Mono- and multilingual GPT-3 models for Hungarian

Zijian Győző Yang, László János Laki, Tamás Váradi, and Gábor Prószéky

Hungarian Research Centre for Linguistics, H-1068 Budapest, Benczúr str. 33.
www.nytud.hu
{yang.zijian.gyozo,laki.laszlo,varadi.tamas,proszeky.gabor}@nytud.hu

Abstract. In recent years, the growth in size of Transformer-based language models has accelerated significantly. Global technology companies are training larger and larger models that require enormous resources and training data. With these experiments, they aim to demonstrate that sufficiently large models with abundant training data can solve any natural language processing task even without fine-tuning. It may not be feasible to compete directly in this race, but there is an opportunity to conduct experiments in the direction of larger models in their shadow. Our aim is to train large language models for Hungarian. According to the knowledge transfer researches, a language model can adapt valuable knowledge from other languages. Furthermore, in order for the model to be able to solve translation tasks, it also needs multilingual knowledge. In our research, we trained a Hungarian monolingual and a Hungarian-English-Chinese trilingual 6.7 billion parameter GPT language model with more than 1TB text data. In our experiments, we also fine-tuned our model with the prompts provided by the Stanford Alpaca dataset. Thus, employing this methodology, an instruct GPT was built, which, as far as we know, is the first multilingual large language model in this region that can follow instructions.

Keywords: GPT-3, multilingual large language model, instruct GPT

1 Introduction

In recent years, there has been a race among major research centers and companies to develop larger and more parameter-rich language models. In 2021, when Microsoft and NVIDIA jointly created the Megatron-Turing NLG model with 530 billion parameters [27], the question was raised in an article[1] whether this competition could be the new Moore's Law. These studies attempt to demonstrate that with a large enough model trained on extensive data, a single large model can solve any language technology task without fine-tuning, relying solely on prompt programming. However, this competition requires enormous resources that only the largest global technology companies and research centers can afford. In the recent days, the GPT-4 model [20] with 1 trillion parameters was released.

Currently, the best-performing language models for the Hungarian language are PULI BERT-Large [36] and huBERT [18]. Although the HILBERT model [8] is larger

[1] https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model

in size than huBERT, its performance in available tests has been lower than huBERT, mainly due to being trained on less data. In June 2022, the HILANCO consortium introduced the HILANCO-GPTX, a 6.7 billion-parameter English-Hungarian bilingual GPT-3 model[2].

In our research, we trained a monolingual (Hungarian) and a trilingual (Hungarian-English-Chinese) GPT language model with 6.7 billion parameters. Our models are called **PULI**, which is a small-medium breed of Hungarian herding dog. Both of our models are freely available for research purposes on our Hugging Face page[3]:

– **NYTK/PULI-GPT-3SX**: Hungarian monolingual model.
– **NYTK/PULI-GPTrio**: Hungarian-English-Chinese trilingual model.

2 Related Work

Currently, one of the largest models in the world is the recently released GPT-4 with more than 1 trillion parameters [20]. The GPT-4 is a large multimodal model which can accept image and text inputs and produce text outputs. Among text-only large language models (LLM), one of the largest model is the PaLM (Pathways Language Model) by Google, which has 540 billion parameters [7]. In addition to increasing its size, the model introduced the Pathways architecture, which aims to enable the model to learn multiple tasks simultaneously. The Pathways architecture implements a modified version of the traditional transformer [32] architecture with only a decoder. The modifications are drawn from recent developments in the field, such as the SwiGLU activation function [25], parallel layering [33] in transformer blocks, RoPE embedding [28], and the use of SentencePiece [13]. Despite being slightly smaller in size compared to the PaLM model, it is a serious competitor to the Megatron-Turing NLG model mentioned in the introduction [27]. At this scale, the difference of 10 billion parameters goes unnoticed, but for example, no one has been able to train a model with 10 billion parameters for Hungarian language yet. In terms of parameter count, the PaLM model is still three times larger than the milestone GPT-3 [5], which generated significant attention in both the press and the natural language processing community upon its release. What made GPT-3 novel was that it was trained with a massive amount of data and had an order of magnitude more parameters than the state-of-the-art at that time. The model was capable of generating text that was similar to human writing. Moreover, without fine-tuning, using prompt programming with few-shot examples or even no examples at all, it could solve various natural language processing tasks in a zero-shot manner. GPT-3 models available in different sizes named as Davinci, Curie, Babbage, or Ada, each specializing in different types of tasks. In addition to the models mentioned so far, it's worth mentioning the Wu Dao 2.0 model[4]. The Wu Dao 2.0 model was introduced by the Beijing Academy of Artificial Intelligence (BAAI) in 2021. It is currently the largest neural model with 1.750 trillion parameters, this model is also a multimodal

[2] https://hilanco.github.io
[3] https://huggingface.co/NYTK
[4] https://towardsdatascience.com/gpt-3-scared-you-meet-wu-dao-2-0-a-monster-of-1-75-trillion-parameters-832cd83db484

model. Comparing this model with other language models is challenging because it was trained not only on text but also on images. The model was trained on the Pile English dataset [9], as well as 1.2TB of Chinese text and 2.5TB of images. The training was conducted using the FastMoE [10] system. The model has achieved 'state-of-the-art' results in multiple tasks. Over the past few years, models have been introduced one after another with increasing frequency. The predecessors of Megatron-Turing NLG are also worth mentioning, such as the 17.2 billion parameter Turing-NLG[5] or the 8.3 billion parameter Megatron-LM [26] models. In recent weeks, the Meta AI has published the LLaMA models [31], which is a collection of foundation language models ranging from 7B to 65B parameters. During the training, only publicly available datasets were used, the LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B.

Recently, many experiments were made in the field of building instruct and chat models based on large language models. One of the most successful chat applications is ChatGPT, which integrates reinforcement learning into the fine-tuning process [23]. Creating instruction and chat prompts need a huge amount of human effort, thus automatic extraction methods became popular. Wang et al. [34] used the *text-davinci-001* model to generate instructions for the task. Similar experiment was conducted by R. Taori et al. [30] with some modifications. LM-SYS implemented the FastChat [16], and its' adaptations [6] that provide fine-tuning processes and data to build chat application from language models.

3 Corpora

The first part of our research is collecting the training data. Based on previous studies, it is widely accepted that training large models requires a correspondingly large amount of data.

For training our Hungarian monolingual models, we utilized corpora from the sources that are described in Table 2. The text was not tokenized, and numbers were kept in their raw form. The text was not tokenized during the model training either. In the corpus, each line represents a paragraph, and empty lines are used to separate documents. The texts of the corpus consist of the following sources:

- **Webcorpus 2.0**: The Webcorpus 2.0 [17] was colleted by Dávid Márk Nemeskey from the Common Crawl[6] database. The data is from 2013 to April of 2019. The corpus contains more than 9 billion tokens. For our training process, we used the non-tokenized version.
- **Wikipedia**: The Hungarian Wikipedia, part of the Webcorpus 2.0.
- **Common Crawl (CC)**: Most of our Hungarian text was collected from the Common Crawl database. Since the Webcorpus 2.0 contains text only until April of 2019, we collected the the data that was created afterwards. For downloading and

[5] https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft

[6] https://commoncrawl.org

boilerplate-cleaning, we used the modified CC downloader script[7] that was originally implemented by Balázs Indig [11]. The CC collection consists of two parts:
  - .hu domain: Collection only from .hu domain.
  - non .hu domain: Collection from other than .hu domain, but in Hungarian language.
- **neticle**: Text collection from public social media posts and comments, which was collected by Neticle Kft.
- **JSI**: The Jožef Stefan Institute in Slovenia has been collecting news from internet sources (RSS feeds) in multiple languages since 2013 for the purposes of the different web services. We have utilized the Hungarian content.
- **araneum**: Araneum Hungaricum Maium[8] [1, 2, 24] corpus was compiled by Vladimir Benko.
- **hutenten**: The huTenTen corpus is part of the TenTen corpus family developed by Lexical Computing LLC [12], and serves as the Hungarian reference corpus for the SketchEngine platform. The corpus was compiled by Lexical Computing LLC based on a collection carried out in 2013 [29], and the Hungarian language analysis was conducted using the MNSZ1 [22] code by Csaba Oravecz, and the emMorph [19] code by Lexical Computing LLC.
- **news/press**: The dataset was collected for our previous research, primarily for text summarization tasks. It includes articles and their leads collected from various news portals, including index.hu, nol.hu, and HVG. There may be overlaps with the data collected from CC, and duplicate data was handled at the end of the process.
- **MNSZ2**: The MNSZ2 [21] is the renewed second version of MNSZ1, which contains more than 1 billion words.
- **OpenSubtitles**: The OpenSubtitles [15] is a collection of translated movie subtitles. In our research, we used the Hungarian monolingual subtitles.

At the end of the collection process, all the collected texts from various sources were concatenated, and then converted into a document-level `jsonline` format, where each line represents a json object containing a `text` field that contains the text of a document, preserving line breaks. Document-level deduplication (`uniq`) and random shuffling were performed on this json file.

For training the monolingual Hungarian model, the training corpus did not contain CC (non .hu), MNSZ2 and OpenSubtitles subcorpora (see sign $*$ in Table 1) yet.

In the case of trilingual model, beside the Hungarian corpus, the training corpus consists of texts from the following sources:

- English: The first 1/3 of The Pile [9] corpus. From index 00 to 09. In Table 2, we separately showed the Github data.
- Chinese (zh):
  - **Wu Dao 2.0** [37]: Public available part (200 GB).
  - **Common Crawl**: Custom collection from .cn domain from 2018 to 2022. we used the same script as for the Hungarian collection.
  - **Chinese Wikipedia**: Downloaded from the brightmart Github [35].

[7] https://github.com/DavidNemeskey/cc_corpus
[8] http://ucts.uniba.sk/aranea_about/_hungaricum.html

	Document	Paragraph	Word
Webcorpus 2.0	9 240 709	171 239 297	8 051 677 190
Wikipedia	418 622	6 804 115	124 982 493
CC (.hu) 2019–2022	28 902 005	690 761 866	20 860 935 871
∗ CC (non .hu) 2019–2022	11 685 663	387 600 105	10 877 153 207
neticle	30 471 970	85 351 213	1 112 740 383
jsi	4 023 083	32 363 186	1 077 066 597
araneum	3 727 984	31 721 824	1 329 200 470
hutenten	6 447 787	164 654 976	2 670 682 031
news/press	3 009 073	12 606 903	1 058 656 664
∗ MNSZ2	1 879	58 654 846	846 089 645
∗ OpenSubtitles	130 831	103 579 701	471 393 322

Table 1. Statistics of the Hungarian corpus

In Table 2, the main characteristics of the corpora from the three languages are described. In our research, we tried to balance the three languages.

	Document	Paragraph	Word / Character	Size (GB)
Hungarian	86 008 464	1 499 319 836	Word: 41 508 933 801	314
English	64 192 842	2 538 238 213	Word: 61 906 491 823	391
Github	6 018 366	-	-	33
Chinese	111 262 633	3 824 592 151	zh chars: 98 693 705 456 non zh token: 12 072 234 774	340

Table 2. Statistics of the trilingual training corpora

For building our models, we trained custom vocabularies:

– Hungarian model: the size of Hungarian monolingual vocabulary is 50 000.
– Trilingual model: Considering the variety of Chinese characters, the size of our vocabulary is 150 016.

4 Training models

To pretrain our GPT models, we used the GPT-NeoX implementation [3]. GPT-NeoX is a project by EleutherAI[9] with the aim of training large-scale language models, similar to GPT-3. Their implementation is based on NVIDIA Megatron-LM and DeepSpeed technologies. They have implemented various GPT-3-like configurations, ranging from small models (e.g., 160 million parameters) to large ones (175 billion parameters). In our research, we used a relatively small configuration with 6.7 billion parameters. We trained the model using an NVIDIA GDX A100 box containing 8 A100 (80GB) GPUs. The training was performed without modifying the hyperparameters, except for the

[9] https://www.eleuther.ai

micro batch size, which was empirically set to 16 (to fit within the 80GB GPU memory). The training information for the models are showed in Table 3.

	training steps	final lm loss	val lm loss	val lm loss ppl	training time
PULI-3SX (Hungarian)	150 000	2.03	2.17	8.76	3 weeks
PULI-GPTrio (trilingual)	400 000	2.22	2.25	9.47	3 months

Table 3. Main training information of models

4.1 Instruct trilingual GPT

Recently, many experiments were released in field of building instruct and chat models based on large language models. Create instruction and chat prompts need a huge amount of human effort, thus automatic extraction methods became popular. In our first step, we used the Stanford Alpaca implementation and data for fine-tuning our trilingual model. We used the same prompt template as the Stanford Alpaca. Based on the experiments conducted with ChatGPT, our hypothesis is that the model, after fine-tuning solely on English data, will also be able to follow instructions given in Hungarian language, thanks to transfer learning. Our instruct model can be tested on our demo site[10].

5 Evaluation and Results

We evaluated our Instruct trilingual model on Hungarian benchmark corpora released in 2022, the HuLU (Hungarian Language Understanding Benchmark Kit) [14] corpora. We applied measurements on Hungarian Corpus of Linguistic Acceptability (HuCOLA), Hungarian version of the Stanford Sentiment Treebank (HuSST) and Hungarian Recognizing Textual Entailment dataset (HuRTE) tasks.

In the case of monolingual model, we conducted few-shot learning to solve the tasks. We tried different set of hyper-parameters and prompts, we achieved the highest performance with the following settings in the different tasks:

- HuCOLA: prompt #: 27; temperature: 0.1; top-p: 0.12; top-k: 10;
- HuSST: prompt #: 29; temperature: 0.3; top-p: 0.1; top-k: 10;
- HuRTE: prompt #: 15; temperature: 0.3; top-p: 0.4; top-k: 10;

For choosing examples for prompts, we chose the first # record from the training set.

In the zero-shot experiments, we compared our model with the ChatGPT (gpt-3.5-turbo) and the GPT-3 (text-davinci-003) models [5].

In Table 4, you can see the results of our models. Since the test sets are unbalanced, we used the balanced accuracy [4] instead of accuracy metric.

In the few-shot experiments, the monolingual model could outperforme the trilingual model in all cases. It is an expected result from a monolingual model in a monolingual Hungarian task.

[10] https://juniper.nytud.hu/demo/gptrio

	HuCOLA	HuSST	HuRTE
PULI-3SX few-shot	54.27	64.27	57.42
PULI-GPTrio few-shot	52.71	61.58	54.54
ChatGPT zero-shot	49.10	36.99	50.26
text-davinci-001 zero-shot	50.78	35.48	49.06
Instruct PULI-GPTrio zero-shot	52.12	59.20	58.14

Table 4. Performance of the models on HuLU benchmark

One significant findings is that our fine-tuned instruct model achieved competitive or higher results with zero-shot learning. In the zero-shot experiments, in all cases, our instruct model could outperformed the ChatGPT and the GPT-3 models.

In our multilingual experiments, the most expected result is that our model could gain knowledge from other languages. In our instruction task, in the current stage, we fine-tuned our trilingual model only on the English Alpaca data set. Although the model may not have been specifically fine-tuned on Hungarian or Chinese prompts, it is still capable of understanding and following Hungarian or Chinese instructions, and can effectively solve tasks in the given languages.

Instruction:
Ki a legerősebb szuperhős?
(Who is the strongest superhero?)

Response:
A legerősebb szuperhős egyértelműen Thor, a Marvel-képregények egyik legerősebb és legismertebb alakja. Thor a Marvel Univerzum legerősebb és leghatalmasabb hősei közé tartozik, ereje és gyorsasága a legnagyobb hősökével vetekszik. A Marvel Univerzum legerősebb szuperhőseinek listáján a második helyet a Marvel kapitány, a harmadik helyet pedig a Vasember foglalja el.
(The strongest superhero is undoubtedly Thor, one of the most powerful and well-known characters in Marvel comics. Thor ranks among the mightiest and most formidable heroes in the Marvel Universe, with strength and speed that rivals that of the greatest heroes. Captain Marvel takes the second place on the list of the most powerful superheroes in the Marvel Universe, followed by Iron Man in third place.)

Instruction:
谁是最强壮的超级英雄?
(Who is the strongest superhero?)

Input:
汉语回答。
(Give the answer in Chinese.)

Response:
绿巨人（Hulk）是世界上最强壮的超级英雄。
(Hulk is the strongest superhero in the world.)

Table 5. A text generation sample for Hungarian and Chinese instructions

We illustrate this phenomenon in Table 5, Table 6 and Table 7. In the examples, the original instruction did not contain English translation, thus, for better understanding, we provided the translation in brackets with italic font. We did not provide English samples.

In Table 5, you can see a sample for text generation. In the case of Chinese, we had to add an additional instruction in the input part: 汉语回答。, which means 'Give the answer in Chinese.'. Without this instruction, the answer were given in English. Based on this sample, we can observe that after fine-tuning the model in English-only prompt data set, it has become capable of providing relevant answers, but in many cases, it responded in the English language.

Instruction:
Listázd ki a személyneveket az alábbi szövegből!
(List the personal names from the text below.)

Input:
Dr. Kovács Péternek van egy autója és szereti Győzikét!
(Dr. Péter Kovács has a car and likes Győzike!)

Response:
Dr. Kovács Péter, Győzike.

Instruction:
列出以下句子中的汉语人名！
(List the Chinese names in the following sentences!)

Input:
这部由瑞恩·高斯林(Ryan Gosling) 和艾玛·斯通(Emma Stone) 主演的电影获得了各大奖项的提名。
(This film, starring Ryan Gosling and Emma Stone, has been nominated for various awards.)

Response:
瑞恩·高斯林(Ryan Gosling) 、艾玛·斯通(Emma Stone)

Table 6. A NER sample for Hungarian and Chinese instructions

In Table 6, you can see an sample for solving named entity recognition (NER) task. In this example we ask our model to list the person names from the text. In the Hungarian sample, in the sentence you can see the name 'Győzikét', which is in the accusative form. The lemma of this name is 'Győzike', which was correctly extracted by our model. We also tried this task with ChatGPT, but it returned 'Győziké', an incorrect lemma, as it ignores the quite regular variation in the stem vowel. This example demonstrates the added value of using the larger amount of data for training model in a given language.

One of the biggest advantages of the transfer learning is that our model can generate programming code (see in Table 7) with Hungarian instruct, which our monolingual model cannot do, since the Hungarian corpus does not contain programming codes.

Instruction:
Generálj egy python kódot, ami összead két számot!
(Generate a python code that add two numbers.)

Response:
```
def add_two_numbers(a, b):
   return a + b
```
Table 7. Sample for code generating

Instruction:
`Fordítsd le magyar nyelvre!`
(Translate it into Hungarian language.)

Input:
`The find also grants insight into the evolution of feathers in birds.`

Response:
`A talált lelet emellett betekintést nyújt a madarak tollazatának alakulásába is.`

Instruction:
把它翻译成中文
(Translate it into Chinese.)

Input:
`The find also grants insight into the evolution of feathers in birds.`

Response:
它还揭示鸟类的羽毛是如何演变的。

Table 8. Samples for translation

Last but not least, a multilingual task like translation also requires multilingual knowledge, which our model has become capable of acquiring (see Table 8).

6 Conclusion and Future Work

The present paper described two GPT-3 models, one monolingual (PULI-GPT-3SX – Hungarian), the other trilingual (PULI-GPTrio – Hungarian, English and Chinese) each having 6.7 billion parameters. The trilingual model was further fine-tuned with the Alpaca instruction dataset to produce an instruct GPT model. The results proved that our hypothesis about cross-lingual transfer learning operating with the trilingual model was correct. Despite their size, the performance of our models were quite promising. Our instruct GPT model in zero-shot experiments outperformed ChatGPT on some Hungarian tests in the HuLU benchmark datasets.

So far, we have only applied experiments with fine-tuning our trilingual model on Stanford Alpaca prompts. We plan to proceed by translating and cleaning the Standord Alpaca prompts into Hungarian. Using the translated prompts, we plan to fine-tune our monolingual model to be able to follow Hungarian instructions with high quality. In addition, we will fine-tune our model for chatting. Last but not least, we will improve the performance of our models by applying reinforcement learning methods.

References

1. Benko, V.: Aranea: Yet Another Family of (Comparable) Web Corpora. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) Text, Speech and Dialogue. pp. 247–256. Springer International Publishing, Cham (2014)
2. Benko, V.: Compatible Sketch Grammars for Comparable Corpora. In: Abel, A., Vettori, C., Ralli, N. (eds.) Proceedings of the 16th EURALEX International Congress. pp. 417–430. EURAC research, Bolzano, Italy (jul 2014)
3. Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U.S., Purohit, S., Reynolds, L., Tow, J., Wang, B., Weinbach, S.: GPT-NeoX-20B: An Open-Source Autoregressive Language Model. In: Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models (2022), https://arxiv.org/abs/2204.06745
4. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. pp. 3121–3124 (2010). https://doi.org/10.1109/ICPR.2010.764
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
6. Chenghao Fan, Z.L., Tian, J.: Chinese-vicuna: A chinese instruction-following llama-based model (2023), https://github.com/Facico/Chinese-Vicuna
7. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: PaLM: Scaling Language Modeling with Pathways (2022)
8. Feldmann, Á., Hajdu, R., Indig, B., Sass, B., Makrai, M., Mittelholcz, I., Halász, D., Yang, Z.Gy., Váradi, T.: HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 29–36. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2021)
9. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., Leahy, C.: The Pile: An 800GB Dataset of Diverse Text for Language Modeling (2020)

10. He, J., Qiu, J., Zeng, A., Yang, Z., Zhai, J., Tang, J.: FastMoE: A Fast Mixture-of-Expert Training System (2021)
11. Indig, B.: Közös crawlnak is egy korpusz a vége – Korpuszépítés a CommonCrawl .hu domainjából. In: Vincze, V. (ed.) XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018). p. 125–134. Szegedi Tudományegyetem Informatikai Intézet, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2018)
12. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. In: 7th International Corpus Linguistics Conference CL 2013. pp. 125–127. Lancaster (2013)
13. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2012, `https://aclanthology.org/D18-2012`
14. Ligeti-Nagy, N., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Laki, L.J., Vadász, N., Yang, Z.Gy., Váradi, T.: HuLU: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. p. 431–446. JATEPress, Szeged (2022)
15. Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 923–929. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016)
16. LM-SYS: Fastchat (vicuna: An open-source chatbot) (2023), `https://github.com/lm-sys/FastChat`
17. Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020)
18. Nemeskey, D.M.: Introducing huBERT. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 3–14. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2021)
19. Novák, A., Siklósi, B., Oravecz, Cs.: A New Integrated Open-source Morphological Analyzer for Hungarian. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
20. OpenAI: GPT-4 Technical Report (2023)
21. Oravecz, C., Váradi, T., Sass, B.: The Hungarian Gigaword corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1719–1723. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014)
22. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1719–1723. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014)
23. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022)
24. Rychlý, P.: Manatee/Bonito - A Modular Corpus Manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 65–70. Masarykova univerzita, Brno (2007)

25. Shazeer, N.: GLU Variants Improve Transformer (2020)
26. Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., Catanzaro, B.: Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism (2019)
27. Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R.Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., Catanzaro, B.: Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model (2022)
28. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: RoFormer: Enhanced Transformer with Rotary Position Embedding (2021)
29. Suchomel, V., Pomikálek, J.: Efficient Web Crawling for Large Text Corpora. In: Kilgarriff, A., Sharoff, S. (eds.) Proceedings of the seventh Web as Corpus Workshop (WAC7). pp. 39–43. Lyon (2012)
30. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca` (2023)
31. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is All you Need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)
33. Wang, B., Komatsuzaki, A.: GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax` (May 2021)
34. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language model with self generated instructions (2022)
35. Xu, B.: Nlp chinese corpus: Large scale chinese corpus for nlp (Sep 2019). https://doi.org/10.5281/zenodo.3402023, `https://doi.org/10.5281/zenodo.3402023`
36. Yang, Z.G., Dodé, R., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Kőrös, A., Laki, L.J., Ligeti-Nagy, N., Vadász, N., Váradi, T.: Jönnek a nagyok! BERT-large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. In: XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023). pp. 247–262. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Hungary (2023)
37. Yuan, S., Zhao, H., Du, Z., Ding, M., Liu, X., Cen, Y., Zou, X., Yang, Z., Tang, J.: Wudaocorpora: A super large-scale chinese corpora for pre-training language models. AI Open **2**, 65–68 (2021). https://doi.org/https://doi.org/10.1016/j.aiopen.2021.06.001