Exploring the Linguistic Efficiency of Large Language Models in Armenian Discourse

Capstone Thesis

Author: Anahit Navoyan
Bachelor of Science in Data Science
American University of Armenia

Supervisor: Aram Butavyan
American University of Armenia

Abstract

This capstone project investigates the performance of a Generative Pre-training Transformer (GPT) processing the Armenian language, a low-resource language in the field of natural language processing (NLP). Although large language models (LLMs) like GPT have proven to be effective and are widely used in processing multiple languages, their performance is somewhat questionable when it comes to Armenian. This is because Armenian is a language with limited available linguistic data and has unique structural characteristics. Through a series of experiments involving different NLP tasks, such as extractive question answering, reasoning, and knowledge access, this study assesses the strengths and limitations of the GPT model. While the results suggest that GPT handles basic tasks well, the performance sharply declines when applied to deep linguistic understanding and context-based problems. Additionally, the research highlights the critical role of high-quality translations and structured prompts in improving the model's performance for a specific language. The improvements proposed here could significantly enhance the accessibility and effectiveness of GPT models for Armenian and other similar languages, making these tools more applicable in diverse digital communications such as automated customer support, content creation, and educational technologies. This research helps to ensure that lesser-spoken languages are not overlooked in the digital age by evaluating these models comprehensively to improve their accuracy and effectiveness in processing low-resource languages.

1 Introduction

In recent years, large language models (LLMs) have significantly advanced the field of natural language processing (NLP) by transforming how machines understand and generate human language. These models have achieved remarkable performance across a range of NLP tasks, setting new industry benchmarks. By analyzing extensive datasets, LLMs learn complex linguistic patterns and structures and demonstrate strong language understanding. However, their training data is primarily focused on the English language, and even other resource-rich languages receive significantly less attention compared to English, highlighting the bias in language model performance. This leads to uneven performance across languages, particularly in those that are considered low-resource languages due to limited training data. This issue has drawn considerable attention from researchers, as low-resource languages provide a unique opportunity for experimentation and exploration.

The introduction of the GPT-3 model marked a significant milestone in the field of NLP. Although its training dataset is predominantly in English, GPT-3's unique capability to learn from a few examples through textual interactions sets it apart from the other models Brown et al. [2020]. This model can generalize across multiple languages, including those with fewer resources, due to its training on a high-quality, broadly diverse dataset that has not been intentionally filtered by language Armengol-Estapé et al. [2021]. However, English still constitutes approximately 93% of the data by word count OpenAI [2023]. This huge imbalance underlines the ongoing challenges that low-resource languages face in being represented in big NLP models.

This paper concentrates on Armenian, a language with distinct linguistic complexities and scarce digital resources. It aims to investigate how effectively the GPT model processes Armenian by evaluating its performance. Investigating the effectiveness of LLMs for the Armenian language is crucial for assessing the current state of Armenian NLP and identifying strategies for improvement. This study seeks to assess how effectively Large Language Models perform for the Armenian language, with the goal of identifying areas for improvement in the field.

The findings provide significant insights into the capabilities and limitations of LLMs in handling Armenian language tasks. While some tasks are managed effectively, others need substantial improvements. For instance, the results show that tasks not heavily dependent on linguistic nuances are handled sufficiently, whereas those requiring deeper linguistic understanding present considerable challenges. The structure of the paper is organized as follows: the next section reviews related work in the field, highlighting previous studies on the multilingual capabilities of LLMs, practices applied for low-resource languages, and specific research examining ongoing efforts in Armenian NLP. Section 3 describes the methodology, detailing the task categorization, the experimental setup, and the specific metrics for evaluating model performance. Section 4 presents the data used for the study. Section 5 discusses the results, analyzing the model's performance across different tasks. Finally, the conclusion summarizes the key findings and suggests directions for future research.

2 Related Work

Numerous studies and research have been conducted to address the issue of the dominance of English-centricity in LLMs and expand their capacity to generate responses in multiple languages. Current NLP predominantly focuses on approximately 10 to 20 languages with high resources, leaving thousands of others underrepresented. Consequently, the primary concern of the researchers is to balance this gap. The problem of low-resource languages is identified as one of the biggest open problems of the contemporary NLP. Various methods and techniques already exist that try to deal with low-resource scenarios, but one major challenge involves the limited availability of labeled data for training. This limitation involves several dimensions, including the lack of task-specific labeled data, which requires manual annotation by domain experts, as well as the lack of unlabeled language-specific text data and auxiliary data sources Hedderich et al. [2020].

The Armenian language, as a low-resource language, presents wide opportunities for NLP development, research, and experimentation. Avetisyan and Broneske's examination of Armenian NLP in their 2023 paper represents both the specific gaps and progress of the language within this context. The Armenian language has distinctive characteristics, such as its alphabetic script featuring unique characters and symbols. This difference poses challenges for digital representation and the application of international standard techniques. Adapting Armenian requires accurate font development and encoding to ensure proper rendering and functionality. Moreover, as in other low-resource languages, the primary challenge remains the limited availability of large labeled datasets. This scarcity of resources limits the full utilization of LLMs for the Armenian NLP tasks. Despite technical constraints, the Armenian language also faces non-technical hardships, such as its limited number of speakers and lack of commercial investment and interest in Armenian NLP Avetisyan and Broneske [2023].

Recent studies have begun to explore the application of LLMs for the Armenian language, focusing on areas such as cross-lingual plagiarism detection and sentence alignment using BERT family models Avetisyan et al. [2023], Ter-Hovhannisyan and Avetisyan [2022]. Another study was conducted to improve the understanding of diverse languages by fine-tuning multilingual BERT (mBERT) Kulshreshtha et al. [2020]. Further research has been directed toward improving LLM efficiency for the Armenian language through various optimization techniques, including pruning, quantization, and Byte Pair Encoding (BPE). Karamyan and Karamyan [2022].

There is limited research specifically evaluating the performance of LLMs on tasks in the Armenian language. However, there is extensive literature on low-resource languages and the multilingual capabilities of LLMs. These insights could be applied to the study of the Armenian language.

In their respective studies, Zhang et al. [2023] and Armengol-Estapé et al. [2021] explore the multilingual capabilities of LLMs, particularly low-resource language conditions. Zhang et al. [2023] proposed a comprehensive framework for assessing the multilingual abilities of LLMs, focusing on cross-language generalization and categorizing tasks into Reasoning, Knowledge Access, and Articulation—each influenced by the model's linguistic capabilities. Meanwhile, Armengol-Estapé et al. [2021] explored GPT-3's capacity specifically with the Catalan language, which is similarly underrepresented in training data like Armenian. Their findings suggest that even minimal data may be sufficient for GPT-3 to demonstrate effective natural language understanding (NLU) and natural language generation (NLG) in zero-shot and few-shot scenarios, confirming its multilingual abilities despite the language's scarce presence in the training corpus.

In the further investigation of multilingualism in LLMs, Kew et al. [2023] explores the adaptability of models primarily trained in English to handle multiple languages. Their research employs instruction-tuning, applying it to various models across different linguistic scenarios. This technique involves improving the models' ability to follow complex instructions in multiple languages, which is tested through real-world tasks such as open-ended chat, extractive question answering, natural language inference, and commonsense reasoning. The findings from the study illustrate the potential of multilingual training to facilitate knowledge transfer across linguistic barriers, increasing the model's overall effectiveness. Although the effectiveness varies by task. Tasks that involve open-ended responses or extractive question answering showed improvement with multilingual instruction, whereas those requiring strict output constraints experienced less benefit. The study also suggests prompt engineering techniques for optimizing model performance. By carefully designing prompts that guide the models' generation processes.

3 Methodology

The methodology employed is aligned with the approach proposed by Zhang et al. [2023] for analyzing the multilingual capabilities of LLMs. To determine the areas in which the Armenian language has better performance, tasks were organized into three categories: natural language understanding (NLU), reasoning, and knowledge access. Although there are three categories, the study includes four specific tasks. Extractive questions answering, multiple-choice question answering, math reasoning, and knowledge-based question answering. The multiple-choice question answering and math reasoning tasks both fall under the reasoning category, as they involve answering questions based on logical and common sense reasoning. Detailed descriptions of each task are provided later in the paper.

In language technologies, tasks are typically categorized into two types, which are language-dependent and language-independent. Language-independent tasks do not require specific knowledge of a language's syntactic or semantic properties because they rely on the universal principles that apply to various languages. On the contrary, language-dependent tasks require a deep understanding of specific linguistic features, and completing these tasks requires the model's ability to accurately capture linguistic syntax, semantics, and nuances and then generate text based on these rules. According to this, the tasks were chosen based on this distinction, too.

Natural Language Understanding Tasks: These tasks are designed to assess the model's proficiency in comprehending and interpreting human language. NLU tasks are highly language-dependent and evaluate the model's ability to understand semantics, syntax, and the contextual nuances of the language. The NLU task implemented in this study is extractive question answering. This involves providing the model with a passage of the text and then giving a question related to that passage. The model is expected to generate an answer directly driven from the given passage.

Reasoning Tasks: Reasoning tasks are generally considered to be minimally influenced by languages, as they primarily require logical and rational thinking. These tasks involve solving problems based on available information and logic, using universal mathematical symbols and logical principles. Therefore, the performance of the model in these tasks tends to rely on its general reasoning ability and common sense, which are typically language-independent. In this study, the reasoning tasks selected include math reasoning and multiple choice common sense question answering, both of which test the model's ability to process and respond based on logical deductions.

Knowledge Access Tasks: Knowledge access tasks evaluate the model's ability to formulate responses from its stored knowledge base that was extracted from the training data. These tasks often involve knowledge-based question answering, named entity recognition and require the model to retrieve and apply information learned from the training. These tasks are less dependent on language specifics, as the model is expected to access and use the knowledge irrespective of the language in which the query is posed. For this part, the knowledge access task involves providing the model with open-ended general questions, to which it must formulate and deliver appropriate responses based on its pre-existing knowledge.

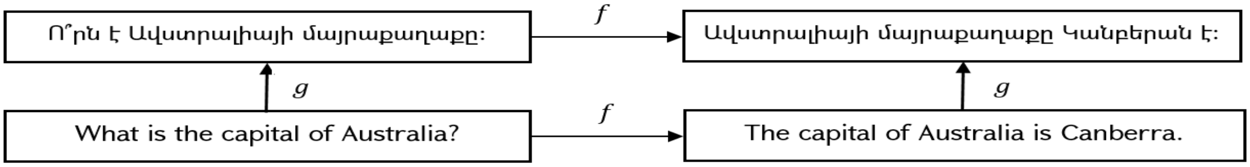

Ո՞րն է Ավստրալիայի մայրաքաղաքը:
f
Ավստրալիայի մայրաքաղաքը Կանբերան է:
g
g
f
What is the capital of Australia?
The capital of Australia is Canberra.

Figure 1: TE task visualization

Another way of categorizing the tasks is the Translation Equivariant (TE) and Translation Variant (TV) techniques Zhang et al. [2023]. TE tasks are characterized by their consistency in outcomes regardless of language translation. For instance, the model is expected to deliver the same response whether a question is posed in English or Armenian. In contrast, TV tasks are sensitive to the language of the input. The same question asked in different languages might result in different results, posing challenges for language-specific nuances. For this study, only TE tasks were used, under the assumption that the model would produce consistent and similar answers across different languages, thereby highlighting its capacity to maintain stability. The concept of TE task is visually presented in **Figure 1**.

Zhang et al. [2023] also explored interconnected translation techniques, which were Prompt Translation (PT) and Response Back Translation (RBT). PT involved using an LLM to translate non-English prompts into English, assuming that LLMs can handle TE tasks with minimal loss of information. On the other hand, RBT was obtaining output from the LLM and then translating it back to the original language. The similarity between the back-translated output of the LLM and the original source language was then evaluated for accuracy.

In this paper data translation was conducted using both Google Translate and gpt-3.5-turbo model as machine translation tools. The quality of the translation was quantified using the BLEU (Bilingual Evaluation Understudy) score, which evaluates the similarity of the machine-translated text to the reference texts. It measures the correspondence of the machine-generated text to reference translations, focusing on the precision of n-grams. Closer to one values represent more similar texts, while 0 means that the translated output has no overlap with the reference text. The BLEU score is defined as the following:

$$\text{BLEU Score} = BP \cdot \exp\left(\sum_{i=1}^{N}(w_i \cdot \ln(p_i))\right)$$

Where,

- BP stands for Brevity Penalty.
- w_i is the weight for n-gram precision of order i.
- p_i is the n-gram modified precision score of order i.
- N is the maximum n-gram order to consider.

The formula for the Brevity Penalty (BP) is given by:

$$BP = \exp\left(1 - \frac{c}{r}\right)$$

Where,

- r is the length of the candidate translation.
- c is the average length of the reference translations.

The data translation process started with converting the source text into Armenian using Google Translate. Subsequently, both the original English and the translated Armenian questions were input into the model via the OpenAI API using the gpt-3.5-turbo model. This step employed prompt engineering techniques as suggested by Kew et al. Kew et al. [2023], aiming to ensure the precision of the generated responses. Responses generated in Armenian were then back-translated into English. The text similarity between these responses in both languages was assessed by comparing them to each other and against the correct reference answers using the BERT similarity metric. This metric, based on the 'bert-large-uncased' version of the BERT model, employs cosine similarity calculations on BERT embeddings to measure the closeness of semantic content between two texts by calculating the cosine of the angle between their embeddings in a high-dimensional space Devlin et al. [2018].

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Additionally, the experimental setup incorporated various prompting strategies, including zero-shot, one-shot, and few-shot, to evaluate the model's performance across different levels of prompt engineering techniques. The system instructions varied between English and Armenian to further test the model's effectiveness under changing linguistic conditions.

The entire pipeline of the process is visually summarized in the **Figure 2** below. This illustration provides a comprehensive overview of the sequential steps and methodologies employed in the experiment.

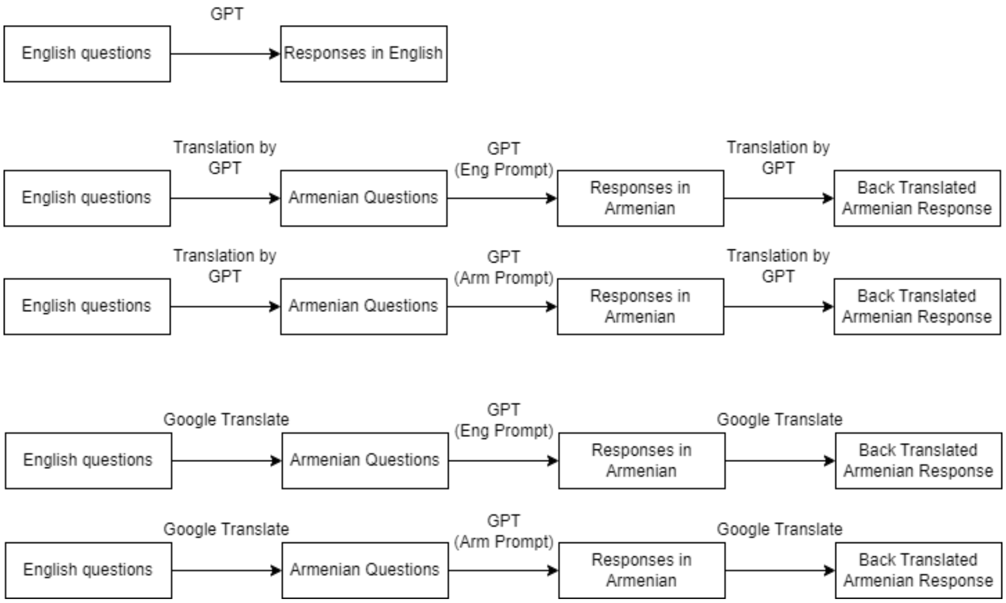


Figure 2: Pipeline

4 Data

The datasets used in this paper are accessible through the Hugging Face library, each designed to evaluate different aspects of the model's performance. The tasks include *extractive question answering*, *multiple-choice question answering*, *knowledge-based question answering*, and *math reasoning*.

- Stanford Question Answering Dataset (SQuAD) Rajpurkar et al. [2016]: A reading comprehension dataset developed by Stanford, consisting of over 100,000 question-answer pairs derived from more than 500 Wikipedia articles. Questions are designed by crowdworkers and require identifying text segments from the articles that answer the questions, including some that are deliberately unanswerable.
- Commonsense Q&A Talmor et al. [2019]: A multiple-choice dataset requiring the application of various commonsense knowledge types to select the correct answers among multiple options. It includes approximately 12,000 questions, each with one correct answer and four distractors.
- WebQuestions Berant et al. [2013]: Dataset for knowledge access. This dataset includes 6,642 question-answer pairs that rely on the model's ability to utilize knowledge bases extracted during training to generate accurate responses.
- Grade School Math 8K (GSM8K) Cobbe et al. [2021]: This dataset targets reasoning abilities with 8,500 diverse mathematical word problems suitable for grade school levels. Each problem requires logical and rational thinking based on available information and logical principles. Tasks take between 2 and 8 steps to solve, and solutions involve performing a sequence of elementary calculations and arithmetic operations.

The datasets consist of separate training and testing segments. For this paper, only a part of the training data was used, more specifically, randomly sampled 50 instances from each dataset were used to evaluate model performance. Each dataset includes answer keys, which serve as the ground truth for evaluating the accuracy.

5 Experiments and Results

5.1 Extractive Question Answering

Experiments: Extractive Question Answering is NLU task that requires the model to identify relevant answers within the given context passages provided in the prompt. To evaluate the effectiveness of gpt-3.5-turbo model in these tasks, 50 instances were randomly selected from the Armenian translation of the SQuAD dataset. Each instance consisted of a paragraph with its corresponding question-answer pair. The study experimented with different prompting setups: zero-shot, one-shot, and few-shot, all using English for the system messages, and a variation where the system instructions were provided in Armenian to evaluate the impact of instruction language on the model performance. Zero-shot approach evaluated the model's fundamental understanding and its ability to generate responses based solely on the provided input data without any prior examples. In contrast, the one-shot and few-shot approaches involve providing the model with examples prior to answering the main questions. These examples were supposed to guide the model's response structure and increase the accuracy and relevance of the question. To compare the accuracy of these responses, the cosine similarity of BERT embeddings was calculated between the generated responses.

Results: The translation of the dataset was carried out by using Google Translate and gpt-3.5-turbo model. Each instance, including the contextual passage and corresponding questions, was first translated from English into Armenian and subsequently back-translated into English. This allowed to evaluate the translation quality by comparing between the original English texts and their back-translated versions. The BLEU scores for translation done by Google Translate consistently showed higher accuracy. The average BLEU score for Google Translate was 0.61 for the context and 0.55 for the question. These suggest that Google Translate was effective in preserving the meaning and structure of the original texts during the translation process. In contrast, the GPT scored significantly lower, with average BLEU scores of 0.09 and 0.06 for the context and question. The prompt used for the translation was used from Kew et al. [2023] paper and ensured the maintenance of the structural integrity and semantic accuracy of the original texts, emphasizing the need for translation that are both grammatically correct and terminologically precise, suitable even for the expert readers in the target language.

In Figure 3, the visual representation of these BLEU scores illustrates the differences in translation quality between Google Translate (GT) and the gpt-3.5-turbo model(GPT).

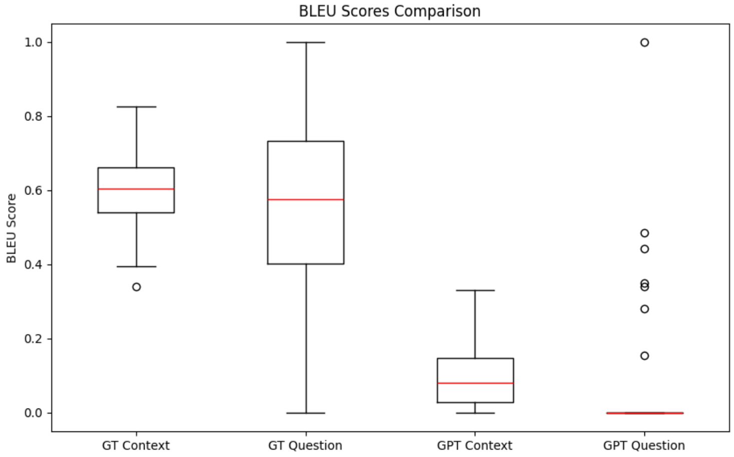
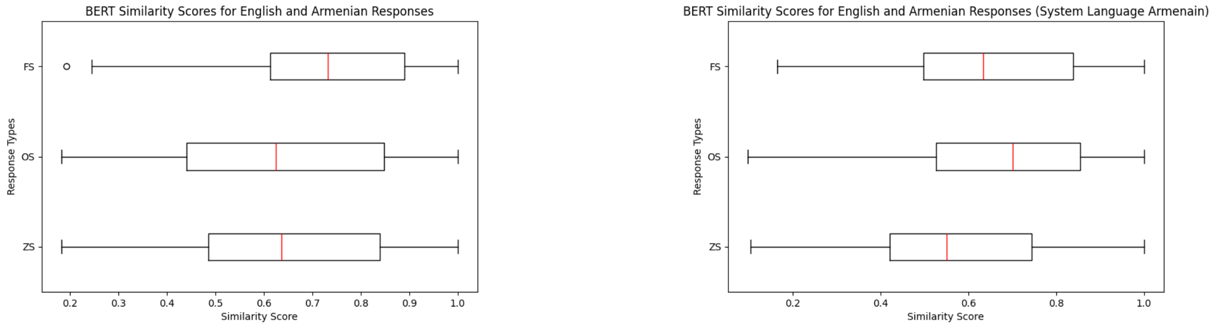


Figure 3: BLEU scores

Early tests showed that gpt-3.5-turbo's translations for Armenian were significantly less accurate compared to Google Translate. Given the higher accuracy achieved by Google Translate, it has been determined that subsequent steps that require translation will be completed using Google Translate.

For the three prompting setups, zero-shot, one-shot, and few-shot—results were analyzed under two conditions: one where the instruction language was English, and the other where it was Armenian. In both scenarios, the English text generated by the model was compared with the Armenian text that had been generated and then back-translated to English. The structure and quality of the examples provided for one-shot and few-shot scenarios played a crucial role in the outcomes. Examples were carefully prepared and presented separately from the main dataset to guide the model's response generation. When examples were given in Armenian, several challenges emerged. Complex text structures and poor-quality translations, especially those with incorrect punctuation, often affected the model's ability to generate responses. However, modifying the examples either by reducing their number or simplifying the content from a long and complex passage to a brief sentence allowed the response generation to proceed smoothly. This observation highlighted the sensitivity of the model to the quality and configuration of input data. Adjustments to the examples used in Armenian showed how small changes could significantly impact the model's performance.



(a) English Instruction Language

(b) Armenian Instruction Language

Figure 4: Comparison of model performance with different instruction languages for SQuAD

Figure 4 presents the visual comparison between two experimental scenarios. One is using English instruction, and the other is using Armenian instruction. The first results coincide with the expectations; the few-shot prompting approach demonstrates better performance compared to another setup under the condition of English instructions. Interestingly, when the instruction language is switched to Armenian, the accuracy of the one-shot prompting setup excels that observed in the few-shot and zero-shot configurations. These unexpected results show the influence of the instruction language on the model's performance. Even minimal contextual guidance can significantly improve the model's effectiveness when processing less familiar languages for the model.

All results for the extractive question answering task, including the means of the similarities between two sentences and their standard deviations (presented in brackets), are detailed in Table 1. The data indicate that the one-shot prompting strategy, where the example is provided to the model prior to the main query, is not performing well for this task. In three out of four instances, where one-shot prompting was used it resulted in lower similarity scores. The impact of the number of examples provided becomes visible when comparing model-generated responses to reference answers. Specifically, the few-shot approach slightly increases the model's accuracy for the Armenian language. What makes this experiment surprising is the fact that the score for one-shot prompting is reduced compared with zero-shot prompting, but it can be directly linked to the quality or the structure of the provided example. Consequently, the few-shot approach allows the model to understand the context better and give improved responses compared to the zero-shot. The results indicate the importance of high quality examples to influence how the model generates answers.

5.2 Multiple Choice Question Answering

Experiments: For this analysis, 50 instances were randomly selected from the Commonsense Q&A dataset, specifically designed to test multiple-choice question-answering skills based on common sense reasoning. Each instance in the dataset contains a question followed by five possible answer choices, with only one being correct. To adapt this dataset for the Armenian language assessment, both the questions and the answer choices were translated into Armenian. The instruction given to the model was taken from Kew et al. [2023] and was as follows: "You will be presented with a question and several possible answers. An example has been provided to guide you. After reviewing the example, please choose the most suitable option from 'A,' 'B,' 'C,' 'D,' or 'E' for the real question based on your best judgment." Since the answer choices contained Latin letters, they were separated from the generated responses for further analysis. The experimental setup for this task was similar to the other sections of the study, including zero-shot, one-shot, and few-shot prompting strategies with different instruction languages.

Results: The structure of examples in one-shot and few-shot instances significantly contributed to the results. However, it was also observed that in cases of giving examples to guide the model, it increased the cases of hallucinations, particularly for the Armenian language. For instance, when shown an example, the model often generated responses that exactly matched the example provided, disregarding the question. Alternatively, it sometimes repeated the question and gave it as an answer or generated text unrelated to the task, which was selecting a single letter from multiple choices.

To calculate the accuracy of the generated answers, the letters were separated from the responses because, in some instances, the choices were accompanied by statements explaining the answers or repeating the question itself. Accuracy was calculated by determining the proportion of correct responses. This involves computing the mean of true/false values, where "true" indicates a correct answer and "false" indicates an incorrect one. The results are then presented as percentages. Here again, the one-shot prompting was not efficient compared to zero and few-shots. The accuracy for English zero-shot responses measured against the answer key was 76%, while for Armenian, it was 26% with English instruction and only 4% with Armenian instruction. One interesting finding is that when instructions are given in Armenian, the accuracy significantly increases from 4% in the zero-shot setup to 6% in the one-shot and to 20% in the few-shot setup. Aside from this, for multiple choice questions answering tasks when the instruction language was English, the model performed the best.

The distribution of few-shot answers is shown in Figure 5, which illustrates the differences between English and Armenian when instructions are provided in the respective languages.

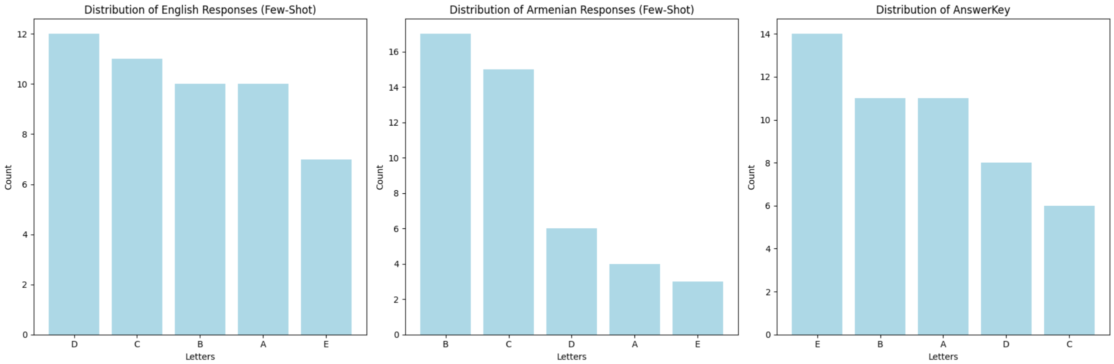


Figure 5: Distribution of Few-Shot Responses Compared to Answer Keys by Language

5.3 Knowledge-based Question Answering

Experiments: The methodology employed in the extractive question-answering task was similarly applied to the knowledge base question answering, specifically using the WebQuestions dataset. Fifty instances from this dataset were translated into Armenian and given to the get-3.5-turbo model to generate responses in zero-shot, one-shot, and few-shot setups, each with their respective instruction languages.

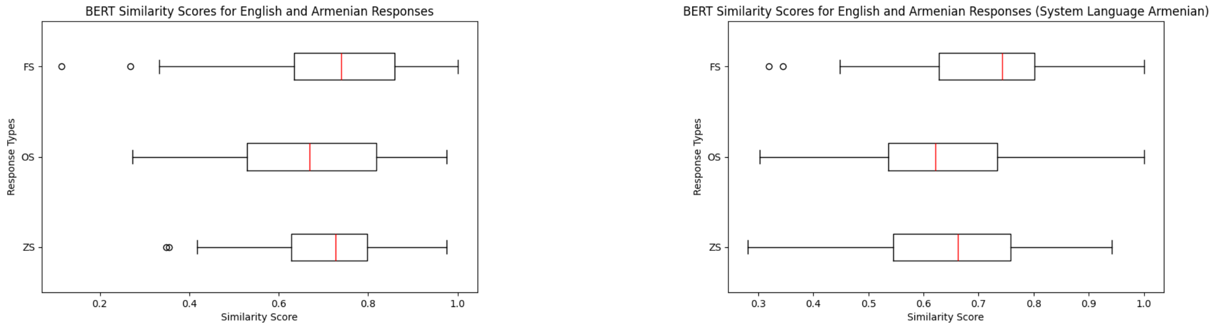
Results: As shown in Table 1, the results display the comparison of cosine similarity of BERT embeddings between generated answers text and reference texts with generated answers. From zero-shot to few-shot setups, a slight increase in similarity is observed between English-generated and Armenian-generated text, indicating improved alignment between the languages as more context is provided. Compared to the extractive question answering task, the standard deviation within these similarities is much lower, suggesting that the model produced more consistent answers across both languages. This consistency may be coming from the nature of the knowledge base question-answering task, which is less dependent on language nuances than NLU tasks. However, when comparing the generated answers to the reference text, changes in similarity are not vivid. Particularly, in the case of Armenian, the scores remain consistent across all three setups. An interesting observation is that both English and Armenian demonstrate nearly identical performance when compared with reference answers, highlighting that providing more examples does not significantly influence the consistency of outcomes in this task.

The visual representation of language comparison between Armenian and English is also shown in Figure 6.

5.4 Math Reasoning

Experiments: The experimental setup for the previously mentioned tasks is mirrored here for the math reasoning task, using the GSM8K dataset. This dataset included mathematical problems along with step-by-step solutions. Fifty instances were randomly selected from the larger dataset and translated into Armenian. Additionally, the steps required to solve each problem were isolated as an answer key, with the solutions recorded separately.

Results: The results given in Table 1 for the GSM8K dataset display a more consistent pattern across different setups. In all scenarios, one-shot prompting surpasses zero-shot prompting in performance, and few-shot prompting outperforms both, aligning with the expected outcomes. This consistency is linked to the nature of reasoning tasks, which are not language-dependent, especially for Math reasoning cases, which use universal mathematical operations and understandings that depend on the model, not on the specific language. The success in performance improvement is evident in the BERT scores, where the similarity score for the Armenian language with its reference increases significantly from 0.52 to 0.83, marking the highest improvement recorded among all tasks performed for the Armenian language. Similarly, for English, the score improves from 0.67 to 0.88. Both languages show a small variation in

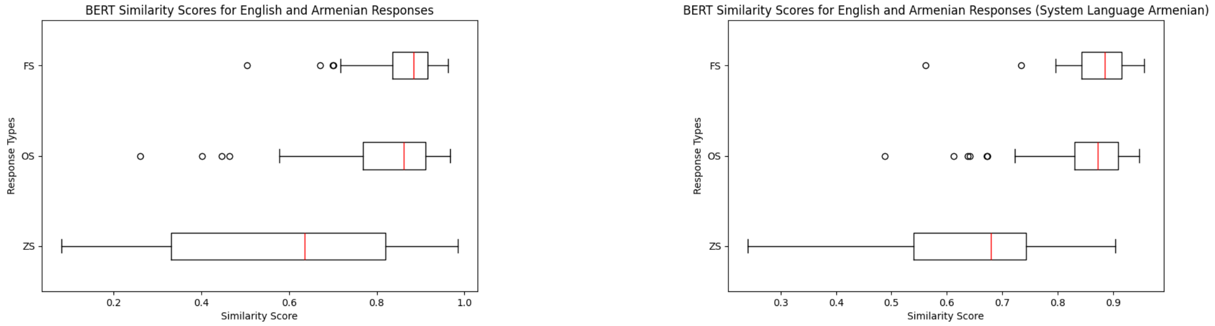


(a) English Instruction Language

(b) Armenian Instruction Language

Figure 6: Comparison of model performance with different instruction languages for WebQuestions

results, illustrating the effectiveness of few-shot prompting in improving model performance in math reasoning tasks by providing additional examples. The boxplots of these scores are represented in the Figure 7



(a) English Instruction Language

(b) Armenian Instruction Language

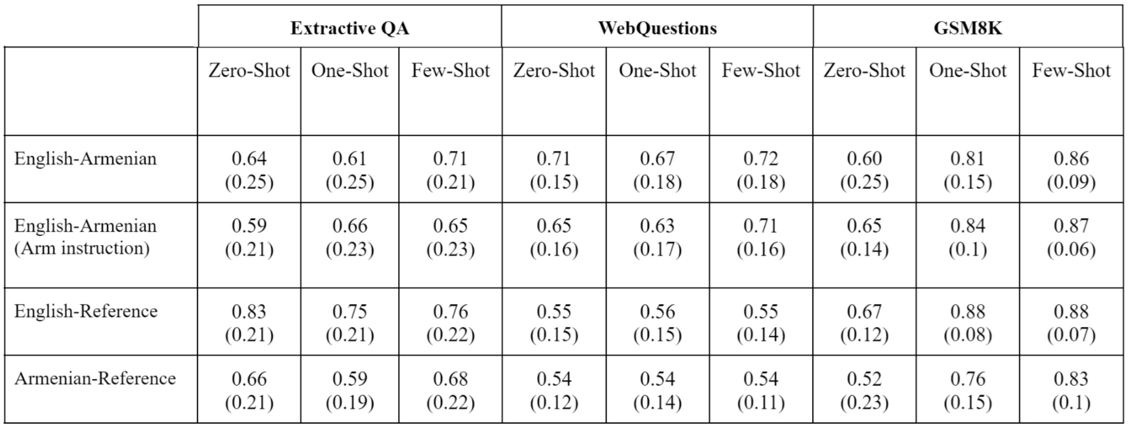
Figure 7: Comparison of model performance with different instruction languages for GSM8K

6 Discussion

The effectiveness of the one-shot and few-shot approaches heavily depends on the quality and structure of the examples provided. Properly crafted examples can lead to more accurate answers, while poorly designed ones may mislead the model, resulting in irrelevant outputs. The quantity of given examples also proved to be crucial. In certain instances, a larger number of examples overwhelmed the model, preventing it from generating any answers. This suggests that there should be a balance between providing the model with examples in order to guide it.

The other challenge was connected to translation quality. While Google Translate performed sufficiently as a machine translation tool, it also generated bad-quality translations. Issues included unnatural sentence structures, grammatical errors, and inaccuracies in translating proper names, which could mislead the model. To increase the translation quality, it would be beneficial for human translators to review and refine the machine-generated translations of the datasets in Armenian.

An important constraint during the study was the cost limitation relating to applying the GPT model and OpenAI API. These costs, therefore, constrained the experiment in terms of the volume of data processed and also constricted the scale of the experiment. However, if more resources were to be made available, then most likely, a future study with the application of such methods would use a larger dataset because it would increase the reliability and validity of the results. Applying different models from the GPT family could also provide broader insights and help investigate how model architecture affects the quality of text generated in Armenian. Besides cost limitations, the process was also time-consuming, especially when instructions were given in Armenian, which added complexity to the experimental setup.



	Extractive QA			WebQuestions			GSM8K		
	Zero-Shot	One-Shot	Few-Shot	Zero-Shot	One-Shot	Few-Shot	Zero-Shot	One-Shot	Few-Shot
English-Armenian	0.64 (0.25)	0.61 (0.25)	0.71 (0.21)	0.71 (0.15)	0.67 (0.18)	0.72 (0.18)	0.60 (0.25)	0.81 (0.15)	0.86 (0.09)
English-Armenian (Arm instruction)	0.59 (0.21)	0.66 (0.23)	0.65 (0.23)	0.65 (0.16)	0.63 (0.17)	0.71 (0.16)	0.65 (0.14)	0.84 (0.1)	0.87 (0.06)
English-Reference	0.83 (0.21)	0.75 (0.21)	0.76 (0.22)	0.55 (0.15)	0.56 (0.15)	0.55 (0.14)	0.67 (0.12)	0.88 (0.08)	0.88 (0.07)
Armenian-Reference	0.66 (0.21)	0.59 (0.19)	0.68 (0.22)	0.54 (0.12)	0.54 (0.14)	0.54 (0.11)	0.52 (0.23)	0.76 (0.15)	0.83 (0.1)

Table 1: BERT Scores: Language Comparison and Reference Evaluation (Means and Standard Deviations in Parentheses)

As a direction for future research, it would be beneficial to apply fine-tuning, particularly for tasks that demonstrated poor or unclear results. The case of the knowledge base question-answering task, where the Armenian language results remained unclear and low regardless of providing examples for guidance, suggests more targeted techniques to solve the issue. Considering the mixed outcomes observed in this task, a decent number of examples from the WebQuestions dataset were translated using machine translation and subsequently refined with human feedback to prepare for potential improvements to the language model in upcoming studies.

7 Conclusion

This study has examined the performance of LLM's, specifically the GPT-3.5 model, in processing Armenian, a low-resource language. The investigations revealed significant inconsistencies in model performance across different linguistic tasks, which were predominantly influenced by the structure and quantity of examples provided and the quality of translations.

The findings indicate that although state-of-the-art LLM like GPT-3.5 is capable of processing multiple languages to some degree, its performance significantly declines with less-resourced languages like Armenian. Tasks that do not rely heavily on language specifics tend to perform better, while the model struggles with tasks that require an in-depth understanding of the language itself. The experiments showed that few-shot learning improves results by giving the model more context, which helps it handle tasks that require a detailed understanding of language nuances. Nevertheless the picture was different in each task. One-shot learning did not meet expectations, with the majority of tasks recording the lowest scores under this setup. Specifically, the extractive question-answering task exhibited average performance, where zero-shot and few-shot approaches proved most effective. In knowledge-based question answering, the provided examples had little impact on response quality. For reasoning tasks, math reasoning showed the best performance compared to the others, while the commonsense multiple-choice answering was most efficient with a zero-shot setup, and providing additional examples sometimes led to hallucinating, complicating the task and making it confusing.

However, using high-quality translations and well-structured input examples is crucial. Translation errors, especially from tools like Google Translate, pose additional challenges and affect the model's accuracy. These issues highlight the need to improve translation methods and involve human review to ensure the accuracy of translations.

For future research, it would be beneficial to work on fine-tuning LLMs for the Armenian language. This might include using specialized datasets to improve training or creating new methods better suited to the challenges of low-resource languages. Also, finding cost-effective models that provide high accuracy without needing many resources could make NLP technology more accessible for the Armenian language.

In conclusion, this capstone project highlights the capabilities and limitations of the GPT-3.5 model in processing Armenian. It will provide a setting for future advancements in the Armenian NLP. By continuing to refine these models and tackling the identified challenges, there is a potential to improve both the accessibility and efficacy of GPT, as well as other large language models, in managing Armenian. Such continuous improvements promise to improve the performance and usability of language technologies, ensuring that Armenian and similar languages are better represented and more useful in today's digital world.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. On the multilingual capabilities of very large-scale english language models. *arXiv preprint arXiv:2108.13349*, 2021.

OpenAI. Gpt-3: Language models are few-shot learners. `https://github.com/openai/gpt-3`, 2023.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*, 2020.

Hayastan Avetisyan and David Broneske. Large language models and low-resource languages: An examination of armenian nlp. *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 199–210, 2023.

Karen Avetisyan, Arthur Malajyan, Tsolak Ghukasyan, and Arutyun Avetisyan. A simple and effective method of cross-lingual plagiarism detection. *arXiv preprint arXiv:2304.01352*, 2023.

Tatevik Ter-Hovhannisyan and Karen Avetisyan. Transformer-based multilingual language models in cross-lingual plagiarism detection. In *2022 Ivannikov Memorial Workshop (IVMEM)*, pages 72–80. IEEE, 2022.

Saurabh Kulshreshtha, José Luis Redondo-García, and Ching-Yun Chang. Cross-lingual alignment methods for multilingual bert: A comparative study. *arXiv preprint arXiv:2009.14304*, 2020.

Davit S Karamyan and Tigran S Karamyan. Compact n-gram language models for armenian. *Mathematical Problems of Computer Science*, 57:30–38, 2022.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. Don't trust gpt when your question is not in english. *arXiv preprint arXiv:2305.16339*, 2023.

Tannon Kew, Florian Schottmann, and Rico Sennrich. Turning english-centric llms into polyglots: How much multilinguality is needed? *arXiv preprint arXiv:2312.12683*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi:10.18653/v1/D16-1264. URL `https://aclanthology.org/D16-1264`.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1421. URL `https://aclanthology.org/N19-1421`.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D13-1160`.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.