Toward Hungarian-Centric Language Understanding: Hungarian-Adapted PULI Large Language Models

Zijian Győző Yang, Ágnes Bánfi, Réka Dodé, Gergő Ferenczi, Flóra Földesi, Enikő Héja, Mariann Lengyel, Gábor Madarász, Mátyás Osváth, Bence Sárossy, Kristóf Varga, Noémi Ligeti-Nagy
ELTE Research Centre for Linguistics
Budapest, Hungary
(family name).(given name)@nytud.hu

Abstract

1 Background

In recent months, large language models (LLMs) have drawn significant attention as companies race to develop systems capable of solving diverse NLP tasks. OpenAI's ChatGPT demonstrated unprecedented capabilities through multi-stage fine-tuning.

2 Methods and Experiments

For Hungarian, the first instruction-following model, based on LLaMA-2, was introduced by Z. G. Yang et al., 2024, showing improved performance on local benchmarks. However, it still lacked natural conversational abilities. Conversational fine-tuning requires resource-intensive, multi-turn datasets, which are scarce for Hungarian. Building on Yang et al.'s approach, we used continual pre-training to adapt the Llama 3.1 8B Instruct (Grattafiori et al., 2024) and Qwen2.5 7B Instruct (A. Yang et al., 2024) models for Hungarian.

For continual pre-training, we applied the same methods and used the same corpora as Z. G. Yang et al., 2024. However, during training, we observed an increase in grammaticality errors. To address this, we concluded with a Hungarian-only dataset, aiming to improve both linguistic accuracy and the model's knowledge of recent events. This final dataset consisted of carefully selected Hungarian articles published within the last six years. The corpora used for the final stage of training were:

- Hungarian Wikipedia (2019–2025): 174,945,601 words
- Hungarian news articles (2019–2025): 622,275,769 words, in the following domains:
  - daily news, celebrity, tax, aviation, economy, family, healthcare, fashion, education, culture, art, lifestyle, music, movies, law

After continual pre-training, we performed supervised fine-tuning (SFT) using multilingual instruction datasets that included 44,626 Hungarian instruction segments. For the Llama model, we supplemented this with 85,775 English prompts, drawing from the Stanford Alpaca dataset (Taori et al., 2023) and LLaMA-Instruct[1] datasets. For the Qwen-based model, we additionally incorporated 50,130 Chinese prompts from the neo_sft_phase2[2] dataset, alongside the English ones.

The main characteristics of continual pre-training corpora are shown in Table 1.

To evaluate and compare the effectiveness of Hungarian language adaptation, we performed SFT on Llama and Qwen models at various stages of training. The following SFT variants were included in our evaluation:

[1] https://huggingface.co/datasets/togethercomputer/llama-instruct
[2] https://huggingface.co/datasets/m-a-p/neo_sft_phase2

	Documents	Words	Average document length 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 (Characters)	16 398.17 / 14 767 (Characters)
Final stage (hu)	1 411 979	797 221 370	443.65 / 217

Table 1: Corpus characteristics for continual pre-training

	bias	toxicity	relevance	faithfulness	summary
Llama Original	73.47	89	65	81	63.27
Llama Original SFT	88.78	94	78	99	57.14
Llama CP SFT	81.63	82	91	98	26.53
Llama CP-H-1 SFT	85.71	85	89	99	28.57
Llama CP-H-2 SFT	90.82	88	93	100	34.69
Qwen Original	87.76	98	53	98	30.61
Qwen Original SFT	80.61	86	56	100	32.65
Qwen CP SFT	78.57	85	87	98	36.73
Qwen CP-H-1 SFT	81.63	91	90	99	22.45
Qwen CP-H-2 SFT	80.61	82	89	99	24.49

Table 2: HuGME evaluation part 1

- **Llama/Qwen Original**: The original Llama 3.1 8B Instruct and Qwen2.5 7B Instruct models without any fine-tuning.
- **Llama/Qwen Original SFT**: The original models fine-tuned directly with supervised instruction data.
- **Llama/Qwen CP SFT**: Models that underwent continual pre-training followed by supervised fine-tuning.
- **Llama/Qwen CP-H-1 SFT**: Models with continual pre-training, followed by one epoch of Hungarian-only training, then supervised fine-tuning.
- **Llama/Qwen CP-H-2 SFT**: Models with continual pre-training, followed by two epochs of Hungarian-only training, then supervised fine-tuning.

For evaluation and analysis, we tested the models on the HuGME benchmarks (Ligeti-Nagy et al., 2025), covering the following tasks:

- Bias, Toxicity, Relevance, Faithfulness, Summary, Prompt Alignment, Readability, Spell Checking, TruthfulQA, and MMLU.

3 Results and Evaluation

Figure 1 and Figure 2 show the evaluation of models on the HuGME benchmarks. A key observation is that the Llama and Qwen models respond differently to the applied training strategies. For the Llama variants, continual pre-training followed by Hungarian-only fine-tuning (CP-H) consistently improved performance across most metrics, with the best results observed after two epochs of Hungarian-specific training. In contrast, the Qwen models showed limited additional

	prompt alignment	readability	spell checking	truthfulQA	MMLU
Llama Original	29	70.7	94.895	23.03	46.63
Llama Original SFT	32	77.3	94.078	48.59	43.46
Llama CP SFT	27	79.7	100	25.44	46.98
Llama CP-H-1 SFT	32	68	100	28.78	45.44
Llama CP-H-2 SFT	35	75.9	100	38.55	43.89
Qwen Original	33	76.1	93.32	40.03	46.17
Qwen Original SFT	35	76.5	94.308	33.07	50.26
Qwen CP SFT	44	73.9	94.55	51.14	58.14
Qwen CP-H-1 SFT	62	77.3	94.78	50.07	57.69
Qwen CP-H-2 SFT	51	68.7	94.58	50.87	57.5

Table 3: HuGME evaluation part 2

benefits from further training. While CP-H-1 SFT brought noticeable gains, the second Hungarian epoch (CP-H-2 SFT) did not lead to further improvements and even resulted in slightly lower performance in several metrics. This pattern suggests the possibility of overfitting or a reduction in generalization capacity at that stage of training.

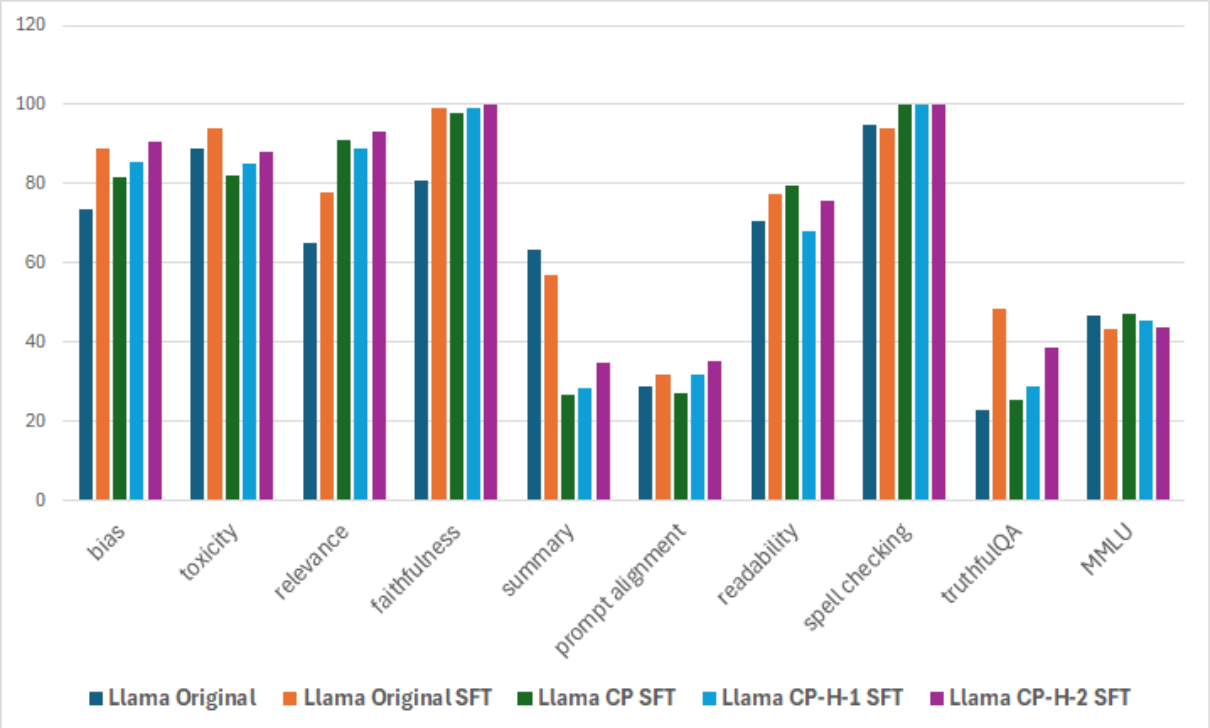

Figure 1: Llama models evaluation on HuGME benchmarks

The continual pre-training process may impair the model's original instruction-following or task-solving abilities (such as toxicity, summary). However, these capabilities are potentially recoverable through reinforcement learning methods, which were not applied in our current experiments. More importantly, core language skills—such as spell checking, readability, and prompt alignment—demonstrated significant performance gains, highlighting the effectiveness of the training strategy in improving foundational linguistic abilities.

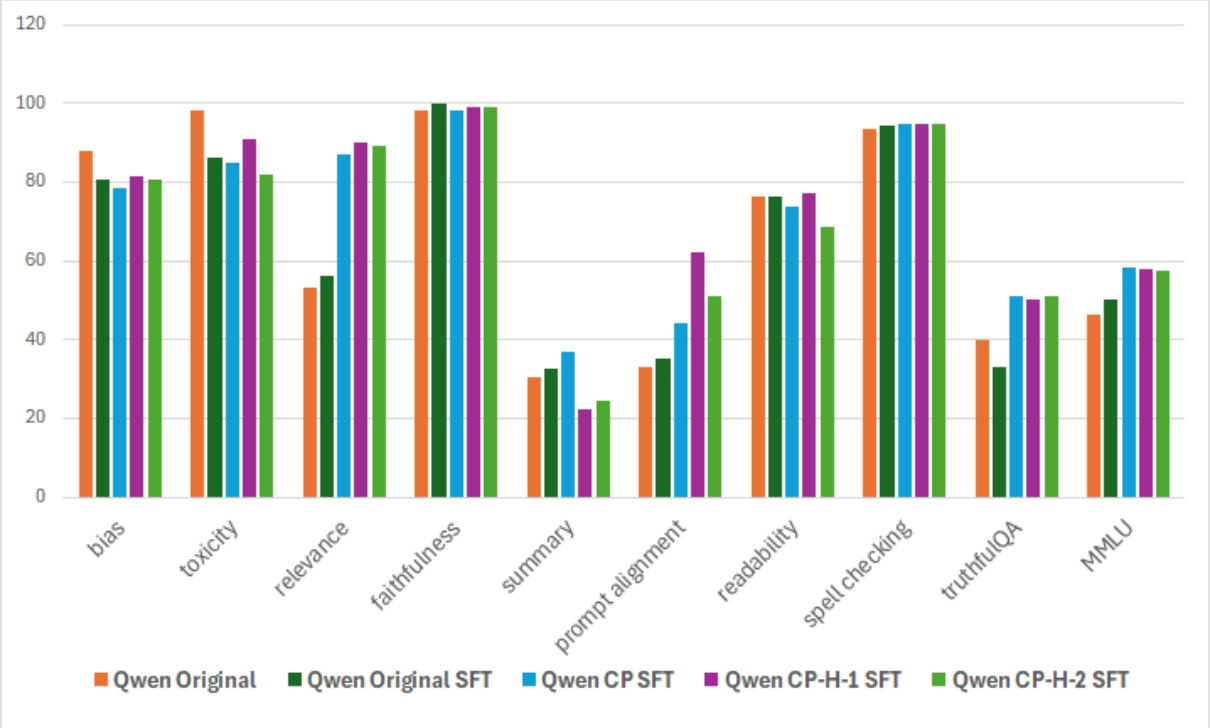

Figure 2: Qwen and Qwen models evaluation on HuGME benchmarks

Fine-tuning significantly enhanced relevance and faithfulness. Llama Original had relevance/faithfulness scores of 65/81, improving to 93/100 in CP-H-2 SFT. Qwen Original also showed similar gains: from 53/98 to 89/99 in CP-H-2 SFT. CP and Hungarian fine-tuning steps appear crucial, especially in Llama, where they yielded the sharpest improvements.

In the spell checking task, more Llama CP variants achieved perfect scores (100), while the Qwen variants remained in the 93–95 range. This result is expected, as the Qwen model is primarily Chinese-English centric, making it more distant from Hungarian compared to the English-centric Llama model.

This evaluation demonstrates that while multilingual specialization can enhance alignment and fairness metrics, it must be carefully balanced against open-ended generation and general-domain reasoning capabilities.

	HuCOLA	HuRTE	HuSST	MT en-hu	MT hu-en
Llama Original	57.64	77.46	70.85	19.23 / 50.95	34.92 / 61.08
Llama CP	60.76	69.38	71.23	22.81 / 53.98	34.30 / 60.11
Llama CP-H-1	65.17	75.72	71.88	24.27 / 55.56	34.82 / 61.10
Llama CP-H-2	**65.61**	**76.27**	**72.74**	**25.23 / 56.08**	**35.34 / 61.34**
Qwen Original	57.21	62.45	61.51	9.83 / 38.38	25.38 / 55.52
Qwen CP	61.89	76.20	70.87	20.70 / 51.49	32.10 / 59.76
Qwen CP-H-1	**67.40**	**79.52**	**73.67**	**21.45 / 52.34**	**32.48 / 60.00**
Qwen CP-H-2	63.85	76.92	73.34	20.66 / 52.30	32.25 / 59.83

Table 4: Zero-shot evaluation

4 Conclusion

Overall, the results indicate that language adaptation through continual pre-training and Hungarian-specific fine-tuning is effective, particularly in improving core linguistic capabilities. However, the impact of the second Hungarian epoch varies across models. While Llama benefits consistently from two Hungarian-specific epochs, showing continued improvements across multiple metrics, the Qwen model demonstrates signs of performance degradation after the second epoch. This suggests that while language adaptation is generally beneficial, its effectiveness depends on the underlying architecture and pre-training distribution of the base model, highlighting the importance of model-specific tuning strategies.

References

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). The llama 3 herd of models. https://arxiv.org/abs/2407.21783

Ligeti-Nagy, N., Madarász, G., Földesi, F., Lengyel, M., Osváth, M., Sárossy, B., Varga, K., Yang, Z. G., Héja, E., Váradi, T., & Prószéky, G. (2025). HuGME: A benchmark system for evaluating Hungarian generative LLMs [In press]. *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM 2025)*.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Stanford Alpaca: An Instruction-following LLaMA model. %5Curl%7Bhttps://github.com/tatsu-lab/stanford%5C_alpaca%7D

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., … Fan, Z. (2024). Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yang, Z. G., Dodé, R., Ferenczi, G., Hatvani, P., Héja, E., Madarász, G., Ligeti-Nagy, N., Sárossy, B., Szaniszló, Z., Váradi, T., Verebélyi, T., & Prószéky, G. (2024). The first instruct-following large language models for hungarian. *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) Proceedings*, 247–252.