

PULI Chat

Our 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

HUN-REN Hungarian Research Centre for Linguistics
(familyname).(givenname)@nytud.hun-ren.hu

Abstract

In recent months, 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 by OpenAI showcased unprecedented capabilities, achieved through a multi-stage fine-tuning process.

For Hungarian, the first instruction-following model was introduced by Yang et al. [5], based on the LLaMA-2 architecture [3]. 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 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. [5], we applied continual pre-training to the Llama 3.1 8B Instruct [1] and Qwen2.5 7B Instruct [4] 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**.

Table 1. Performance of the PULI-LlumiX-Llama-3.1 and PULI Trio Q models, on the HuLU benchmarks

	HuCOLA (MCC)	HuCoPA (MCC)	HuRTE (MCC)	HuSST (ACC)
PULI LlumiX (Llama 2)	70.3	64.2	68.2	81.9
PULI-LlumiX-Llama-3.1 base	69.2	74.1	71.9	82.2
PULI-LlumiX-Llama-3.1 chat	72.54	74.1	76.0	83.1
PULI Trio Q chat	62.9	67.1	68.8	82.1

For continual pre-training, we used the same corpora employed in the research by Yang et al. [5], but we also included the Hungarian Wikipedia to strengthen the model’s knowledge of the Hungarian language. During the pre-training experiments, we observed that the model generated more grammatical 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. Therefore, the dataset consisted of carefully selected Hungarian articles published within the last six years. Altogether, our pre-training corpus contains 9,704,312,719 words.

For fine-tuning, we significantly expanded the instruction dataset used in the research by Yang et al. [5], and replaced the Stanford Alpaca instruction template with the Llama 3 and Qwen2 chat templates to ensure consistency with each base model’s instruction format. Our final fine-tuning dataset contains a total of 44,626 segments. Both the pre-training and fine-tuning processes were conducted using the LLaMA-Factory implementation.

Table 1 presents the performance of our new models. For evaluation, we further fine-tuned the models on four HuLU benchmarks [2] using the LoRA and full parameter fine-tuning method. As the results show, in most cases, our new models outperformed the previous LlumiX versions. Furthermore, by fine-tuning the chat model specifically on the benchmark tasks, we achieved significantly higher results across all tasks.

A sample conversation is shown in Table 2, demonstrating the model’s 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 2, the model consistently provided three relevant features for each entity, and correctly responded with four when requested.

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. [5], 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

Table 2. Example snippet of a conversation

User:	mondj három tulajdonságot a kutyákról (‘tell me three characteristics about dogs.’)
Assistant:	1. A kutyák hűségesek és odaadóak gazdáik iránt. 2. A kutyák képesek felismerni és reagálni az emberi érzelmekre. 3. A kutyák nagyszerű képességgel rendelkeznek a szagokat ...
User:	csigák? (‘snails?’)
Assistant:	1. A csigák ügyesek a mászásban és kiválóan képesek ... 2. A csigák képesek kemény héjú otthonukat önmagukban ... 3. A csigák nagyon válogatósak és csak bizonyos növényeket ...
...	...
User:	és a transzcendens? (‘and the transcendent?’)
Assistant:	1. A transzcendens olyan fogalom, amely ... 2. A transzcendens élmények olyan intenzív ... 3. A transzcendens élmények gyakran meghatározó ...
User:	legyen négy (‘let there be four’)
Assistant:	1. A transzcendens olyan fogalom, amely ... 2. A transzcendens élmények olyan intenzív ... 3. A transzcendens élmények gyakran meghatározó ... 4. A transzcendens több mint a szó szoros értelmében vett ...

LlumiX versions.

References

- [1] A. GRATTAFIORI ET AL.: *The Llama 3 Herd of Models*, 2024, arXiv: [2407.21783](https://arxiv.org/abs/2407.21783) [cs.AI], URL: <https://arxiv.org/abs/2407.21783>.
- [2] 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/>.
- [3] 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].
- [4] A. YANG ET AL.: *Qwen2 Technical Report*, arXiv preprint arXiv:2407.10671 (2024).
- [5] 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.