53.47        | 54.30        | 54.96        |
| PULI LlumiX                | 55.84        | 61.49        | 68.53        |
| PULI-LlumiX-Llama-3.1      | <b>71.26</b> | <b>72.63</b> | 69.38        |
| PULI Trio Q                | 63.85        | 62.98        | <b>70.79</b> |
|                            | Zero-shot    |              |              |
| PULI 3SX Instruct          | 61.76        | 52.09        | 46.27        |
| PULI Trio Instruct         | 51.97        | 74.54        | 54.50        |
| PULI LlumiX Instruct       | 66.98        | 74.54        | 70.06        |
| PULI-LlumiX-Llama-3.1 Chat | 72.98        | <b>86.76</b> | <b>79.44</b> |
| PULI Trio Q Chat           | <b>76.68</b> | 86.10        | 78.04        |

In our second evaluation phase, we conducted few-shot and zero-shot experiments. To ensure a fair comparison with the previous work of Yang et al. [35], we evaluated our models on HuCOLA, HuRTE, and HuSST using the same methodologies and metrics (balanced accuracy). Performance improved consistently across both few-shot and zero-shot tasks from PULI 3SX to the latest models. The PULI-LlumiX-Llama-3.1 Chat and PULI Trio Q models outperformed all previous iterations, with PULI-LlumiX-Llama-3.1 Chat achieving the highest scores on the HuCOLA, HuRTE, and HuSST benchmarks compared to PULI Trio Q. This underscores the effectiveness of continual pre-training and conversational fine-tuning on Hungarian datasets in enhancing model generalization. Interestingly, in the HuCOLA task, the Qwen-based PULI Trio Q model achieved the best result, demonstrating strong adaptability to specific tasks, even when those tasks are linguistically specialized.

Table 4 presents the few-shot and zero-shot experiments on English–Hungarian (en–hu) and Hungarian–English (hu–en) machine translation tasks. The evaluation metrics used were SacreBLEU [23] and ChrF [22], reported as BLEU / ChrF scores. The results show that the latest models exhibit significantly improved translation capabilities compared to the earlier PULI versions. In both few-shot and zero-shot translation tasks, the PULI-LlumiX-Llama-3.1 Chat and PULI Trio Q Chat models displayed marked improvements over their predecessors. The Chat variants significantly outperformed the earlier LlumiX and Trio models, particularly in the zero-shot setting, with PULI-LlumiX-Llama-3.1 Chat achieving the highest BLEU and ChrF scores. This suggests that conversational fine-tuning not only enhanced dialogue handling but also contributed to stronger cross-lingual transfer and translation capabilities.

**Table 4.** Comparison of PULI multilingual models on machine translation task.

|                            | en-hu                | hu-en                |
|----------------------------|----------------------|----------------------|
|                            | Few-shot             |                      |
| PULI Trio                  | 14.42 / 45.12        | 19.69 / 48.27        |
| PULI LlumiX                | 13.73 / 42.30        | 15.43 / 35.84        |
| PULI-LlumiX-Llama-3.1      | 19.31 / 50.58        | 29.27 / 56.77        |
| PULI Trio Q                | <b>21.24 / 52.68</b> | <b>31.42 / 59.41</b> |
|                            | Zero-shot            |                      |
| PULI Trio Instruct         | 24.86 / 52.51        | 26.88 / 53.34        |
| PULI LlumiX Instruct       | 18.32 / 50.63        | 27.85 / 56.56        |
| PULI-LlumiX-Llama-3.1 Chat | <b>37.79 / 63.02</b> | <b>43.62 / 66.79</b> |
| PULI Trio Q Chat           | 30.08 / 58.39        | 38.45 / 63.58        |

In the third phase of our evaluation, we performed fine-tuning experiments on our LLMs. Table 5 presents the fine-tuning performance of our new models. For evaluation, we further fine-tuned the models on four HuLU benchmarks [15] using

the LoRA [13] and full parameter fine-tuning method. The evaluation metrics were the original HuLU metrics, Matthews Correlation Coefficient (MCC) and Accuracy (ACC), respectively. We experimented with various hyperparameters; in the results tables, we report only the best results achieved. The results from HuLU benchmarks reinforce the superior performance of the conversationally fine-tuned models. PULI-LlumiX-Llama-3.1 Chat consistently achieved the best scores across all tasks, notably surpassing previous models in HuCOLA, HuCoPA, HuRTE, and HuSST. PULI Trio Q Chat, while trailing slightly behind LlumiX, still outperformed all older PULI models. This demonstrates the clear advantage of fine-tuning with a carefully designed conversational dataset, especially when paired with continual pre-training.

**Table 5.** Performance of the PULI-LlumiX-Llama-3.1 and PULI Trio Q models, on the HuLU benchmarks.

|                            | HuCOLA<br>(MCC) | HuCoPA<br>(MCC) | HuRTE<br>(MCC) | HuSST<br>(ACC) |
|----------------------------|-----------------|-----------------|----------------|----------------|
| PULI 3SX Instruct          | 59.5            | 5.3             | 55.5           | 79.7           |
| PULI Trio Instruct         | 63.1            | 44.5            | 59.2           | 79.1           |
| PULI LlumiX Instruct       | 70.3            | 73.4            | 68.2           | 81.9           |
| PULI-LlumiX-Llama-3.1 Base | 69.2            | 74.1            | 71.9           | 82.2           |
| PULI-LlumiX-Llama-3.1 Chat | <b>72.54</b>    | <b>74.1</b>     | <b>76.0</b>    | <b>83.1</b>    |
| PULI Trio Q Chat           | 64.7            | 67.1            | 68.8           | 82.1           |

We also conducted an experiment in which we fine-tuned the PULI-LlumiX-Llama-3.1 Base and PULI-LlumiX-Llama-3.1 Chat models. As expected, the Chat model achieved higher results in this comparison, likely because it had already encountered some examples from the HuLU benchmarks during its conversational training.

A sample conversation is shown in Table 6, demonstrating the model’s (PULI-LlumiX-Llama-3.1 Chat) ability to follow instructions and maintain context across multiple turns. In this example, we asked the model to list three characteristics for various entities: dog, snail, table, carrot, love, and the transcendent. At the end of the conversation, we asked it to list four characteristics instead. As shown in Table 6, the model consistently provided three relevant features for each entity, and correctly responded with four when requested.

## 6. Conclusion

In this work, we introduce **PULI-LlumiX-Llama-3.1 Chat** and **PULI Trio Q Chat**, our first conversational Hungarian large language models, based on the Llama 3.1 and Qwen2 architectures. Building upon the foundation laid by Yang et al. [34], we expanded both the pre-training and fine-tuning datasets. Our new model achieved notable improvements across multiple HuLU benchmarks and

demonstrated enhanced conversational capabilities for Hungarian compared to previous Llumix versions.

**Table 6.** Example snippet of a conversation.

|            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| User:      | mondj három tulajdonságot a kutyákról<br>(‘tell me three characteristics about dogs.’)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Assistant: | 1. A kutyák hűségesek és odaadóak gazdáik iránt.<br>2. A kutyák képesek felismerni és reagálni az emberi érzelmekre.<br>3. A kutyák nagyszerű képességgel rendelkeznek a szagokat és zajokat érzékelni.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| User:      | csigák? (‘snails?’)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Assistant: | 1. A csigák ügyesek a mászásban és kiválóan képesek felmászni a felületeken.<br>2. A csigák képesek kemény héjú otthonukat önmagukban visszahúzni és védelmet nyújtani számukra.<br>3. A csigák nagyon válogatósak és csak bizonyos növényeket táplálkoznak, melyek létfontosságú tápanyagokat biztosítanak számukra.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| ...        | ...                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| User:      | és a transzcendens? (‘and the transcendent?’)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| Assistant: | 1. A transzcendens olyan fogalom, amely a mindennapi élettől való fizikai vagy szellemi távolságra utal, és gyakran kapcsolódik a spirituális vagy misztikus tapasztalatokhoz.<br>2. A transzcendens élmények olyan intenzív és mély élmények, amelyek túlmutatnak az érzékelhető világon, és gyakran egyfajta megvilágosodással, békeélménnyel vagy egyesülési élménnyel járnak.<br>3. A transzcendens élmények gyakran meghatározó és megváltoztató hatással lehetnek az emberek életére, inspirálva őket a személyes fejlődésre vagy a vallási gyakorlatokba való mélyebb elmerülésre.                                                                                                                                                                                                                       |
| User:      | legyen négy (‘let there be four’)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| Assistant: | 1. A transzcendens olyan fogalom, amely a mindennapi élettől való fizikai vagy szellemi távolságra utal, és gyakran kapcsolódik a spirituális vagy misztikus tapasztalatokhoz.<br>2. A transzcendens élmények olyan intenzív és mély élmények, amelyek túlmutatnak az érzékelhető világon, és gyakran egyfajta megvilágosodással, békeélménnyel vagy egyesülési élménnyel járnak.<br>3. A transzcendens élmények gyakran meghatározó és megváltoztató hatással lehetnek az emberek életére, inspirálva őket a személyes fejlődésre vagy a vallási gyakorlatokba való mélyebb elmerülésre.<br>4. A transzcendens több mint a szó szoros értelmében vett élmény lehet, hanem egyfajta tudatállapot vagy spirituális tudatosság, mely gyakran meditáció vagy transzcendentális gyakorlatok segítségével érhető el. |

These results also contribute to a possible framework of developing conversational models for small languages through transfer learning and the application of

continual pre-praining methods. Still, it is worth noting that while our research delivered measurable improvements, the results were limited by the need for manual dataset curation, which remains a resource-intensive and time-consuming process.

In the future, we plan to evaluate our conversational models using LLM evaluation benchmarks, as two new benchmarks for Hungarian, HuGME [16] and Open-HuEval [32], have been released in recent weeks.

## References

- [1] J. ÁCS: *Exploring bert's vocabulary*, 2019, URL: <https://juditacs.github.io/2019/02/19/bert-tokenization-stats.html> (visited on 02/12/2025).
- [2] ANTHROPIC: *Claude 3: Haiku, Sonnet, and Opus — the next generation of Claude models*, blog post / online announcement, Available at <https://www.anthropic.com/research/claude-3-family>, 2024.
- [3] B. BARTA, D. LAKATOS, A. NAGY, M. K. NYIST, J. ÁCS: *HunSum-1: an Abstractive Summarization Dataset for Hungarian*, in: XIX. Hungarian Computational Linguistics Conference, Szeged, Hungary: University of Szeged, 2023, pp. 231–243.
- [4] S. BLACK, S. BIDERMAN, E. HALLAHAN, Q. ANTHONY, L. GAO, L. GOLDING, H. HE, C. LEAHY, K. McDONELL, J. PHANG, M. PIELER, U. S. PRASHANTH, S. PUROHIT, L. REYNOLDS, J. TOW, B. WANG, S. WEINBACH: *GPT-NeoX-20B: An Open-Source Autoregressive Language Model*, in: Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, ed. by A. FAN, S. ILIC, T. WOLF, M. GALLÉ, virtual+Dublin: Association for Computational Linguistics, May 2022, pp. 95–136, DOI: [10.18653/v1/2022.bigscience-1.9](https://doi.org/10.18653/v1/2022.bigscience-1.9), URL: <https://aclanthology.org/2022.bigscience-1.9/>.
- [5] M. CONOVER, M. HAYES, A. MATHUR, J. XIE, J. WAN, S. SHAH, A. GHODSI, P. WENDELL, M. ZAHARIA, R. XIN: *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*, 2023, URL: <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm> (visited on 06/30/2023).
- [6] Z. CSAKI, B. LI, J. L. LI, Q. XU, P. PAWAKAPAN, L. ZHANG, Y. DU, H. ZHAO, C. HU, U. THAKKER: *SambaLingo: Teaching Large Language Models New Languages*, in: Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024), ed. by J. SÄLEVÄ, A. OWODUNNI, Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 1–21, DOI: [10.18653/v1/2024.mrl-1.1](https://doi.org/10.18653/v1/2024.mrl-1.1), URL: <https://aclanthology.org/2024.mrl-1.1/>.
- [7] G. DEEPMIND: *Introducing Gemini: our most capable and general model*, <https://blog.google/technology/ai/google-gemini-ai/>, Accessed: 2025-09-29, 2023.
- [8] DEEPSEEK-AI: *DeepSeek LLM: Scaling Open-Source Language Models with Longtermism*, arXiv preprint arXiv:2401.02954 (2024), URL: <https://github.com/deepseek-ai/DeepSeek-LLM>.
- [9] F. FELLETÁR, G. GOSZTOLYA, Z. G. YANG, I. HOFFMANN, A. BABARCZY, Z. S. UNOKA: *Associations between coherence and temporal parameters of narrative speech production in borderline personality disorder*, Journal of Psychiatric Research 188 (2025), pp. 218–228, ISSN: 0022-3956, DOI: [10.1016/j.jpsychires.2025.05.063](https://doi.org/10.1016/j.jpsychires.2025.05.063), URL: <https://www.sciencedirect.com/science/article/pii/S0022395625003632>.
- [10] A. GRATTAFIORI ET AL.: *The Llama 3 Herd of Models*, 2024, arXiv: [2407.21783 \[cs.AI\]](https://arxiv.org/abs/2407.21783), URL: <https://arxiv.org/abs/2407.21783>.
- [11] S. GUGGER, L. DEBUT, T. WOLF, P. SCHMID, Z. MUELLER, S. MANGRULKAR, M. SUN, B. BOSSAN: *Accelerate: Training and inference at scale made simple, efficient and adaptable*. <https://github.com/huggingface/accelerate>, 2022.

- [12] P.-L. HSU, Y. DAI, V. KOTHAPALLI, Q. SONG, S. TANG, S. ZHU, S. SHIMIZU, S. SAHNI, H. NING, Y. CHEN, Z. WANG: *Liger-Kernel: Efficient Triton Kernels for LLM Training*, in: *Championing Open-source Development in ML Workshop @ ICML25*, 2025, URL: <https://openreview.net/forum?id=36SjAIT42G>.
- [13] E. J. HU, yelong SHEN, P. WALLIS, Z. ALLEN-ZHU, Y. LI, S. WANG, L. WANG, W. CHEN: *LoRA: Low-Rank Adaptation of Large Language Models*, in: *International Conference on Learning Representations*, 2022, URL: <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [14] A. KÖPF, Y. KILCHER, D. VON RÜTTE, S. ANAGNOSTIDIS, Z.-R. TAM, K. STEVENS, A. BARHOUM, N. M. DUC, O. STANLEY, R. NAGYFI, S. ES, S. SURI, D. GLUSHKOV, A. DANTU-LURI, A. MAGUIRE, C. SCHUHMAN, H. NGUYEN, A. MATTICK: *OpenAssistant conversations - democratizing large language model alignment*, in: *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, New Orleans, LA, USA: Curran Associates Inc., 2023.
- [15] N. LIGETI-NAGY, G. FERENCZI, E. HÉJA, L. J. LAKI, N. VADÁSZ, Z. G. YANG, T. VÁRADI: *HuLU: Hungarian Language Understanding Benchmark Kit*, in: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ed. by N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI, N. XUE, Torino, Italia: ELRA and ICCL, May 2024, pp. 8360–8371, URL: <https://aclanthology.org/2024.lrec-main.733/>.
- [16] N. LIGETI-NAGY, G. MADARASZ, F. FOLDESI, M. LENGVEL, M. OSVATH, B. SAROSSY, K. VARGA, G. Z. YANG, E. HÉJA, T. VÁRADI, G. PRÓSZÉKY: *HuGME: A benchmark system for evaluating Hungarian generative LLMs*, in: *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM<sup>2</sup>)*, ed. by O. ARVIV, M. CLINCIU, K. Dhole, R. DROR, S. GEHRMANN, E. HABBA, I. ITZHAK, S. MILLE, Y. PERLITZ, E. SANTUS, J. SEDOC, M. SHMUELI SCHEUER, G. STANOVSKY, O. TAFJORD, Vienna, Austria and virtual meeting: Association for Computational Linguistics, July 2025, pp. 385–403, ISBN: 979-8-89176-261-9, URL: <https://aclanthology.org/2025.gem-1.32/>.
- [17] S. LONGPRE, Y. LU, J. DAIBER: *MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering*, *Transactions of the Association for Computational Linguistics* 9 (2021), ed. by B. ROARK, A. NENKOVA, pp. 1389–1406, DOI: 10.1162/tacl\_a\_00433, URL: <https://aclanthology.org/2021.tacl-1.82/>.
- [18] P. H. MARTINS, P. FERNANDES, J. ALVES, N. M. GUERREIRO, R. REI, D. M. ALVES, J. POMBAL, A. FARAJAN, M. FAYSSE, M. KLIMASZEWSKI, P. COLOMBO, B. HADDOW, J. G. C. DE SOUZA, A. BIRCH, A. F. T. MARTINS: *EuroLLM: Multilingual Language Models for Europe*, 2024, arXiv: 2409.16235 [cs.CL], URL: <https://arxiv.org/abs/2409.16235>.
- [19] T. NGUYEN, C. V. NGUYEN, V. D. LAI, H. MAN, N. T. NGO, F. DERNONCOURT, R. A. ROSSI, T. H. NGUYEN: *CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages*, in: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ed. by N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI, N. XUE, Torino, Italia: ELRA and ICCL, May 2024, pp. 4226–4237, URL: <https://aclanthology.org/2024.lrec-main.377>.
- [20] A. NOVÁK, B. NOVÁK, T. ZOMBORI, G. SZABÓ, Z. SZÁNTÓ, R. FARKAS: *A Question Answering Benchmark Database for Hungarian*, in: *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, ed. by J. PRANGE, A. FRIEDRICH, Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 188–198, DOI: 10.18653/v1/2023.law-1.19, URL: <https://aclanthology.org/2023.law-1.19/>.
- [21] OPENAI: *Introducing ChatGPT*, <https://openai.com/blog/chatgpt>, Accessed: 2025-07-24, 2022.
- [22] M. POPOVIĆ: *chrF: character n-gram F-score for automatic MT evaluation*, in: *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 392–395, DOI: 10.18653/v1/W15-3049, URL: <https://aclanthology.org/W15-3049>.

- [23] M. POST: *A Call for Clarity in Reporting BLEU Scores*, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191, URL: <https://www.aclweb.org/anthology/W18-6319>.
- [24] N. PRÓTÁR, D. M. NEMESKEY: *HunSimpleNews: Az első autentikus magyar nyelvű szövegek-ből álló szövegegyszerűsítési korpusz [The first text simplification corpus consisting of authentic Hungarian-language texts]*, in: XXI. Hungarian Computational Linguistics Conference, Szeged, Hungary, 2025, pp. 197–218.
- [25] E. SIMON, N. VADÁSZ: *Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus*, in: Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings, ed. by K. EKSTEIN, F. PÁRTL, M. KONOPÍK, vol. 12848, Lecture Notes in Computer Science, Springer, 2021, pp. 222–234, DOI: [10.1007/978-3-030-83527-9\\_19](https://doi.org/10.1007/978-3-030-83527-9_19).
- [26] S. SINGH, F. VARGUS, D. D’SOUZA, B. KARLSSON, A. MAHENDIRAN, W.-Y. KO, H. SHANDILYA, J. PATEL, D. MATACIUNAS, L. O’MAHONY, M. ZHANG, R. HETTIARACHCHI, J. WILSON, M. MACHADO, L. MOURA, D. KRZEMIŃSKI, H. FADAEI, I. ERGUN, I. OKOH, A. ALAAGIB, O. MUDANNAYAKE, Z. ALYAFEAI, V. CHIEN, S. RUDER, S. GUTHIKONDA, E. ALGHAMDI, S. GEHRMANN, N. MUENNIGHOFF, M. BARTOLO, J. KREUTZER, A. ÜSTÜN, M. FADAEI, S. HOOKER: *Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning*, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ed. by L.-W. KU, A. MARTINS, V. SRIKUMAR, Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 11521–11567, DOI: [10.18653/v1/2024.acl-long.620](https://doi.org/10.18653/v1/2024.acl-long.620), URL: <https://aclanthology.org/2024.acl-long.620/>.
- [27] R. TAORI, I. GULRAJANI, T. ZHANG, Y. DUBOIS, X. LI, C. GUESTRIN, P. LIANG, T. B. HASHIMOTO: *Stanford Alpaca: An Instruction-following LLaMA model*, 2023, URL: [https://github.com/tatsu-lab/stanford5C\\_alpaca7D](https://github.com/tatsu-lab/stanford5C_alpaca7D).
- [28] G. TEAM: *Gemma 3* (2025), URL: <https://arxiv.org/abs/2503.19786>.
- [29] H. TOUVRON ET AL.: *Llama 2: Open Foundation and Fine-Tuned Chat Models* (2023), arXiv: [2307.09288](https://arxiv.org/abs/2307.09288) [cs.CL].
- [30] H. TOUVRON ET AL.: *Llama 2: Open Foundation and Fine-Tuned Chat Models*, 2023, arXiv: [2307.09288](https://arxiv.org/abs/2307.09288) [cs.CL], URL: <https://arxiv.org/abs/2307.09288>.
- [31] A. YANG ET AL.: *Qwen2 Technical Report*, arXiv preprint arXiv:2407.10671 (2024).
- [32] H. YANG, X. WEI, J. WU, N. LIGETI-NAGY, J. SUN, Y. WANG, G. Z. YANG, J. GAO, J. WANG, B. JIANG, S. WANG, N. YU, Z. ZHANG, S. HONG, H. LIU, W. LI, S. ZHANG, D. LIN, L. WU, G. PRÓSZÉKY, C. HE: *OpenHuEval: Evaluating Large Language Model on Hungarian Specifics*, in: Findings of the Association for Computational Linguistics: ACL 2025, ed. by W. CHE, J. NABENDE, E. SHUTOVA, M. T. PILEHVAR, Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 7464–7520, ISBN: 979-8-89176-256-5, URL: <https://aclanthology.org/2025.findings-acl.390/>.
- [33] Z. G. YANG: *Neural text summarization for Hungarian*, Acta Linguistica Academica 69.4 (2022), pp. 474–500, DOI: [10.1556/2062.2022.00577](https://doi.org/10.1556/2062.2022.00577), URL: <https://akjournals.com/view/journals/2062/69/4/article-p474.xml>.
- [34] Z. G. YANG, R. DODÉ, G. FERENCZI, P. HATVANI, E. HÉJA, G. MADARÁSZ, N. LIGETI-NAGY, B. SÁROSSY, Z. SZANISZLÓ, T. VÁRADI, T. VEREBÉLYI, G. PRÓSZÉKY: *The First Instruction-Following Large Language Models for Hungarian*, in: 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) Proceedings, Debrecen, Hungary: University of Debrecen, 2024, pp. 247–252, ISBN: 9798350387889.
- [35] Z. G. YANG, R. DODÉ, G. FERENCZI, H. PÉTER, E. HÉJA, M. LENGYEL, N. LIGETI-NAGY, G. MADARÁSZ, B. SÁROSSY, K. VARGA, T. VARGA, T. VÁRADI: *PULI LlmIX modell - Egy folytatólagosan előtanított nagy nyelvi modell [PULI LlmIX Model – A Continual Pre-Trained Large Language Model]*, in: XXI. Hungarian Computational Linguistics Conference, Szeged, Hungary: University of Szeged, 2025, pp. 153–167.

- [36] Z. G. YANG, L. J. LAKI, T. VÁRADI, G. PRÓSZÉKY: *Mono- and multilingual GPT-3 models for Hungarian*, in: Text, Speech, and Dialogue, Lecture Notes in Computer Science, Plzeň, Czech Republic: Springer Nature Switzerland, 2023, pp. 94–104, ISBN: 978-3-031-40498-6.
- [37] S. YUAN, H. ZHAO, Z. DU, M. DING, X. LIU, Y. CEN, X. ZOU, Z. YANG, J. TANG: *WuDao-Corpora: A super large-scale Chinese corpora for pre-training language models*, AI Open 2 (2021), pp. 65–68, ISSN: 2666-6510, DOI: [10.1016/j.aiopen.2021.06.001](https://doi.org/10.1016/j.aiopen.2021.06.001).
- [38] W. ZHAO, X. REN, J. HESSEL, C. CARDIE, Y. CHOI, Y. DENG: *WildChat: 1M ChatGPT Interaction Logs in the Wild*, in: The Twelfth International Conference on Learning Representations, 2024, URL: <https://openreview.net/forum?id=Bl8u7ZR1bM>.
- [39] Y. ZHAO, A. GU, R. VARMA, L. LUO, C.-C. HUANG, M. XU, L. WRIGHT, H. SHOJANAZERI, M. OTT, S. SHLEIFER, A. DESMAISON, C. BALIOGLU, P. DAMANIA, B. NGUYEN, G. CHAUHAN, Y. HAO, A. MATHEWS, S. LI: *PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel*, 2023, arXiv: [2304.11277](https://arxiv.org/abs/2304.11277) [cs.DC], URL: <https://arxiv.org/abs/2304.11277>.
- [40] Y. ZHENG, R. ZHANG, J. ZHANG, Y. YE, Z. LUO: *LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models*, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), ed. by Y. CAO, Y. FENG, D. XIONG, Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 400–410, DOI: [10.18653/v1/2024.acl-demos.38](https://doi.org/10.18653/v1/2024.acl-demos.38), URL: <https://aclanthology.org/2024.acl-demos.38/>.