Pretraining GPT-style models in Hungarian

Kristóf Szentmihályi*, Dávid Márk Nemeskey†, András Márk Szekeres*, Bence György Gortka†, Balázs Indig†, Gábor Palkó†, Balázs Nagy‡, László Vidács‡

***Abstract*—In this paper, we introduce two bilingual large language models, named OTP-1.5B and OTP-13B, designed with a focus on both English and Hungarian languages. Both models utilize an 8k token context window and are trained on a dataset of 640 billion tokens, allowing the models to capture a broad range of linguistic nuances and generalize effectively across languages. Notably, their performance in Hungarian is on par with the results reported for equivalent models in English, such as GPT-3, marking a significant breakthrough in bilingual model development and evaluation. This achievement demonstrates the viability of scaling language models to perform robustly in less-resourced languages like Hungarian, a critical advancement for linguistic diversity in AI research. In addition, we introduce new benchmark datasets specifically designed to evaluate Hungarian language capabilities, addressing a significant gap in the current landscape of multilingual large language model (LLM) assessment. These benchmarks offer a comprehensive framework for measuring performance in non-English contexts, enabling more precise and culturally relevant evaluations of LLMs. Our models achieve state-of-the-art results in Hungarian, underscoring their proficiency and utility for a wide range of applications in both English and Hungarian language tasks.**

***Index Terms*—Article submission, IEEE, IEEEtran, journal, LATEX, paper, template, typesetting.**

I. INTRODUCTION

WITH the huge success of Transformer models [1], better and better solutions have continuously appeared, achieving state-of-the-art (SoTA) results on the traditional Natural Language Processing (NLP) tasks [2], such as language understanding and generation. Among these, the encoder only BERT model [3] (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) held its prominent first place for a long time in tasks requiring textual understanding. With the discovery of the few- and zero-shot capabilities of the GPT family of models [4], [5], the focus has shifted to Transformer decoders. Decoder-based models proved to be capable of solving these kinds of tasks even in the few-shot setting, and were also better at generating texts. The latest models (e.g., PaLM [6] or GPT-4 [7]) boast human-level performance on various benchmarks.

As most decoder-based models have been trained on English and other high-resource languages, their performance on lower-resource languages generally lag behind. In order to overcome this problem and to be able to fully use the capabilities of GPT-like generative models, we decided to develop Hungarian-specific models.

In parallel with our development, the Hungarian Linguistics Research Institute (NYTUD) also developed models and benchmark data [8]–[10]. By building a better Hungarian-language corpus, we were able to deliver stronger models, as it will be seen in the discussion and comparison of results.

The goal of our project was to create a foundational language model for public purposes that can also be used by smaller enterprises. This could allow smaller players to benefit from the AI revolution and exert a positive effect on the economy. For this reason, we paid special attention to the efficiency of the models and tried to cover all of the capabilities that someone can effectively use in production with the smallest possible model. This is in line with the generic trends: while the field first saw the development of ever larger decoder-based models (e.g., GPT3 [5], Gopher [11] or PaLM [6]), more recently small but highly capable models, such Llama as 3 8B [12] or Mistral 3 7B[1], or even phi-3-mini at 3.8B [13], have become the focus of attention. The importance of these considerations was also seen in the Mistral expert model [14] that was published recently during the development of our models.

This research is the first milestone of a greater multi-year project. Here, we present the development and evaluation phases for our Hungarian foundation models. We will report the results of other milestones of the project, such as adaptation techniques and Hungarian instruction fine-tuning, in forthcoming publications.

The rest of the paper is organized as follows. Section II details our training corpus. Section III discusses our main considerations concerning training and our experiments with tokenizers, while section IV discusses our benchmark datasets. Section V presents and discusses results. Finally, Section VI concludes the paper and delineates possible next steps.

II. CORPUS

It has long been established that pretraining Large Language Models requires huge amounts of textual data. In line with the scaling laws for foundation models [15], [16], the larger the model gets, the more massive its training corpus needs to be; Llama 3.1, the latest LLM to be released, was trained on more than 15 trillion tokens [12].

Still, quantity is only one side of the coin. Initially most general-purpose models were based on web text only [4], [5], [17], usually extracted from Common Crawl. However,

* OTP

† Department of Digital Humanities, Eötvös Loránd University, Budapest, Hungary

‡ Department of Software Engineering, University of Szeged, Szeged, Hungary

Corresponding author: D. Nemeskey (e-mail: nemeskey.david@btk.elte.hu)

[1] https://huggingface.co/mistralai/Mistral-7B-v0.3

unfiltered web datasets are known to exhibit significant quality issues [18], so a lot of effort has gone into creating clean web text corpora [4], [17], [19]. Augmenting web text with smaller but higher quality data sources such as books, news, legal and scientific texts, etc. has also been shown to have a positive effect on downstream performance [20]. As a consequence, most LLMs use a mix of cleaned web data and other, better curated data sources.

When compiling the pretraining dataset for our model, we also aimed at a large, diverse corpus. Unfortunately, smaller languages have neither the web presence, nor the amount of literary sources English does. Hungarian has only about 5B words in any OSCAR [21] release and, with the exception of mC4 [22], is missing from most large-scale multilingual collections as well, such as ROOTS [23] or Occiglot-Fineweb [24]. This prompted us to look for sources not yet tapped into by earlier Hungarian efforts [8], [25].

Our corpus consists of three main sub-corpora supplemented by a collection of miscellaneous, smaller datasets. The three main components are:

1) Web text from Common Crawl;
2) Books and papers from electronic libraries and repositories;
3) News items from online media outlets.

These sub-corpora are discussed in the following sections.

A. Web Text

We compiled our web text corpus from all Common Crawl dumps until the end of 2023. We followed the procedure outlined in [25] with a few modifications. As in the paper, all documents under the `.hu` top level domain (TLD) were downloaded. However, the distribution of Hungarian texts on the web does not correspond to the `.hu` domain. First, there are substantial Hungarian minorities in neighboring countries, chiefly in Romania and Slovakia; second, many businesses and organizations use the original TLDs, such as `.com`. Consequently, we also downloaded the `.ro` and `.sk` TLDs and added the top 1000 domains that had a large ratio of Hungarian pages according to OSCAR.

The downloaded pages were filtered for boilerplate using JusText [26]. Since JusText does not seem to catch JavaScript and cookie warning popups, these have been removed in a separate step with a few hand-crafted regular expression rules. We opted for this solution because the texts followed a relatively small number of templates and we wanted to avoid removing pages whose main content was the JavaScript language itself. The resulting documents were filtered by language on the document level. Finally, all documents shorter than 500 characters (as opposed to the 1500 in [25]) were discarded.

The data was deduplicated on both the URL and document level. As shown in Table I, out of all the cleaning steps, deduplication had the largest effect on the size of the corpus: about 67% of the index and 52% of the filtered documents were found to be redundant. Language filtering had minimal effects on the `.hu` domain, but it is responsible for the huge drop in document number between rows 2 and 3 in the table. Curiously, while we have a separate language filtering step included in row 4, most of the filtering had already been done by JusText. The difference is that our method is paragraph-based and leaves the rest of the document intact. After all filtering steps, we end up with roughly a quarter of the downloaded data.

B. Books and papers

1) Data sources: The second largest part of our corpus consists of edited documents (mostly books and academic journals) collected from publicly available electronic document repositories. Maintained by universities and research institutes, these repositories provide a so-called OAI-PMH [27] endpoint. OAI-PMH is a standard protocol for automatic metadata retrieval. Simple HTTP requests can be used to make queries, which are answered with a list of documents and their metadata. The metadata also includes where the document itself is available. This allows the documents to be downloaded automatically.

A great number of repositories providing metadata through the OAI-PMH protocol are harvested by the Scientific Document Common Search Service[2], which provides searching and browsing in the contents of Hungarian archives. We collected the endpoints and some basic information from here. The same service provides a collection of Hungarian journals qualified by the Committee for Repository Qualification (OJS/OCS Search Engine); this collection was also included in our sub-corpus.

For harvesting OAI-PMH we created our own Python script. Based on the experience we gained in the process, we have continuously improved and enhanced this code. We also saved basic data from the files for later identification.

In addition, we collected data from the Hungarian Electronic Library[3], both from its e-book collection and from the Electronic Periodicals Archive[4]. A separate script selected and downloaded the files from the library's FTP server.

2) Processing downloaded files: These services mostly store the material in PDF format, so part of our job was to process them. The PDF files include born-digital documents exported in PDF format by different text processing software, as well as documents originally published in printed format and later digitized. Most of the digitized files also contain a text layer created by optical character recognition (OCR). We used PyMuPdf [28] to process the PDF files. This allowed us to export various formats (including TXT, JSON, ALTO XML) from the documents. JSON and ALTO XML[5] files contain some useful layout information which helped in document processing.

Similarly to the previous sub-corpus, we deduplicated the documents and performed language detection, both at document and paragraph levels. We used fasttext [29] for language detection and we filtered files which contained mostly foreign language text.

[2] https://tudokk.mtak.hu/
[3] https://www.mek.oszk.hu/
[4] https://www.mek.oszk.hu/
[5] https://www.loc.gov/standards/alto/

TABLE I
EFFECTS OF THE DIFFERENT FILTERING AND DEDUPLICATING STEPS ON THE `.hu` DOMAIN IN CC.

Dataset	Documents	Characters
Index	2 146 224 445	N/A
Deduplicated index	704 743 407	N/A
Boilerplate filtering	205 370 228	929 812 129 807
Language & length filtering	124 514 464	804 876 536 633
Deduplication	59 848 839	241 572 118 452

The layout information allowed us to identify repetitive parts from texts, e.g., to filter headers and footers. As commonly used OCR procedures are not capable of interpreting complex layout information, mathematical formulas, diagrams, table of contents, tables etc. are presented as text fragments in the exported text. Thus, these had to be filtered out as well.

C. ELTE-DH Web Harvesting Corpus

The ELTE-DH Web Harvesting Corpus [30] consists of Hungarian news portals from Hungary and the neighboring countries. The corpus is created to demonstrate a well-defined crawling workflow specialized in archiving text content from all available articles of the selected news portals for digital humanities purposes. In contrast to the crawl(er)s available at that time (e.g., Common Crawl), the configuration is tailored for the actual state of the select portal to facilitate minimal resource usage by not retrieving duplicate or non-mandatory content (e.g., images, scripts, external content) if possible. Keeping these goals in mind, the operators could be certain that all available articles and nothing unnecessary is downloaded from a portal without exception. The homogeneity of the downloaded content made it possible to extract text and metadata at a precision which makes the corpus gold standard quality. The downloaded material is then published in the Zenodo.org[6] repository, which assigns a separate DOI to each dataset, and a meta-trend viewer is created to allow getting further insights into the corpus without technical skills [31].

The Web Harvesting Corpus was crawled between 2018 and 2022. It consists of more than 20 portals, with some of them spanning back to over 20 years. At about 6 million news items published from 1996 to 2020, it is the smallest of the three main sub-corpora listed in Section II; see Table II. However, it is also of comperatively higher quality than Common Crawl, so any incidental duplication between the two were resolved in favor of the Web Harvesting Corpus.

D. Miscellaneous datasets

We also included a set of small, yet high quality datasets in our training mix:

- Anonymized court rulings[7] (`Court`);
- A collection of parliamentary speeches downloaded from parlament.hu (`HuParl`);
- The Hungarian monolingual part of the OpenSubtitles[8] corpus [32] (`OpenSubtitles`);
- The Hungarian Wikipedia[9] (`Wikipedia`) taken from WebCorpus 2 [25].

E. Corpus statistics

Table II shows the final composition of our pretraining corpus. As can be seen, even heavily deduplicated, web text accounts for a little over three fourth of the corpus, with better edited sources making up the rest. Looking at the length distribution of the documents, two outliers emerge: Wikipedia documents seem to be on the short side, which points to a possible problem with text extraction. On the other end of the spectrum, the HuParl documents each contain the minutes of a parliamentary sitting (day). We found no easy way to divide these into smaller parts, so they were included as-is.

All in all, our pretraining corpus is the largest Hungarian corpus to date. Compared to earlier efforts [8], [25], it contains a larger ratio of academic, legal and literary texts. While it has fewer news items than in [8], given the inherently duplicative nature of the genre, the actual coverage should be similar.

III. METHODS

A. Selection of training data

The scaling laws of LLMs, which try to establish the amount of data needed to train models of certain sizes, have been revised multiple times during the last few years. When our project started, the consensus was along the lines of the Chinchilla study [16], which suggested a corpus of a few hundred billion tokens for pre-training a 13B model. While this number is an order of magnitude larger than the size of our entire Hungarian corpus, it has been shown previously that the strategy of training for multiple epochs by repeating data may help in data-constrained regimes [33]. Accordingly, we upsampled our Hungarian corpus 4 times and then added an equal amount of English text from the Pile [20].

More recent developments [34]–[36] have shown that LLMs could greatly benefit from yet larger training corpora and longer training than what those earlier guidelines suggested. Unfortunately only the largest companies can afford the resources required for these extended trainings.

B. Model architecture

We used the same model architecture as OpenAI did in developing the GPT3 model [5]. The goal of the first milestone was to acquire similar capabilities as OpenAI achieved, but in Hungarian language.

[6] https://zenodo.org
[7] https://eakta.birosag.hu/anonimizalt-hatarozatok
[8] http://www.opensubtitles.org/
[9] https://hu.wikipedia.org/wiki/Kezdőlap

TABLE II
SIZES AND RATIOS OF THE DIFFERENT SUB-CORPORA

Dataset	Subset	Documents	Words	Avg. document length	Percentage
Common Crawl	.hu	53 803 209	29 317 226 270	544.90	75.92%
	.ro	576 944	288 134 477	499.41	
	.sk	454 408	153 976 537	338.85	
	.com	3 029 769	1 756 970 221	579.90	
Repo	EPA	300 790	1 611 420 639	5357.29	17.07%
	Books	32 830	807 341 016	24 591.56	
	OJS	27 702	58 105 913	2097.53	
	OAI	349 398	4 612 470 677	13 201.19	
News		5 974 635	1 720 655 367	287.99	4.15%
Court		198 296	669 503 351	3376.28	2.86%
HuParl		1707	84 756 485	49 652.31	
OpenSubtitles		88 519	306 833 037	3466.30	
Wikipedia		418 621	124 982 503	298.56	
Sum		65 256 828	41 512 376 493	636.14	100%

C. Tokenizers

Since the model was designed to be multilingual based on its training data, its tokenizer had to be produced accordingly. We achieved this by sampling a smaller unit from the English and Hungarian corpora, then using the BPE [37] algorithm to create our dictionary containing 52k tokens. We paid special attention to make it optimal for both Hungarian and English. This way we can efficiently encode information for both and converge to the bilingual distribution. During the research, we also tried an approach suggested by our partner SambaNova, where we modified an English tokenizer by changing the last 4 000 vocabulary elements to Hungarian. The tokenizers produce the following fertility [38] measurements:

TABLE III
FERTILITY VALUES OF THE TOKENIZERS

Tokenizer	Hungarian	English
native for English	3.29	1.15
native for Hungarian	1.38	1.92
bilingual	1.88	1.50
modified	2.36	1.64

During the calculation, we used the RegExpTokenizer class of the nltk library [39]. We split the sample texts from https://universaldependencies.org/ into words (r"\w+"). By iterating through the list, we tokenize all the words, and then averaged the quantities associated with the words. The lower the value, the faster the training, since we can encode more information in one batch.

If we modify the calculation in a much simpler way and measure how many tokens are needed to cover a text for a given tokenizer, or in other words, how many characters it compresses on average, then the table looks like this:

TABLE IV
CHARACTER COMPRESSION CAPABILITY OF EACH TOKENIZER

Tokenizer	Hungarian	English
native for English	2.03	3.79
native for Hungarian	4.22	2.52
bilingual	4.02	3.78
modified	3.21	3.06

D. Hardware and training environment

We performed the training on a special hardware consisting of 96 SN10 RDUs [40]. We used SambaStudio [41] to perform the training. In this setup we only have control over the data; the hardware specifics and the model architecture are fixed.

IV. BENCHMARK DATASETS

The major breakthrough of the GPT3 model was its ability to generalize well on unseen tasks, as illustrated by several examples in the original publication [5]. These results were unimaginable without task-specific fine-tuning at the time. This means that to measure the capabilities of a GPT3 model, we need benchmarks to measure commonsense reasoning, grammatical excellence, translation quality and logical reasoning. Creating a large, high-quality benchmark dataset is a complex task and usually requires extensive research. For this reason, we tried to find a Hungarian equivalent for a subset of the listed measurement points, as we did not want to unnecessarily invest enormous energy in the creation of a new one. Fortunately, the Hungarian Research Centre for Linguistics already made efforts in the direction of measurement, so as a logical first step we examined these datasets [10].

A. HuLU experiences

Evaluating newly trained language models is an important and usually elaborate task designed to understand model performance, capabilities, and limitations. A common and well-accepted way of evaluation is to use benchmarks which provide a standardized framework for fair evaluation.

Benchmarks ensure that the performance of the different models or model versions is measured in a comparable way. For English models, probably the most widely used such benchmarks are the GLUE [42] and SuperGLUE [43] collections of language tasks. The tasks in these benchmarks have been selected to assess a model's ability to understand and process language. In GLUE, tasks include – among others – sentiment analysis, sentence similarity measurements, and question answering exercises. SuperGLUE, aiming at improving upon some limitations in the previous benchmark and addressing challenges introduced by the more advanced

models, includes more difficult tasks such as Common Sense reasoning or the Winograd Schema Challenge.

For the first larger Hungarian models, the practice was to evaluate using the Szeged Treebank [44], which is a large manually annotated treebank for the Hungarian language. It has multiple variations including a treebank annotated for noun phrases and clauses, a treebank that contains a deep phrase-structured syntactic analysis for all sentences, and a dependency parsing treebank. A subcorpus of the Szeged Treebank is the Named Entity Corpus for Hungarian dataset [45], which has also been used in model evaluations, for example the widely used huBERT model [46] has been evaluated in a NER task based on this dataset.

Recently, a new benchmark, the Hungarian Language Understanding Benchmark Kit, HuLU [10] has been introduced. Inspired by the GLUE and SuperGLUE benchmarks, HuLU was designed to evaluate the performance of Hungarian language models using similar tasks as in the English benchmarks. The corpora of HuLU were selected from GLUE and SuperGLUE, four of them (HuSST, HuCoPA, HuRTE, and HuWNLI) are translated from the English counterparts, while the rest (HuCOLA, HuRC, and HuCommitmentBank) are from Hungarian sources.

We used to HuLU benchmark to evaluate earlier model versions developed during this work, as well as the commonly used Hungarian encoder models, namely HuBERT and PULI-BERT-Large [47]. HuBERT's model structure is equivalent of the structure of the English BERT-Base model, while PULI-BERT-Large is a Megatron [48] BERT large model. HuBERT and the PULI-BERT models were augmented with the necessary head layers for classification or multiple choice tasks and fully fine-tuned for the task at hand. We used the corresponding Hugging Face[10] libraries for the fine-tuning experiments.

Since our main purpose in this work is to develop a larger language model based on the GPT model architecture, we also investigated the HuLU tasks with a selection of such models available for the Hungarian language. In these training experiments, we fine-tuned only the top layer of the models, and kept the language model weights frozen. Prompt-based zero/one/few-shot learning was also applied, where samples were selected randomly from the training split of the corresponding dataset.

As one can see in Tables V and VI, there are a few discrepancies when comparing the accuracies obtained with the different models. First, for the translated corpora – HuSST and HuCoPA – the Hungarian BERT-like models significantly underperform the English BERT models that are fine-tuned for the corresponding English corpora, SST and CoPA. This behavior could be caused by several factors, such as the language difference, the particularities in task definition and the underlying machine learning task, and the differences in the number of examples in the training and evaluation datasets. Furthermore, Table V shows the scaling of accuracies with increasing model sizes. In all Hungarian cases, the BERT-Base variant outperforms the large variant, which could be caused by the limited size of training samples in the benchmark sets.

Inconsistencies also occur in Table VI, when comparing performances for larger models, both with traditional fine-tuning and low-shot learning. In case of foundation models, the expectation is that few-shot outperforms one- and zero-shot results. However, this tendency cannot be seen undoubtedly for the HuLU benchmark tasks. For example, HuSST performs especially poorly with all models compared to fine-tuning experiments. Moreover, with the exception of the two examined OTP models, the 1.5B and 13B variants, all benchmarks performed worse in few-shot experiments than in the one-shot ones or even the zero-shot ones.

Apparently, in many cases, low-shot results are around the random guess baseline. Other experiments show that an exclusive English model yields very similar accuracies to those trained on Hungarian language. Both in zero-shot and one-shot settings, GPT2-XL gives more accurate results for HuSST than the Hungarian models. For HuCOLA, the fine-tuned English model also gives very similar results to Hungarian models. Considering HuSST, the best results for zero-shot and one-shot learning are obtained with the English model, while in the few-shot setting the tendency is completely different, where the best result is obtained with the largest model. There is also a discrepancy in the PULI models. In the one-shot setting of both HuSST and HuCOLA the 350 million parameter PULI-GPT-2 outperforms the much larger, 6.7 billion parameter size PULI-GPT-3 model. After examining the model outputs in detail, it can also be seen that the distribution of correct answers is not similar either for the two PULI models.

In conclusion, based on the above results, the HuLU benchmarks show quite a few uncertainties both in model size scaling experiments, and also in low-shot learning experiments. Since our aim is to evaluate large foundation models as efficiently and accurately as possible, these uncertainties are clearly unacceptable and make the development of a new custom Hungarian benchmark necessary.

B. Handcrafted datasets

To help us compare various models of roughly similar sizes, we required Hungarian language benchmarks which were sensitive enough to show the fine differences between the models. Therefore, we created small but balanced datasets. Since we had a huge Hungarian collection of data, we decided to start from that, instead of translating English benchmarks. We wanted to create tasks that could be easily solved by native Hungarian speakers but are challenging enough for moderate sized models. Consequently, we had to rely on human creativity during the creation of these datasets. We created three such datasets: sentiment, interpretation and integrity. Since the goal was only the evaluation measurement and ranking without proper model training, we created 200 evenly distributed examples for each type of datasets.

Our first benchmark set is a sentiment analysis dataset, which was intended to be a trivial task for any Hungarian-speaking adult, so it can serve as an entry level benchmark. We filtered our Hungarian corpus for forum posts and extracted

[10] https://huggingface.co/

TABLE V
REFERENCE FINE-TUNING EXPERIMENTS FOR A SELECTION OF HULU TASKS WITH THE HUBERT AND PULI-BERT-LARGE MODELS. ALL VALUES ARE ACCURACIES.

Model	HuSST		HuCoPA		HuRTE		HuCOLA	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
HuBERT	**0.743**	**0.657**	**0.640**	**0.639**	**0.757**	**0.747**	**0.891**	**0.821**
PULI-BERT-Large	0.724	0.598	0.570	0.570	–	–	0.849	0.745
BERT-Base[a]	0.935	–	**0.740**	–	0.664	–	0.521	–
BERT-Large[a]	**0.949**	–	0.706	–	**0.701**	–	**0.605**	–

[a] Results obtained for the corresponding English benchmark with the English BERT model variants base and large.

TABLE VI
EVALUATING SEVERAL LARGER GPT-LIKE MODELS USING THE HULU BENCHMARKS. ALL VALUES ARE ACCURACIES.

Model	HuSST	HuCOLA	HuWNLI	HuRTE
LM head only fine-tuning				
PULI-GPT-2 [47]	60.17	78.46	46.66	55.96
PULI-GPT-3 [47]	**65.06**	78.90	36.66	**62.96**
GPTrio [9]	64.97	**79.23**	50.00	62.55
OTP 1.5B	31.33	77.80	**51.66**	48.55
OTP 13B adapted	58.28	77.58	48.33	56.79
GPT2-XL[4]	19.39	78.35	30.00	51.85
Zero-shot				
PULI-GPT-2	3.69	71.76	43.33	55.14
PULI-GPT-3	19.48	21.54	43.33	**55.97**
GPTrio	10.30	31.21	**46.67**	55.14
OTP 1.5B	2.83	**74.51**	**46.67**	55.56
OTP 13B adapted	3.44	25.60	44.44	55.00
GPT2-XL	**40.17**	21.54	**46.67**	55.56
One-shot				
PULI-GPT-2	20.06	69.23	40.00	51.85
PULI-GPT-3	3.43	64.84	43.33	**57.20**
GPTrio	31.93	69.01	40.00	47.74
OTP 1.5B	33.30	67.25	43.33	52.67
OTP 13B adapted	20.17	**69.45**	**58.33**	47.74
GPT2-XL	**33.39**	69.23	48.33	46.91
Few-shot				
PULI-GPT-2	12.10	30.77	45.00	54.32
PULI-GPT-3	19.57	43.85	45.00	**55.97**
GPTrio	29.36	65.38	46.67	50.62
OTP 1.5B	37.77	75.71	45.00	54.73
OTP 13B adapted	**43.78**	**76.70**	**50.00**	45.27
GPT2-XL	9.79	46.91	48.33	46.91

comments. We pre-categorized the comments by using a list of words for positive and negative sentiments. Naturally, the sole appearance of positive or negative words is not enough to properly categorize posts, so the result of the pre-categorization was thoroughly reviewed by human annotators.

The second benchmark was a more complex task of categorizing a text passage into a pre-determined list of topics (public life, sport, lifestyle, science, economy). The aim here was to measure the ability of the model to recognize the general meaning of texts. We created this dataset from the news portal section of our corpus, and interpreted the tags attached to news articles as the categories of this task. The results were also overseen and validated by humans.

Our third benchmark was inspired by the HellaSwag [49] dataset, which measures the model's ability to find a suitable continuation for a given text. Compared to HellaSwag, our benchmark dataset is a much more difficult task, as it contains longer text than HellaSwag for both prompts and answers, and logical reasoning is required to solve its tasks. We created this dataset from so-called choose-your-own-adventure books. The essence of this kind of literature is that it is divided into several short numbered chapters which do not progress linearly and the reader's decisions determine the development of the story. In other words, the chapters of the book are basically the nodes of a directed graph. Since we are aware of the possible directions from one node to another, we were able to automatically generate the benchmark dataset. However, we found chapters which were too short to use or not clear enough to categorize even for native speakers, so we filtered those out.

Based on our experiences, the production and expansion of high-quality benchmark data is a time-consuming process and requires a lot of attention, as well as human validation.

A representative and meaningful evaluation can only be created with high-quality data. We were able to evaluate and rank our models with these datasets. However, the last task was too difficult for our current models, so we also set a future goal for improvement that we would like to achieve with further model developments.

C. Translated datasets

With the datasets presented above we were able to analyze the Hungarian language capabilities of the models, but the goal of the project was to train a GPT3 level model. In order to evaluate this model, we had to be able to measure performances on similar tasks used for the validation of the original GPT3 model. Therefore, we translated the first 200 examples of the most widely used English datasets measuring commonsense reasoning into Hungarian in an analogous manner to previous datasets. We included the Winogrande [50], PiQA [51] and Lambada [52] datasets. We first tried to speed up the translation by automatically pre-translating the tasks using the NYTUD machine translator model [53]. Unfortunately, our experience was that Google Translate was significantly better despite the fact that its results had to be rewritten by native speakers. It is worth noting that creating a Hungarian version of Lambada was particularly challenging because of how Hungarian conjugation works.

D. LM eval harness integration

To ensure that our performance measurements are strictly deterministic and reproducible, as well as parallel to those used by the scientific community, we integrated them into the `lm-eval-harness`[11] project [54].

[11] https://github.com/EleutherAI/lm-evaluation-harness

V. RESULTS AND DISCUSSION

Our first milestone was to create a model comparable to GPT2 in size and architecture, before moving onto the realm of larger models. Such a small model allowed faster iterations with the option of testing various approaches. Since the GPT2 was trained on 40 GB of data, we thought that Webcorpus2 [25] would be sufficient for training when used in 2 epochs. Unfortunately, we could not approach the perplexity score of 8.83 even in general loss, which was the score of the OpenAI's 1.5B model in Lambada task. Data quality is critical in each phase of the trainings, and since OpenAI filtered their training data, we assumed that the poor result of our training was due to the quality of the scraped Hungarian data.

Our Hungarian data increased considerably after the arrival of Webcorpus3, however after tokenization, we saw that we are far below the guidelines [5], [16], [55], in terms of training corpus. Our goal was to create a dataset twice the size of the corpus used for training GPT3. In order to achieve this, we created a mixture where the Hungarian corpus was used for 4 epochs, and after each Hungarian part we inserted English texts from The Pile [20]. We shuffled the dataset sectionally, and finally we got a bilingual dataset with 640B tokens. This did not cause overfitting [33], but it did increase the Hungarian capabilites of the models, and it also increased the complexity of the training as we trained the model to achieve bilingualism.

With this bilingual corpus, we trained a model with 1.5 billion parameters and 1024 context window, which produced very good benchmark numbers compared to previous tests. The model is labelled as `OTP-1.5B-1k`.

We also made an experiment where we took a model previously trained on an English-only corpus (300B tokens of the C4 corpus) and did a continual pretraining to adapt it to Hungarian language. This model is labelled as `OTP-13B-2k-adapted`. Before the continual pretraining, we modified its tokenizer to have a bit better support for the Hungarian language: the last 4 000 of the original 52 000 tokens were replaced with Hungarian ones. The continual pretraining corpus contained the Webcorpus3 (84B token) and an equal amount of English text. The goal was to shorten the training time and thus the environmental impact by using an "off-the shelf" checkpoint.

We found that the English capabilities of the model decrease drastically during a pure Hungarian training. When using a bilingual corpus with the language ratio of 1:1, the Hungarian capabilities can be built up with little degradation happening to the English capabilities. When using a Hungarian to English ratio of 3:1, less compute is needed to achieve the results, at the cost of weaker capabilities at the end [56].

The results were not bad, but we were not satisfied with the improvements when compared to the 1.5B model. Also, the Byte Pair Encoding (BPE) merge rules were strongly suboptimal to the Hungarian language. So we decided to train a larger model from scratch. At the same time, we planned to extend the context window of the models to prepare them for future RAG use cases. We modified our training strategy because we divided the entire training into 2 parts [57]. One when we did the main part of the training with 500B tokens in 2k context window and another when we trained the model to be able to handle 8k context windows with larger sequences that contained 140B tokens. With the help of the strategy, we expected stable [58] training. We first validated this method by training a 1.5 billion parameter model, `OTP-1.5B-8k`. It is worth noting here that the training of the 13B model `OTP-13B-8k` took 4 months to complete.

This approach, using the same bilingual corpus, achieved better results than adapting a pre-trained English model, and also had a larger context window. It was trained on roughly 25% more tokens than the combined English-only and bilingual training of the previous model, but even more importantly, its token vocabulary was optimized from the very beginning for these languages, instead of being modified just for the continual pretraining, so as expected, it outperformed the adapted model. For this reason, we believe that much better results can be achieved in adaptation as well if we approach the training with a vocabulary that is close to optimal for each language.

Evaluation metrics obtained for the presented benchmarks are summarized in Tables VII and VIII. Total averages are presented in Table IX.

In terms of general capabilities, it can be said that bilingualism developed in parallel due to the training data; this phenomenon can also be seen in the results of the NYTUD model.

We experienced a slight difference in performance between our models and GPT3, and we assumed that it could be derived from two differences. On the one hand, our models are bilingual, and their training goal was twice as complex as that of the GPT3 model. On the other hand, the Hungarian corpus is still a small dataset compared to the English materials, moreover, its general quality may have been lower compared to the GPT3's training data. If we examine the numbers from the perspective of the quality of two target languages and focusing on the Hungarian capability, our model gives much better results compared to GPT3, and at the time of development it produced SOTA results in Hungarian. Moreover, quite notably, the 1.5 billion parameter model approached the results of NYTUD's 7 billion parameter model.

Even with a smaller corpus, better results can be achieved with a different training strategy. It is proven by the 8k extension and by the results of the Phi3 models, where they achieved the same performance with much less data than their competitors [13]. We believe that optimizing the composition and sequence of training data during pre-training or adaptation may be an important research in the future to reduce environmental impact and training costs.

Results from hand-crafted datasets, shown in Table X, suggest that we should convert them to generative measurement, as simple discriminative measurement can greatly bias the results.

Llama2 showed promising values in the results, but it seems that it is suboptimal for the Hungarian language due to its high perplexity value, and the tokenization algorithm also broke the Hungarian text into very small parts. Since it achieved very good results on the MMLU [59] benchmarks when it

TABLE VII
EVALUATION RESULTS IN HUNGARIAN

	Hungarian				
Model	Lambada (ppl)	Lambada (acc)	PiQA	Winogrande	*Average*
OTP 1.5B 1k	7.33	60.50	66.00	53.00	*59.83*
OTP 13B 2k adapted	5.42	61.00	69.00	55.50	*61.83*
OTP 1.5B 8k	6.88	61.00	62.50	55.00	*59.50*
OTP 13B 8k	**4.94**	**65.50**	**72.50**	**63.00**	***67.00***
NYTUD GPTrio 7B	6.58	59.50	70.00	55.50	*61.67*
GPT3 13B	-	-	-	-	**-**
GPT3 175B	11.63	0.00	66.50	53.50	*40.00*
Llama2 7B	35.11	45	58.5	53.5	*52.33*

TABLE VIII
EVALUATION RESULTS IN ENGLISH

	English				
Model	Lambada (ppl)	Lambada (acc)	PiQA	Winogrande	*Average*
OTP 1.5B 1k	7.60	56.20	68.93	56.67	*60.60*
OTP 13B 2k adapted	4.71	65.38	76.88	63.93	*68.73*
OTP 1.5B 8k	6.97	56.74	69.64	57.22	*61.20*
OTP 13B 8k	4.23	67.16	75.73	62.98	*68.62*
NYTUD GPTrio 7B	6.73	59.97	71.55	55.01	*62.18*
GPT3 13B	3.56	72.50	78.50	67.90	*72.97*
GPT3 175B	**3.00**	**76.2**	**80.50**	**70.20**	***75.63***
Llama2 7B	3.39	73.89	79.11	68.98	*74.00*

TABLE IX
AGGREGATED AVERAGE SCORES ACHIEVED ON THE FOUR DATASETS, PER LANGUAGE AND COMBINED

Model	Hungarian	English	*Combined*
OTP 1.5B 1k	59.83	60.60	*60.22*
OTP 13B 2k adapted	61.83	68.73	*65.28*
OTP 1.5B 8k	59.50	61.20	*60.35*
OTP 13B 8k	**67.00**	68.62	***67.81***
NYTUD GPTrio 7B	61.67	62.18	*61.92*
GPT3 13B	-	72.97	**-**
GPT3 175B	40.00	75.63	*57.82*
Llama2 7B	52.33	**76.50**	*64.42*

TABLE X
HANDCRAFTED BENCHMARKS

Model	Sentiment	Interpretation	Integrity	*Average*
OTP 1.5B 1k	97.00	**86.00**	27.5	***70.17***
OTP 13B 2k adapted	85.00	80.50	27.5	*64.33*
OTP 1.5B 8k	96.00	72.50	27.5	*65.33*
OTP 13B 8k	**98.00**	74.50	27.5	*66.67*
NYTUD GPTrio 7B	96.00	68.5	27.5	*64.00*
GPT3 13B	-	-	-	**-**
GPT3 175B	80.50	71.00	27.5	*59.67*
Llama2 7B	80.00	71.00	27.5	*59.50*

was released, considering its size, we focused our research on exploiting this capability.

VI. CONCLUSION

We created two GPT3-level bilingual models focusing on making their Hungarian language skills at least as good as their English language skills, with much less digitally available Hungarian data. Thanks to our benchmarks, now it is possible to measure the Hungarian capability of LLMs. We were able to measure that the two languages developed in parallel during the training, which resulted in our models generating the most grammatically correct Hungarian texts. This result would be unimaginable without our Hungarian corpus.

Despite the results, these models still lag behind the latest LLMs in terms of general skills and knowledge. Building a larger, more capable model would require a larger Hungarian corpus and more powerful hardware, both of which are beyond our means. Therefore, in the future we are going to focus our research capacities on model adaptation and instruction fine-tuning instead.

While we addressed the problem of the lack of benchmarks in this paper to some extent, the Hungarian LLM scene could benefit from Hungarian equivalents to some of the benchmarks used in LLM leaderboards (such as the)[12]. We leave such work for future research.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. arXiv: 1804.07461 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1804.07461.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: 1810.04805 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1810.04805.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners*, 2019.

[5] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[12] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, OpenLLMLeaderboard

[6] A. Chowdhery, S. Narang, J. Devlin, *et al*., "Palm: Scaling language modeling with pathways," *J. Mach. Learn. Res*., vol. 24, no. 1, Mar. 2024, ISSN: 1532-4435.

[7] OpenAI, J. Achiam, S. Adler, *et al., Gpt-4 technical report*, 2024. arXiv: 2303. 08774 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2303.08774.

[8] Z. G. Yang, R. Dodé, G. Ferenczi, *et al*., "Jönnek a nagyok! BERT-large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre," in *XIX. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2023, pp. 247–262.

[9] Z. G. Yang, L. J. Laki, T. Váradi, and G. Prószéky, "Mono- and multilingual gpt-3 models for hungarian," in *Text, Speech, and Dialogue*, ser. Lecture Notes in Computer Science, Plzeň, Czech Republic: Springer Nature Switzerland, 2023, pp. 94–104, ISBN: 978-3-031-40498-6.

[10] N. Ligeti-Nagy, G. Ferenczi, E. Héja, *et al*., "HuLU: Hungarian language understanding benchmark kit," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 8360–8371. [Online]. Available: https://aclanthology.org/2024.lrec-main.733.

[11] J. W. Rae, S. Borgeaud, T. Cai, *et al*., "Scaling language models: Methods, analysis & insights from training gopher," *CoRR*, vol. abs/2112.11446, 2021. [Online]. Available: https://arxiv.org/abs/2112.11446.

[12] A. Dubey, A. Jauhri, A. Pandey, *et al., The llama 3 herd of models*, 2024. arXiv: 2407.21783 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2407.21783.

[13] M. Abdin, S. A. Jacobs, A. A. Awan, *et al., Phi-3 technical report: A highly capable language model locally on your phone*, 2024. arXiv: 2404.14219 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2404.14219.

[14] A. Q. Jiang, A. Sablayrolles, A. Roux, *et al*., Mixtral of experts, 2024. arXiv: 2401.04088 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2401.04088.

[15] J. Kaplan, S. McCandlish, T. Henighan, *et al., Scaling laws for neural language models, 2020*. arXiv: 2001.08361 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2001.08361.

[16] J. Hoffmann, S. Borgeaud, A. Mensch, *et al*., *Training compute-optimal large language models*, 2022. arXiv: 2203.15556 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2203.15556.

[17] C. Raffel, N. Shazeer, A. Roberts, *et al*., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[18] T. H. Trinh and Q. V. Le, *A simple method for commonsense reasoning*, 2019. arXiv: 1806.02847 [cs.AI]. [Online]. Available: https://arxiv.org/abs/1806.02847.

[19] G. Penedo, H. Kydlíček, L. B. allal, *et al*., *The fineweb datasets: Decanting the web for the finest text data at scale*, 2024. arXiv: 2406.17557 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2406.17557.

[20] L. Gao, S. Biderman, S. Black, *et al*., *The pile: An 800gb dataset of diverse text for language modeling*, 2020. arXiv: 2101.00027 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2101.00027.

[21] J. Abadji, P. Ortiz Suarez, L. Romary, and B. Sagot, "Towards a cleaner document-oriented multilingual crawled corpus," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, *et al*., Eds., Marseille, France: European Language Resources Association, Jun. 2022, pp. 4344–4355. [Online]. Available: https://aclanthology.org/2022.lrec-1.463.

[22] L. Xue, N. Constant, A. Roberts, *et al*., "MT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, *et al*., Eds., Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. [Online]. Available: https://aclanthology.org/2021.naacl-main.41.

[23] H. Laurençon, L. Saulnier, T. Wang, *et al*., "The big-science roots corpus: A 1.6tb composite multilingual dataset," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 31 809–31 826. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/ce9e92e3de2372a4b93353eb7f3dc0bd-Paper-Datasets_and_Benchmarks.pdf.

[24] M. Ostendorff. "Announcing occiglot-fineweb." (2024), [Online]. Available: https://huggingface.co/blog/malteos/occiglot-fineweb (visited on 08/16/2024).

[25] D. M. Nemeskey, "Natural language processing methods for language modeling," Ph.D. dissertation, Eötvös Loránd University, 2020.

[26] J. Pomikálek, "Removing boilerplate and duplicate content from web corpora," Ph.D. dissertation, Faculty of informatics, Masaryk university, Brno, Czech Republic, 2011.

[27] Open Archives Initiative, *Open Archives Initiative Protocol for Metadata Harvesting*, version 2.0, https://www.openarchives.org/pmh/, 2002.

[28] Artifex Software, Inc, PyMuPDF, version 1.21.1, https://pymupdf.readthedocs.io/, 2023.

[29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, L. Lee, M. Johnson, and K. Toutanova, Eds., pp. 135–146, 2017. DOI: 10.1162/tacl_a_00051. [Online]. Available: https://aclanthology.org/Q17-1010.

[30] B. Indig, Á. Knap, Z. Sárközi-Lindner, M. Timári, and G. Palkó, "The ELTE.DH pilot corpus – creating a handcrafted Gigaword web corpus with metadata," English, in *Proceedings of the 12th Web as Corpus Workshop*, A. Barbaresi, F. Bildhauer, R. Schäfer, and E. Stemle, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 33–41, ISBN: 979-10-95546-68-9. [Online]. Available: https://aclanthology.org/2020.wac-1.5.

[31] B. Indig, Z. Sárközi-Lindner, and M. Nagy, "Use the metadata, luke! – an experimental joint metadata search and n-gram trend viewer for personal web archives," in *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, M. Hämäläinen, K. Alnajjar, N. Partanen, and J. Rueter, Eds., Taipei, Taiwan: Association for Computational Linguistics, Nov. 2022, pp. 47–52. [Online]. Available: https://aclanthology.org/2022.nlp4dh-1.7.

[32] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, N. Calzolari, K. Choukri, T. Declerck, *et al*., Eds., Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 923–929. [Online]. Available: https://aclanthology.org/L16-1147.

[33] N. Muennighoff, A. M. Rush, B. Barak, *et al*., *Scaling data-constrained language models*, 2023. arXiv: 2305. 16264 [cs.CL]. [Online]. Available: https://arxiv.org/ abs/2305.16264.

[34] H. Touvron, L. Martin, K. Stone, *et al., Llama 2: Open foundation and fine-tuned chat models*, 2023. arXiv: 2307.09288 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2307.09288.

[35] J. Bai, S. Bai, Y. Chu, et al., *Qwen technical report*, 2023. arXiv: 2309.16609 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2309.16609.

[36] G. Team, T. Mesnard, C. Hardin, *et al., Gemma: Open models based on gemini research and technology*, 2024. arXiv: 2403.08295 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2403.08295.

[37] R. Sennrich, B. Haddow, and A. Birch, *Neural machine translation of rare words with subword units*, 2016. arXiv: 1508.07909 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1508.07909.

[38] J. Ács. "Exploring bert's vocabulary." (2019), [Online]. Available: http://juditacs.github.io/2019/02/19/bert-tokenization-stats.html (visited on 08/21/2024).

[39] S. Bird and E. Loper, "NLTK: The natural language toolkit," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 214–217. [Online]. Available: https://aclanthology.org/P04-3031.

[40] H. Peng, C. Ding, T. Geng, S. Choudhury, K. Barker, and A. Li, *Evaluating emerging ai/ml accelerators: Ipu, rdu, and nvidia/amd gpus*, 2024. arXiv: 2311.04417 [cs.AR]. [Online]. Available: https://arxiv.org/abs/2311.04417.

[41] S. Systems. "Sambastudio introduction." (2024), [Online]. Available: https://docs.sambanova.ai/sambastudio/latest/sambastudio-intro.html (visited on 08/21/2024).

[42] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, T. Linzen, G. Chrupała, and A. Alishahi, Eds., Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. **DOI**: 10.18653/v1/W18-5446. [Online]. Available: https://aclanthology.org/W18-5446.

[43] A. Wang, Y. Pruksachatkun, N. Nangia, *et al*., "Superglue: A stickier benchmark for general-purpose language understanding systems," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[44] D. Csendes, J. Csirik, T. Gyimóthy, and A. Kocsor, "The szeged treebank," in *Text, Speech and Dialogue*, V. Matoušek, P. Mautner, and T. Pavelka, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 123– 131, ISBN: 978-3-540-31817-0.

[45] G. Szarvas, R. Farkas, L. Felföldi, A. Kocsor, and J. Csirik, "A highly accurate named entity corpus for Hungarian," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, N. Calzolari, K. Choukri, A. Gangemi, *et al*., Eds., Genoa, Italy: European Language Resources Association (ELRA), May 2006. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2006/pdf/365_pdf.pdf.

[46] D. M. Nemeskey, "Introducing huBERT," in *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, Szeged, 2021, pp. 3–14.

[47] Z. G. Yang, R. Dodé, G. Ferenczi, *et al*., "Jönnek a nagyok! bert-large, gpt-2 és gpt-3 nyelvmodellek magyar nyelvre," in *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*, Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Intézet, 2023, pp. 247– 262.

[48] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, *Megatron-lm: Training multi-billion parameter language models using model parallelism*, 2020. arXiv: 1909.08053 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1909.08053.

[49] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, *Hellaswag: Can a machine really finish your sentence?* 2019. arXiv: 1905.07830 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1905.07830.

[50] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, *Winogrande: An adversarial winograd schema challenge at scale*, 2019. arXiv: 1907.10641 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1907.10641.

[51] Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi, *Piqa: Reasoning about physical commonsense in natural language*, 2019. arXiv: 1911.11641 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1911.11641.

[52] D. Paperno, G. Kruszewski, A. Lazaridou, *et al*., *The lambada dataset,* Aug. 2016. **DOI**: 10.5281/zenodo.2630551.

[53] Z. G. Yang and L. J. Laki, "Solving hungarian natural language processing tasks with multilingual generative models," *Annales Mathematicae et Informaticae*, vol. 57, pp. 92–106, 2023. **DOI**: 10.33039/ami.2022.11.001.

[54] L. Gao, J. Tow, B. Abbasi, *et al*., *A framework for few-shot language model evaluation*, version v0.4.3, Jul. 2024. **DOI**: 10.5281/zenodo.12608602. [Online]. Available: https://zenodo.org/records/12608602.

[55] B. Workshop, : T. L. Scao, *et al., BLOOM: A 176b-parameter open-access multilingual language model*, 2023. arXiv: 2211.05100 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2211.05100.

[56] Z. Csaki, P. Pawakapan, U. Thakker, and Q. Xu, *Efficiently adapting pretrained language models to new languages*, 2023. arXiv: 2311.05741 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2311.05741.

[57] N. Dey, D. Soboleva, F. Al-Khateeb, *et al*., *Btlm-3b-8k: 7b parameter performance in a 3b parameter model*, 2023. arXiv: 2309.11568 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2309.11568.

[58] C. Li, M. Zhang, and Y. He, *The stability-efficiency dilemma: Investigating sequence length warmup for training gpt models*, 2022. arXiv: 2108.06084 [cs.LG]. [Online]. Available: https://arxiv.org/abs/ 2108.06084.

[59] D. Hendrycks, C. Burns, S. Basart, *et al., Measuring massive multitask language understanding*, 2021. arXiv: 2009.03300 [cs.CY]. [Online]. Available: https://arxiv.org/abs/2009.03300.