Open foundation models for Azerbaijani language

Jafar Isbarov*
The George Washington University
Department of Computer Science
0000-0001-8404-2192
jafar.isbarov@gwmail.gwu.edu

Kavsar Huseynova*
Baku Higher Oil School
Information Technology Department
0009-0007-0362-9591
kavsar.huseynova.std@bhos.edu.az

Elvin Mammadov
Baku Higher Oil School
Information Technology Department
0009-0005-9237-9736
elvin.mammadov.std@bhos.edu.az

Mammad Hajili
Microsoft
0000-0002-9522-2137
mammadhajili@microsoft.com

Abstract

The emergence of multilingual large language models has enabled the development of language understanding and generation systems in Azerbaijani. However, most of the production-grade systems rely on cloud solutions, such as GPT-4. While there have been several attempts to develop open foundation models for Azerbaijani, these works have not found their way into common use due to a lack of systemic benchmarking. This paper encompasses several lines of work that promote open-source foundation models for Azerbaijani. We introduce (1) a large text corpus for Azerbaijani, (2) a family of encoder-only language models trained on this dataset, (3) labeled datasets for evaluating these models, and (4) extensive evaluation that covers all major open-source models with Azerbaijani support.

1 Introduction

Large language models (LLMs) have seen a sudden rise in popularity in recent years. Both open-source and proprietary models have seen wide adoption across various industries. This boost has not been shared equally across different regions, however, mostly due to the slow osmosis of these technologies into low-resource languages. Azerbaijani language falls on the "other" side of this barrier, with its 24 million speakers worldwide.

While some models have a limited understanding of the Azerbaijani language, only paid models offered by OpenAI have seen some level of adoption in the industry. Open-source models are being created with multilingual or Azerbaijani-only capabilities, but the community is not as keen to adopt them. This is possibly due to the limited exploration of these models' potential. This paper encompassed several lines of work that share a common goal - promoting open-source foundational models for Azerbaijani. Our contributions are as follows:

1. DOLLMA: A new text corpus of 651.1 million words in Azerbaijani that can be used for pre-training LLMs.

2. aLLMA: A new family of BERT-class models trained on this dataset from scratch.

3. Three labeled datasets that can be used for benchmarking foundation models in Azerbaijani:

   3.1. AZE-SCI: A text classification dataset.

   3.2. AZE-NSP: A next-sentence prediction dataset.

   3.3. CB-MCQ: A closed-book question-answering dataset.

4. A benchmark for several natural language understanding (NLU) tasks in Azerbaijani. It contains our newly introduced models and other existing open-source alternatives.

[*] Equal contribution

Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024), pages 18–28
August 15, 2024 ©2024 Association for Computational Linguistics

1.1 Foundation Models

While language modeling has a long history, transformer-based large foundation models can be considered a recent phenomenon. These models have a disproportionately high number of trainable parameters, made possible due to the highly parallelizable nature of the transformer architecture. Their development takes place in two stages: Pre-training and fine-tuning. Pre-training is performed on Web-scale text corpora, while fine-tuning is performed on smaller and higher-quality data to adapt the model to a specific task. (Minaee et al., 2024)

Foundation models exist for various modalities, including language, vision, and speech. Language foundation models are usually classified as encoder, decoder, or encoder-decoder models. Encoder models are used for tasks that require language understanding, such as sentiment analysis and extractive question-answering. Encoder-decoder and decoder-only models are better suited for generative tasks, such as machine translation and text summarisation. *Our work concentrates on encoder-only models.* Our main inspiration is the BERT model family by (Devlin et al., 2019) and its derivatives.

In the rest of the paper, a foundation model refers to a language model trained on a vast amount of unlabeled text data that can be fine-tuned for various downstream tasks. A large language model refers to a foundation language model with at least tens of millions of parameters.

1.2 Modeling Azerbaijani

The majority of LLMs are either monolingual English models or multilingual models that do not support Azerbaijani. Very few multilingual models support Azerbaijani, and only recently monolingual Azerbaijani models are beginning to emerge.

This slow progress can be explained by several factors. A smaller market and less investment is an obvious explanation, but the field faces more fundamental challenges that would not be immediately solved by more funding. One of these is the state of digitalization of the language. Most of the electronic books in Azerbaijani are scanned books. Only books published since the 1990s are written in the last version of the Azerbaijani Latin alphabet [1], which creates another barrier. Yet another challenge is the small size of the community that's devoted to the development of open-source language models for Azerbaijani. The challenges regarding digitalization and script differences are further discussed in the third section.

An idea that is often heard regarding Azerbaijani LLMs is that we can simply go for the models developed for Turkish since languages are so similar. Azerbaijani and Turkish languages are not as similar as it is publicly perceived. According to (Salehi and Neysani, 2017), Azerbaijanis scored 56% of receptive intelligibility in spoken Turkish. Differences in written language are not any smaller. Based on the methodology offered by (Gupta et al., 2019), a 44% similarity score has been calculated between the vocabularies of the two languages [2]. Due to these significant differences, Turkish LLMs are not useful in machine learning tasks for Azerbaijani.

The paper is structured as follows. The next section gives a brief overview of previous works on foundational language models, and language modeling on Azerbaijani. The third section introduces DOLLMA, a new text corpus, and outlines the methodology, challenges we faced, and future works. The fourth section introduces aLLMA, a new family of monolingual encoder-only language models. The fifth section introduces several benchmarks for evaluating encoder-only Azerbaijani language models. These benchmarks are used to evaluate newly introduced models, as well as existing alternatives. The sixth section presents these benchmarks' results.

2 Previous works

The use of neural networks for language modeling can be traced back to the early 2000s. (Bengio et al., 2000) and (Mikolov et al., 2010) had created neural networks that outperformed traditional state-of-the-art model. (Schwenk et al., 2006) uses neural networks for machine translation.

These models and their derivatives were task-specific. The idea of creating a foundational language model that could later be adapted (i.e., fine-tuned) to specific tasks was popularized only after the introduction of the transformer architecture by

[1] There was an older version of the Azerbaijani Latin alphabet introduced by the Soviets in 1922. This followed several variations until 1939 when the alphabet was replaced with a Cyrillic alternative. Azerbaijan started the transition to an updated Latin alphabet in 1991, which was completed in 2001.

[2] https://www.ezglot.com/most-similar-languages?l=aze

(Vaswani et al., 2017). The earliest foundational language model that gained wide adoption was BERT by (Devlin et al., 2019) and later variations like RoBERTa (Liu et al., 2019).

BERT was an encoder-only model, therefore more suitable for problems that could be formulated as a subset of the classification problem. Generative foundation models came out around the same time, in the example of GPT-1 (Radford and Narasimhan, 2018), GPT-2 (Radford et al., 2019), and T5 (Raffel et al., 2019). While the GPT series continued with closed-source, enterprise models, other alternatives quickly emerged with superior performance. The most famous of these was the LLaMA series, which directly or indirectly resulted in the development of hundreds of open-source language models. (Touvron et al., 2023).

Early foundation models were trained on English text, but multilingual models quickly emerged. Google had released multilingual BERT alternatives, and mGPT by (Shliazhko et al., 2023) was an early variation of the GPT architecture for multiple languages. XLM-RoBERTa by (Conneau et al., 2020) was a larger and more successful alternative to mGPT and was quickly adopted worldwide.

XLM-RoBERTa was also one of the first (if not the first) foundation models that supported Azerbaijani. We are aware of only one academic work that has concentrated on the development of foundational language models for Azerbaijani. (Ziyaden et al., 2024) have trained a RoBERTa model on the Azerbaijani split of the OSCAR dataset (Ortiz Suárez et al., 2020). This work is a first of its kind for Azerbaijani and a very valuable starting point. However, it does not concentrate on the development of a foundation model. Its main focus is improving model performance by text augmentation. Therefore, they do not perform a systematic evaluation of the model. They have released one RoBERTa model, without different sizes, which is yet another limiting factor in the adoption of the work. Unfortunately, this model has not been included in our evaluation benchmarks because they have not released a tokenizer that is compatible with their model.

There have also been some community attempts to create such open-source models. A series of RoBERTa models were developed by continuing the pre-training phase on a small Azerbaijani dataset (Hajili, 2024c). Alas Development Center has developed a series of decoder-only LLMs for Azerbaijani [3], but they offer no explanation regarding their approach, and the models failed to pass initial sanity checks.

3 Text corpus

A large text corpus is a prerequisite for training a large language model. For reference, GPT-2 and RoBERTa both were trained on OpenWebText (Liu et al., 2019), consisting of 13.5 billion tokens, which is roughly equivalent to 10 billion words. Original BERT models were trained on 3.3. billion words. While these numbers have exploded in recent years, the success of these models suggests that similarly effective models can be trained on similarly sized datasets.

The largest corpora that existed at the beginning of our work were OSCAR, which contained 316 million words in Azerbaijani, and Colossal Clean Crawled Corpus (C4) with 1.7 billion words. Introduced by (Raffel et al., 2020), C4 is one of the most widely used datasets in the pretraining stage of LLMs. C4 is labeled by language and contains 1.83 million documents tagged as Azerbaijani. Upon further inspection, however, we discovered a significant portion of this text is not only in different languages, but also in different alphabets (Armenian, Georgian, and Cyrillic). In addition, the C4 dataset contains a significant amount of informal text. This can be a valuable resource, but it is outside the scope of our work. Considering all of these points, we decided against using it. OSCAR (Ortiz Suárez et al., 2020) dataset is also derived from CommonCrawl. It suffers from the same problems, so it was not included in our corpus either.

Due to these limitations, we decided to curate a new dataset specifically for pre-training LLMs that understand Azerbaijani. This new corpus is called DOLLMA (**D**ataset for **O**pen **L**arge **L**anguage **M**odels in **A**zerbaijani).[4] The first and current version of this dataset contains Azerbaijani Wikipedia, Translated English Wikipedia (incomplete), news, blogs, books, and Azerbaijani laws. This dataset contains about 651.1 million words.[5] New versions

[3] `https://github.com/interneuron-ai/project-barbarossa`

[4] `https://huggingface.co/datasets/allmalab/DOLLMA`

[5] Words were counted with a simple whitespace tokenizer.

Data source	Word count	Upscale	Final count	Source
English Wikipedia	194.0M	4	776.0M	(BHOS AI R&D Center, 2024)
Azerbaijani Wikipedia	40.0M	6	245.0M	(aLLMA Lab, 2024c)
News	238.9M	1	238.9M	BHOS AI R&D Center
Books I	2.5M	20	50.0M	aLLMA Lab
Books II	131.7M	4	526.8M	LocalDoc
Blogs	0.9M	20	17.5M	aLLMA Lab
Azerbaijani laws	44M	6	264M	(aLLMA Lab, 2024e)
Total	651.1M	-	2118.2M	-

Table 1: Data sources used to generate the DOLLMA corpus. English Wikipedia has been translated with open-source models by the BHOS AI team.

of DOLLMA will incorporate the Common Crawl data.

Books. We attempted to create a large book corpus but faced several challenges. Most of the available electronic books in Azerbaijani are scanned copies. Publishers rarely offer electronic books that are suitable for text extraction. As of 9 May 2024, Qanun Publishing, the largest publishing house in Azerbaijan, offers 52 PDFs or EPUBs on its website. The remaining books, which were sampled from the Azerbaijan National Library [6], Children's Library [7], and other sources, are all scanned copies that have occasionally passed through an OCR model. For OCR, Tesseract (Smith, 2007) was chosen due to its multilingual support and open-source availability. We scanned thousands of books and manually sampled and analyzed them. Tesseract failed to capture guillemets, which is widespread in older Azerbaijani books. It also mixed up "m" with "rn" in scanned books. This happened often enough to decrease the quality of the text substantially. Due to these limitations, we decided against using OCR output altogether as training data. Instead, we opted for two datasets:

1. Books I contains a small number of hand-picked books.

2. Books II contains a higher number of books with less detailed processing.

Wikipedia. We used dumps provided by the Wikimedia Foundation to create a new version of Azerbaijani Wikipedia. Both the data (aLLMA Lab, 2024d) and cleaning scripts [8] are publicly available. BHOS AI team leads another initiative where they are using open-source translation models to translate English Wikipedia into Azerbaijani (BHOS AI R&D Center, 2024). While this dataset offers little in terms of linguistic variety, it provides an invaluable knowledge base to train the models. Therefore, it was included in the final corpus.

News. There is an abundance of news datasets for Azerbaijani. However, we decided against using a very large news corpus, since it offers little variety in terms of language. In our experience, models trained on news datasets do not learn the language comprehensively, possibly because the news contains little to no creative writing, first- and second-person narration, and dialogue. Due to these limitations, only two news datasets were included. One contains text scraped from several news platforms, and the other contains news and updates from Azerbaijan National Library. The BHOS AI team provided both datasets.

Blogs. Another data source was blog posts collected from various websites. Instead of scraping a large number of websites for their blogs, several blogs were manually picked due to their high-quality text and informative content.

Laws. The last part consisted of Azerbaijani laws, all of which are publicly available. We have also released this as an independent text corpus (aLLMA Lab, 2024e).

You can see a summary of these sources and their accompanying upscaling ratios in Table 1. Upscaling ratios were decided rather arbitrarily. We decided against upscaling the news since they of-

[6] https://www.millikitabxana.az/

[7] https://www.clb.az/

[8] https://github.com/ceferisbarov/azwiki

fer little linguistic variety. Azerbaijani Wikipedia was upscaled higher than the translated English Wikipedia to account for the lossy translation process. Azerbaijani laws offer higher-quality text than Azerbaijani Wikipedia but offer less variety both in terms of content and form. Considering this, we upscaled them at the same level. Blogs and Books II datasets were hand-picked and constituted the highest-quality text in our corpus. Therefore, their upscaling ratio was the highest. Books II had mediocre quality, mostly due to the challenges of extracting text from PDF files. We upscaled it at the same level as the English Wikipedia.

A major shortcoming of DOLLMA is imbalanced domain distribution. While the dataset contains a substantial amount of text on Azerbaijani laws, it is lacking in terms of first-person narrative, and STEM fields. It is also heavily Azerbaijan-centric, which may or may not be an issue depending on the final goal.

Deduplication has not been performed since none of the sources has the potential of overlapping with another (i.e., Wikipedia and News, or Books and Laws). However, the addition of a deduplication stage is important if this corpus is to be expanded further.

Later versions of DOLLMA will include several major changes:

1. Add deduplication to the pipeline. This will allow us to incorporate potentially overlapping text sources.
2. Create a large-scale book corpus.
3. Improve domain distribution.
4. Incorporate web-scraping datasets such as OSCAR and C4.

We believe that these changes will open up new possibilities for modeling the Azerbaijani language. At the current state, however, taking into account time and hardware limitations, our dataset was sufficient to continue to the modeling stage.

4 Pre-training

Using DOLLMA, we have developed a series of foundational language models called aLLMA (**a** **L**arge **L**anguage **M**odel for **A**zerbaijani). aLLMA has been trained in three sizes: small, base, and large. Base and large correspond to the original BERT models BERT$_{\text{BASE}}$ and BERT$_{\text{LARGE}}$ (Devlin et al., 2019). Small architecture was borrowed from (Bhargava et al., 2021). Architectural details of these models can be found in Table 2. aLLMA-SMALL[9] and aLLMA-BASE[10] have been trained and are included in our benchmarks. aLLMA-LARGE will be released before September, 2024 and the benchmarks will be updated accordingly.

We recognize two alternative approaches to the problem of modeling a low-resource language:

- Continue the pertaining step of an existing multilingual foundation model.
- Pre-train a foundation model from scratch.

aLLMA models were developed with the latter approach. While the benchmarks contain several models that have been trained with the former method, no detailed analysis of the performance difference is provided. This is left as a future research area.

The pre-training task was only masked language modeling. The next sentence prediction task constitutes one of our benchmarks but is not included in the pre-training stage. Training loss of aLLMA-SMALL and aLLMA-BASE models can be found in Figure 1.

One major limitation of the original BERT paper was static masking. If tokens are masked before the training process, then even with multiple epochs, the model will always have to predict the same token. We borrow the idea of dynamic masking from (Liu et al., 2019). Instead of masking tokens before the training, tokens are masked on demand. This results in various masking patterns on the same text samples. Since our model is trained from scratch on an Azerbaijani-only dataset, using existing multilingual tokenizers offered no advantages. A WordPiece tokenizer[11] was trained on a weighted version of DOLLMA, with a vocabulary size of 64k. We have not performed a systematic evaluation to find the optimal vocabulary size. (Kaya and Tantuğ, 2024) have researched the impact of vocabulary size on the performance of Turkish language models. Since both Azerbaijani and Turkish are

[9] https://huggingface.co/allmalab/bert-small-aze

[10] https://huggingface.co/allmalab/bert-base-aze

[11] https://huggingface.co/allmalab/bert-tokenizer-aze

Model	Hidden Size	Num. Attention Heads	Num. Hidden Layers	Num. Parameters
aLLMA-SMALL	512	8	4	45.9M
aLLMA-BASE	768	12	12	135.2M
aLLMA-LARGE	1024	16	24	369.5M

Table 2: Architectural differences among the aLLMA models.

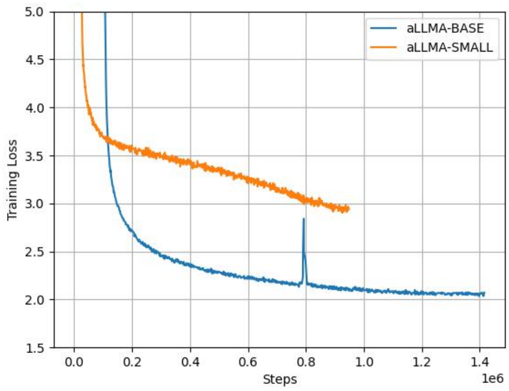

Figure 1: Training loss for aLLMA-SMALL aLLMA-BASE and aLLMA-LARGE models.

agglutinative languages and share similar morphological features, we used the results of this research as a guide. While (Kaya and Tantuğ, 2024) recommends increasing this number further, anything above that would be too computationally expensive for us.

5 Benchmarks

This section presents the tasks that were used to evaluate the natural language understanding capabilities of foundation models in Azerbaijani. All of these tasks are a form of classification since the models are encoder-only. We created three new datasets - text classification (AZE-SCI), closed-book multiple-choice questions (CB-MCQ), and next-sentence prediction (AZE-NSP) as a part of this project. Four more datasets (WikiANN, translated MRPC, translated SQuAD, and LDQuAd) were borrowed from the open-source community.

For each task, all models were trained with the same hyperparameters (learning rate, number of epochs, etc.). In almost all cases, models were undertrained - the project had hardware and time constraints and we were trying to get comparative results rather than functioning models. The source code for all experiments is being released, and the reader can generate better-performing models by simply training longer. Benchmarks have been summarized in Table 3.

5.1 AZE-SCI

AZE-SCI dataset contains titles, topics, and subtopics of dissertations written at Azerbaijani universities and institutes. Subtopics were ignored and only topic labels were used for classification. Being the simplest out of all, this dataset offers a traditional text classification challenge. (Hajili, 2024a)

5.2 AZE-NSP

The next-sentence prediction task allows us to assess the higher-level understanding capabilities of the models. We were unable to find such a dataset in Azerbaijani and decided to build one ourselves. Several books were compiled and split into paragraphs. A sentence pair was extracted from each paragraph and divided into two parts. The second sentence served as the true label, while randomly sampled sentences from other parts of the same book functioned as distractors. Special care was taken to ensure that there was no overlap between this dataset's source text and the pre-training data. (aLLMA Lab, 2024b)

5.3 CB-MCQ

The most challenging task given to the models was a closed-book multiple-choice question-answering dataset, collected from various websites. Its content is mostly middle- and high-school topics, but also contains topics like a driver's exam and state service examination. (aLLMA Lab, 2024a)

All of the tested models failed to learn this model even at a basic level. Due to this, we have decided against testing all models and including them in the leaderboards. This benchmark remains an open challenge for Azerbaijani language modeling. It has been released publicly on the Hugging Face platform to promote further research.

Dataset	Num. of samples	Task	Source
AZE-SCI	5.76k	Text classification	(Hajili, 2024a)
MRPC (translated)	3.67k	Paraphrase identification	(Eljan Mahammadli, 2024)
WikiANN	12k	Named entity recognition	(Pan et al., 2017)
SQuAD (Translated)	54.1k	Extractive QA	(Hajili, 2024d)
LDQuAd	154k	Extractive QA	(LocalDoc, 2024)
AZE-NSP	9.15k	Next sentence prediction	(aLLMA Lab, 2024b)

Table 3: Benchmarks.

5.4 Existing datasets

Several open-source datasets were sampled as an evaluation criterion. Some of these datasets were discarded due to low quality or small size. In the end, we decided on WikiANN, translated SQuAD, LDQuAd, and translated MRPC.

5.4.1 WikiANN

WikiANN is a multilingual named entity recognition dataset sampled from Wikipedia articles (Pan et al., 2017). The dataset contains 12 thousand samples in Azerbaijani. The text is tokenized and location, person, and organization entities are labeled. Since the tokenized version of the dataset does not match our tokenizer, each token was retokenized separately and a tag was assigned to each new token.

5.4.2 SQuAD

Question-answering problems usually demand more robust language understanding and therefore serve as a better criterion than simpler classification tasks. There is no original open-book question-answering dataset in Azerbaijani. The Stanford Question Answering Dataset (SQuAD) is one such dataset in English. We used a translated and reindexed version of the original (Hajili, 2024d).

5.4.3 LDQuAd

LDQuAd is a native Azerbaijani alternative to the SQuAD dataset. It contains 154,000 thousand samples, about 30% of which have no answer. Upon further inspection, we realized that most samples with a "no answer" label actually had a correct answer. It is possible that indices were generated automatically with a string search, and some answers were not found, resulting in mislabeled samples. Due to this, we discarded all samples with no answer. (LocalDoc, 2024)

5.4.4 MRPC

Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) is an English dataset that is used in NLU benchmarks like GLUE. Each sample contains two sentences and a label of whether or not two sentences are paraphrased versions of each other. We used a translated version of the corpus (Eljan Mahammadli, 2024).

6 Results

Initial tests were performed on dozens of foundation models and some were deliberately left out of the final analysis due to their inferior performance. The final benchmark includes four model categories:

Multilingual foundation models. BERT-BASE-MULTI is a multilingual version of the original BERT model. XLM-RoBERTa-BASE and XLM-RoBERTa-LARGE are some of the best-performing multilingual models (Conneau et al., 2020). mDeBERTa-v3-BASE is a multilingual version of DeBERTa v3 model (He et al., 2023)).

Multilingual models further pre-trained for Azerbaijani. BERT-BASE-AZE (Hajili, 2024b) and RoBERTa-BASE-AZE (Hajili, 2024c) have been further pre-trained on a small and high-quality Azerbaijani dataset. Their base models are RoBERTA-BASE, BERT-BASE-MULTI, and DeBERTa-BASE, respectively.

Models pre-trained from scratch. aLLMA-SMALL and aLLMA-BASE are the only monolingual Azerbaijani models. aLLMA-LARGE is still being trained.

Baseline models. The original English-only BERT-BASE was added as a baseline for the multilingual models. BERT-SCRATCH refers to the models trained on a specific task without pre-training weights. It functions as a baseline for all models in the benchmark.

Model name	Size	AZE-SCI	MRPC	WikiANN	SQuAD	AZE-NSP	LDQuAd
XLM-RoBERTa-LARGE	560M	89.76	82.41	92.35	**75.70**	33.46	83.48
mDeBERTa-v3-BASE	279M	87.13	**83.71**	91.87	72.27	**78.84**	85.29
XLM-RoBERTa-BASE	278M	86.99	70.90	90.29	70.97	74.96	85.17
RoBERTa-BASE-AZE	278M	89.17	81.25	91.62	70.36	76.98	85.44
BERT-BASE-AZE	178M	88.80	80.12	**92.35**	69.42	74.12	64.41
BERT-BASE-MULTI	178M	86.88	79.92	91.67	68.92	72.46	83.48
BERT-SCRATCH	135M	73.31	65.36	72.95	16.11	50.73	26.60
BERT-BASE	108M	76.73	75.00	90.94	55.51	62.12	74,88
ALLMA-BASE	135M	**90.84**	79.74	91.26	71.30	75.95	**86.26**
ALLMA-SMALL	46M	88.06	71.77	90.07	59.89	70.23	80.80

Table 4: Azerbaijani NLU benchmark. All metrics are F1 score. Blue models are multilingual. Orange models are multilingual models that have been further pre-trained for Azerbaijani. Green models were trained from scratch only for Azerbaijani. Black models serve as baseline.

You can find the results in Table 4. mDeBERTa-v3-BASE and aLLMA-BASE have the best overall performance. Figure 2 compares the performance of BASE models.[12] aLLMA-BASE outperforms all other models of similar size in 4 out of 6 benchmarks. Comparing BERT-BASE-AZE with BERT-BASE-MULTI shows that further pre-training of multilingual models can result in some performance improvement, but also model collapse (compare their performance in LDQuAd benchmark). However, a more comprehensive analysis is required before we can make generalizations about the effects of continued monolingual pre-training on multilingual models.

BERT-SCRATCH performs particularly well on AZE-SCI, MRPC, and WikiANN tasks. We believe this has two explanations. The first is that these tasks can be solved partially with statistical information from the input text, while this is not possible with the other tasks. The second is that the random baseline in these tasks is relatively high, while SQuAD and LDQuAd have very low random baselines.

These results demonstrate several points regarding foundation models for low-resource languages:

1. *Pre-training from scratch on a monolingual dataset is a viable strategy for building a low-resource LLM*. aLLMA-BASE has competitive performance against larger models de-

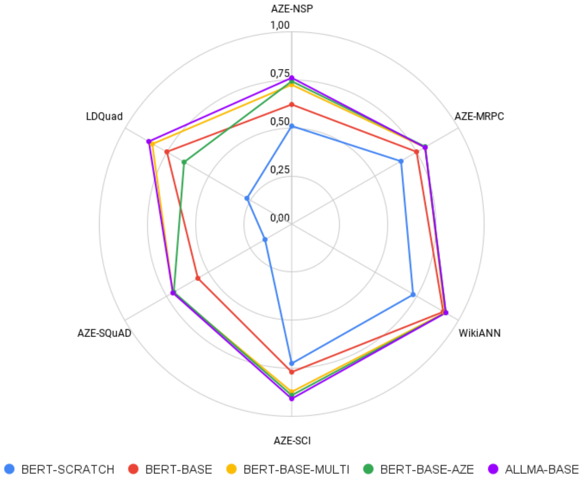

Figure 2: Performance comparison among BERT models of the same configuration. aLLMA-BASE outperforms the other models in 4 out of 6 benchmarks.

[12] The difference in number of parameters between these models is due to varying vocabulary sizes. Otherwise, their architectures are identical.

spite being trained only on the DOLLMA corpus.

2. *Multilingual models offer competitive performance even in languages that they were undertrained for.* Azerbaijani has not been the focus in any of these multilingual models (XLM-RoBERTa, mDeBERTa-v3-BASE, or BERT-BASE-MULTI). Despite this, they outperform most models in some tasks.

3. *Even monolingual English foundation models can be useful for fine-tuning on a downstream task and perform better than training a model from scratch.* BERT-BASE was included in our research as a baseline but exceeded our expectations. This suggests that the state-of-the-art English models can be utilized for certain NLU tasks in Azerbaijani. This remains a potential research area.

It is still possible that we have missed some high-quality models and we are open to feedback regarding this. Our work can be strengthened by finding or creating new benchmarks. We hope that this work will lay the foundations for such developments.

7 Conclusion

Despite some academic and community attempts to create a foundation model for Azerbaijani, this problem has not received systemic treatment. We tackle this issue by introducing a new family of foundation models for the language and benchmarking these models and other existing alternatives. To compensate for the lack of datasets suitable for benchmarking LLMs in Azerbaijani, we introduce text classification, closed-book question-answering, and next-sentence prediction datasets.

This work can be extended in several ways. The simplest improvement would be **training larger models on larger corpora**. Our project does not achieve this due to time and hardware limitations. aLLMA models are not a final product, but an early prototype. A larger training corpus, more advanced hardware, and a better-optimized training process will certainly result in more robust foundation models for Azerbaijani.

A more urgent work, however, is **extending the benchmarks** by creating more labeled task-specific datasets and adding other existing models to the leaderboards.

Including the next-sentence prediction task in the pre-training phase can increase the performance of aLLMA models further.

Another ambitious direction would be using our corpus to **develop a generative foundation model.** This paper concentrated on encoder-only models because it is a simpler problem to solve and it has more immediate applications. Nevertheless, generative language models have wide-ranging industrial applications and demand a systemic treatment.

Acknowledgements

This work has been funded by PRODATA LLC and performed at aLLMA Lab 11 . We would like to thank Haim Dror and Mirakram Aghalarov for the invaluable discussions that led to this project. We would also like to thank the BHOS AI team for their close collaboration throughout the project. All of the data and code that has been created as part of this project is going to be publicly available under permissive licenses.

References

aLLMA Lab. 2024a. az-multiple-choice-questions (revision eb9cd4f).

aLLMA Lab. 2024b. Aze-nsp (revision c59f4f8).

aLLMA Lab. 2024c. azwiki (revision 65d6610).

aLLMA Lab. 2024d. azwiki (revision 65d6610).

aLLMA Lab. 2024e. eqanun (revision 8f99a3a).

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

BHOS AI R&D Center. 2024. Translated_english_wikipedia_on_azerbaijani (revision 077a718).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Eljan Mahammadli. 2024. glue-mrpc-azerbaijani (revision b60caf0).

Prabhakar Gupta, Shaktisingh Shekhawat, and Keshav Kumar. 2019. Unsupervised quality estimation without reference corpus for subtitle machine translation using word embeddings. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 32–38.

Mammad Hajili. 2024a. azsci_topics (revision 26b9a83).

Mammad Hajili. 2024b. bert-base-cased-azerbaijani (revision 0cad0fa).

Mammad Hajili. 2024c. roberta-base-azerbaijani (revision 40f7699).

Mammad Hajili. 2024d. squad-azerbaijani-reindex-translation (revision f48f8fe).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Yiğit Bekir Kaya and A. Cüneyd Tantuğ. 2024. Effect of tokenization granularity for turkish large language models. *Intelligent Systems with Applications*, 21:200335.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

LocalDoc. 2024. Ldquad (revision e082d87).

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proc. Interspeech 2010*, pages 1045–1048.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Asgari Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *ArXiv*, abs/2402.06196.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1).

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Mohammad Salehi and Aydin Neysani. 2017. Receptive intelligibility of turkish to iranian-azerbaijani speakers. *Cogent Education*, 4(1):1326653.

Holger Schwenk, Daniel Dechelotte, and Jean-Luc Gauvain. 2006. Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 723–730, Sydney, Australia. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mgpt: Few-shot learners go multilingual.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Atabay Ziyaden, Amir Yelenov, Fuad Hajiyev, Samir Rustamov, and Alexandr Pak. 2024. Text data augmentation and pre-trained language model for enhancing text classification of low-resource languages. *PeerJ Computer Science*, 10:e1974.