Изображение выглядит как Шрифт, текст, снимок экрана, Графика

Автоматически созданное описание

**MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE REPUBLIC OF KAZAKHSTAN**

**“Kazakh-British Technical University” JSC**

**School of Information Technologies and Engineering**

**ADMITTED TO DEFENCE**

Dean of SITE,

A.Zh. Imanbayev

“ ” 2024

**Theme: «DEVELOPMENT OF METHODS AND ALGORITHMS FOR INTEGRATING COMMON SENSE INTO NATURAL LANGUAGE PROCESSING MODELS»**

**Author: Supervisor:**

Z. Aitakhunova Associate Professor, PhD

“ ” 2024 A. Pak

“ ” 2024

**Major: Norms Compliance Monitor:**

7M06105 Data Science Position, Academic degree

A. Meldebekova

“ ” 2024

**Almaty, 2024**

**ABSTRACT**

One of the main problems of research in natural language processing is that human language is often ambiguous and not fully defined. In language processing, a person relies heavily on their *common sense* and *reasoning* ability to resolve these ambiguities and recover the missing information. On the other hand, natural language processing models based on machine learning lack *common sense* and often make mistakes that might seem trivial to humans. Successful integration of common sense can help reduce such errors and improve natural language processing models. The aim of the study is to develop methods and algorithms for integrating common sense into natural language processing models. The object of research is common sense in natural language processing models. The objectives of the study include the development of algorithms for integrating common sense into natural language processing models for solving math problems, the development of a dataset of mathematical problems in UNT, the solution of which requires common sense and arithmetic reasoning, as well as the integration of common sense into natural language processing models and comparison results on the collected dataset. A detailed study of published scientific articles on the research topic was carried out. Also, an algorithm was developed for integrating common sense into natural language processing models using voting, chain-of-thought and verification. Using the developed algorithm, it was possible to increase the accuracy of solving problems from the dataset to 63.7%, which is 14% more than the baseline result of the LLM.

**The scope and structure of the dissertation:** 61 pages, 9 sections

**Number of illustrations, tables, literary sources:** 16 figures, 4 tables, 22 sources

**List of keywords:** Commonsense reasoning; Natural Language Processing (NLP); Language Models (LM); Large Language Models (LLM)

**Information about publications:** Sagingaliyev, B., Aitakhunova, Z., Shaimerdenova, A., Akhmetov, I., Pak, A., & Jaxylykova, A. (2022). A Bibliometric Review of Methods and Algorithms for Generating Corpora for Learning Vector Word Embeddings. In O. Pichardo Lagunas, J. Martínez-Miranda, & B. Martínez Seis (Eds.), *Advances in Computational Intelligence* (pp. 148–162). Cham: Springer Nature Switzerland.

**АҢДАТПА**

Табиғи тілді өңдеудегі зерттеулердің негізгі мәселелерінің бірі - адам тілі көбінесе анық емес және толық анықталмаған. Тілді өңдеуде адам осы екіұштылықтарды шешу және жетіспейтін ақпаратты толтыру үшін өзінің *парасаттылық* мен пайымдау қабілетіне сүйенеді. Ал машиналық оқытуға негізделген табиғи тілді өңдеу модельдерінде ақыл-ой жетіспейді және адамдар үшін елеусіз болып көрінетін қателіктер жиі жасалады. *Парасаттылықты* сәтті біріктіру мұндай қателерді азайтуға және табиғи тілді өңдеу модельдерін жақсартуға көмектеседі. Зерттеудің мақсаты табиғи тілді өңдеу модельдеріне жалпы *парасаттылықты* біріктіру әдістері мен алгоритмдерін әзірлеу болып табылады. Зерттеу нысаны модельдерді өңдеудегі парасаттылық болып табылады. Зерттеудің міндеттеріне математикалық есептерді шешуге арналған табиғи тілді өңдеу модельдеріне парасаттылықты біріктіру алгоритмдерін жасау; шешуі парасаттылық мен арифметикалық пайымдауды қажет ететін ҰБТ-ның математикалық есептерінің деректер жинағын әзірлеу; жалпы парасаттылықты табиғи тілді өңдеу модельдеріне біріктіру және жиналған деректер жиынтығы бойынша салыстыру нәтижелері кіреді. Зерттеу тақырыбы бойынша жарияланған ғылыми мақалалар жан-жақты зерттелді. Сондай-ақ, дауыс беру, шешімдер тізбегі және шешімдерді тексеру арқылы табиғи тілді өңдеу үлгілеріне парасаттылықты біріктіру алгоритмі әзірленді. Жасалған алгоритмнің көмегімен мәліметтер жиынтығындағы есептерді шешудің дәлдігін 63,7%-ға дейін арттыру мүмкін болды, бұл бастапқы нәтижеден 14%-ға артық.

**Диссертация көлемі мен құрылымы:** 61 бет, 9 секция

**Суреттер, кестелер, пайдаланылған әдеби дереккөздер саны:** 16 иллюстрация, 4 кесте, 22 дереккөз

**Кілт сөздер:** парасаттылық; табиғи тілді өңдеу; үлкен тілдік модельдер; тілдік модельдер

**Жарияланымдар туралы мәліметтер:** Sagingaliyev, B., Aitakhunova, Z., Shaimerdenova, A., Akhmetov, I., Pak, A., & Jaxylykova, A. (2022). A Bibliometric Review of Methods and Algorithms for Generating Corpora for Learning Vector Word Embeddings. In O. Pichardo Lagunas, J. Martínez-Miranda, & B. Martínez Seis (Eds.), *Advances in Computational Intelligence* (pp. 148–162). Cham: Springer Nature Switzerland

**АННОТАЦИЯ**

Одна из основных проблем исследований в обработке естественного языка заключается в том, что человеческий язык часто неоднозначен и не полностью определен. При обработке языка человек в значительной степени полагается на свой *здравый смысл* и способность рассуждать, чтобы разрешить эти неоднозначности и заполнить недостающую информацию. В то время, моделям обработки естественного языка, основанным на машинном обучении, не хватает *здравого смысла*, и они часто допускают ошибки, которые могут показаться тривиальными для людей. Успешное интегрирование здравого смысла может помочь уменьшить такие ошибки и улучшить модели обработки естественного языка. Целью исследования является разработка методов и алгоритмов интеграции здравого смысла в модели обработки естественного языка. Объектом исследования является здравый смысл в моделях обработки. Задачи исследования включают в себя разработку алгоритмов интеграции здравого смысла в модели обработки естественного языка для решения задач по математике, разработку набора данных математических задач ЕНТ, решение которых требует здравого смысла и арифметических рассуждений, а также проведение интеграции здравого смысла в модели обработки естественного языка и сравнение результатов на собранном наборе данных. Было проведено подробное изучение опубликованных научных статей по теме исследования. А также был разработан алгоритм интеграции здравого смысла в модели обработки естественного языка с использованием голосования, цепи решений и верификации решений. С помощью разработанного алгоритма получилось увеличить точность решения задач из набора данных до 63.7%, что на 14% больше изначального результата.

**Объем и структура диссертации:** 61 страниц, 9 секций

**Количество иллюстраций, таблиц, использованных литературных источников:** 16 иллюстраций, 4 таблиц, 22 источников

**Ключевые слова:** здравый смысл; обработка естественного языка, большие языковые модели

**Сведения о публикациях:** Sagingaliyev, B., Aitakhunova, Z., Shaimerdenova, A., Akhmetov, I., Pak, A., & Jaxylykova, A. (2022). A Bibliometric Review of Methods and Algorithms for Generating Corpora for Learning Vector Word Embeddings. In O. Pichardo Lagunas, J. Martínez-Miranda, & B. Martínez Seis (Eds.), *Advances in Computational Intelligence* (pp. 148–162). Cham: Springer Nature Switzerland.

**CONTENTS**

[**INTRODUCTION** 6](#_Toc137641419)

[**1** **A PROBLEM OF FORMALIZING THE COMMON SENSE** 6](#_Toc137641420)

[**2** **DEFINITION OF COMMON SENSE** 11](#_Toc137641421)

[**3** **COMMON SENSE AND LARGE LANGUAGE MODELS (LLM)** 13](#_Toc137641422)

[**4** **COMMONSENSE REASONING INTEGRATION METHODS IN NATURAL LANGUAGE PROCESSING** 14](#_Toc137641423)

[4.1. Zero-Shot Learning Approaches. 14](#_Toc137641424)

[4.2. Few-Shot Learning Approaches. 15](#_Toc137641425)

[4.3. Supervised Fine-Tuning Approaches (SFT) 23](#_Toc137641426)

[**5** **PROJECT DETAILS** 24](#_Toc137641427)

[5.1. Relevance 24](#_Toc137641428)

[5.2. Novelty 25](#_Toc137641429)

[5.3. Research Objectives 25](#_Toc137641430)

[5.4. Scientific and Practical Significance 26](#_Toc137641431)

[**MAIN PART** 27](#_Toc137641432)

[**6** **METHODS** 27](#_Toc137641433)

[**7** **METHODOLOGY** 32](#_Toc137641434)

[7.1. Experimental design 32](#_Toc137641435)

[7.2. Dataset 32](#_Toc137641436)

[7.3. Large Language Model (LLM) 37](#_Toc137641437)

[7.4. Small Language Model as a Verifier 40](#_Toc137641438)

[7.5. Voting 42](#_Toc137641439)

[7.6. Evaluation Metrics 45](#_Toc137641440)

[**8** **RESULTS** 47](#_Toc137641441)

[8.1. Baseline performance. 47](#_Toc137641442)

[8.2. Single answer vs Majority Voting. 48](#_Toc137641443)

[8.3. Confidence score in majority voting. 49](#_Toc137641444)

[8.4. Verifier Model. 50](#_Toc137641445)

[8.5. Majority Voting vs Voting Verifier. 53](#_Toc137641446)

[8.6. Adversarial attacks through incorrect hints. 53](#_Toc137641447)

[**9** **DISCUSSION** 55](#_Toc137641448)

[**CONCLUSION** 59](#_Toc137641449)

[**REFERENCES** 60](#_Toc137641450)

# **INTRODUCTION**

## **A PROBLEM OF FORMALIZING THE COMMON SENSE**

Common sense denotes the shared knowledge and comprehension that individuals within a society possess, equipping them with the capacity to navigate and comprehend the world. It encompasses a wide range of practical knowledge, intuitive reasoning, and everyday understanding regarding the operation of diverse phenomena, adherence to social norms, and fundamental cause-and-effect principles. By drawing upon their prior experiences and understanding of the world, common sense empowers individuals to make well-informed decisions, interpret complex situations, and anticipate probable outcomes.

The formalization of common sense presents a significant challenge within the realms of artificial intelligence and cognitive science. This undertaking involves capturing and representing the extensive repository of implicit knowledge and intuitive reasoning inherent in human cognition. Various approaches have been proposed to formalize common sense, each exhibiting its own set of strengths and limitations. Some notable approaches that have been put forward are described below.

## **DEFINITION OF COMMON SENSE**

Defining common sense in natural language processing (NLP) can be approached both qualitatively and quantitatively.

***Qualitative definitions.*** Its qualitative definition entails capturing the inherent comprehension and knowledge that humans possess instinctively. Here are some qualitative approaches to defining common sense in NLP:

*Everyday Comprehension.* Common sense encompasses the fundamental understanding of everyday concepts, connections, and reasoning that is widely shared among individuals. It encompasses knowledge about the physical world, social norms, cause-and-effect relationships, and general principles that guide human behavior.

*Contextual Interpretation.* Common sense involves the capability to interpret and understand language within its context. This includes grasping implied meanings, metaphors, idiomatic expressions, and pragmatics, which enable effective communication and comprehension.

*Implicit Assumptions and Expectations.* Common sense is rooted in the implicit assumptions and expectations that underlie human understanding. It encompasses expectations about the behavior of objects, events, and people, as well as assumptions about the regularities and predictability of the world.

*Reasoning and Deduction.* Common sense entails the ability to reason and deduce conclusions based on available information. It involves filling in missing details, making logical deductions, and drawing conclusions that align with everyday knowledge and expectations.

## **COMMON SENSE AND LARGE LANGUAGE MODELS (LLM)**

Commonsense reasoning in large language models refers to the ability of these models to understand and generate responses that align with human-like common sense knowledge and reasoning. LMs as GPT can be used to solve math problems and they perform well when the problems are in English and are not very hard. With help of prompting, few-shot learning and chain-of-thought techniques, the performance of the LLMs increases further (e.g.: GPT-4 solves grade 5-6 math problems from GSM8K dataset [2] with 95% accuracy [3]). Large-scale language models like OpenAI's GPT-3, GPT-4 have showcased impressive proficiency in comprehending and generating natural language. Nevertheless, they encounter hurdles when it comes to effectively incorporating and reasoning with commonsense knowledge. Commonsense reasoning involves leveraging background knowledge and implicit understanding to make inferences and draw conclusions in everyday scenarios.

## 

## **COMMONSENSE REASONING INTEGRATION METHODS IN NATURAL LANGUAGE PROCESSING**

## **Zero-Shot Learning Approaches**

***Zero-shot learning and Chain-of-Thought with Text-davinci-002-175B [4][5][6].*** The paper introduces Zero-shot-CoT, a technique that enables multi-hop reasoning across different tasks using a single template, eliminating the need for step-by-step few-shot examples.

## **Few-Shot Learning Approaches**

***DIVERSE [10][11][12].*** The diverse approach, called DIVERSE (diverse verifier on reasoning step), aims to improve the reasoning capabilities of large language models. It introduces three main innovations: diverse prompts, voting verifier, and stepwise verifier.

## **Supervised Fine-Tuning Approaches (SFT)**

***SFT and Reward Models with DeepMind 70B Model [21][22].*** This scholarly article provides a comparative investigation of two separate strategies for guiding large language models (LMs) for reasoning tasks: outcome-oriented supervision, which prioritizes the accuracy of the final solution, and process-oriented supervision, which emphasizes the correctness of each reasoning step leading to the final outcome. The GSM8K task was employed as an experimental platform for this study, a task where the model is given a mathematical problem and is expected to provide both the method of the solution and the final answer.

## **PROJECT DETAILS**

## **Relevance**

One of the main problems of research in natural language processing is that human language is often ambiguous and not fully defined. In language processing, a person relies heavily on their *common sense* and *reasoning* ability to resolve these ambiguities and recover the missing information. On the other hand, natural language processing models based on machine learning lack *common sense* and often make mistakes that might seem trivial to humans. Sometimes they hallucinate, display inconsistencies, and return statements that contradict commonsense reasoning. Integrating commonsense reasoning into NLP models would help address the mentioned problems and increase the performance of models in various downstream tasks and drive NLP closer to the goal of general AI. The task, however, is complicated and challenging. This project tries to improve general commonsense and arithmetic reasoning of NLP models in solving math word problems, which currently remains as a hard task for

# **MAIN PART**

## **METHODS**

This study focuses on enhancing the ability of Large Language Models (LLMs) to engage in Commonsense Reasoning, specifically in the context of solving Unified National Test (UNT) math problems. The objective is to optimize the reasoning skills of LLMs when tackling UNT problems by employing various techniques such as ones outlined in the DiVeRSe paper. **It uses diverse prompts, zero-shot and few-shot learning, and chain-of-thought techniques in obtaining diverse answers from the LLM. Then it introduces a voting verifier as the next step before applying the majority voting. By applying different combinations of these techniques, we try to maximize the performance of the model on the UNT dataset, as well as determine the increase in performance compared to a baseline case where no techniques are used.**

## **METHODOLOGY**

## **Experimental design**

**The methodology for the project consists of the following steps:**

1. **Collecting UNT math problems with answers and solution steps from different sources and combining them into a dataset that will further be referred to as a UNT dataset. Splitting the dataset into training and test sets.**
2. **Applying zero-shot learning and chain-of-thought techniques in creating diverse prompts for the training problems from the UNT dataset.**
3. **Inputting the constructed prompts and using GPT-3.5 LLM to obtain solutions to the problems from the dataset.**
4. **Training a verifier model for reasoning paths by providing examples of correct and incorrect reasoning steps.**
5. **Applying (majority) voting either directly or on the output of the verifier model and selecting the answer.**
6. **Measuring the performance on the training and test sets to evaluate the system.**
7. **Repeating the same steps above with additionally providing incorrect hints along with the questions to the LLM.**

The detailed pipeline of the experiment is illustrated in Figure 1.

## **Dataset**

GSM8K has been a popular dataset that contains 6th grade math problems in English language, and it has been quite challenging for Large Language Models to solve. At the beginning, the models resulted in performance levels ranging from approximately 30% to 40% accuracy on the dataset. Over time, with the evolution of LLMs and application of various novel techniques, the performance of the LLMs on solving the tasks from the dataset has been steadily increasing and GPT-4 was reported to score 95% accuracy on the dataset, demonstrating a remarkable progress.

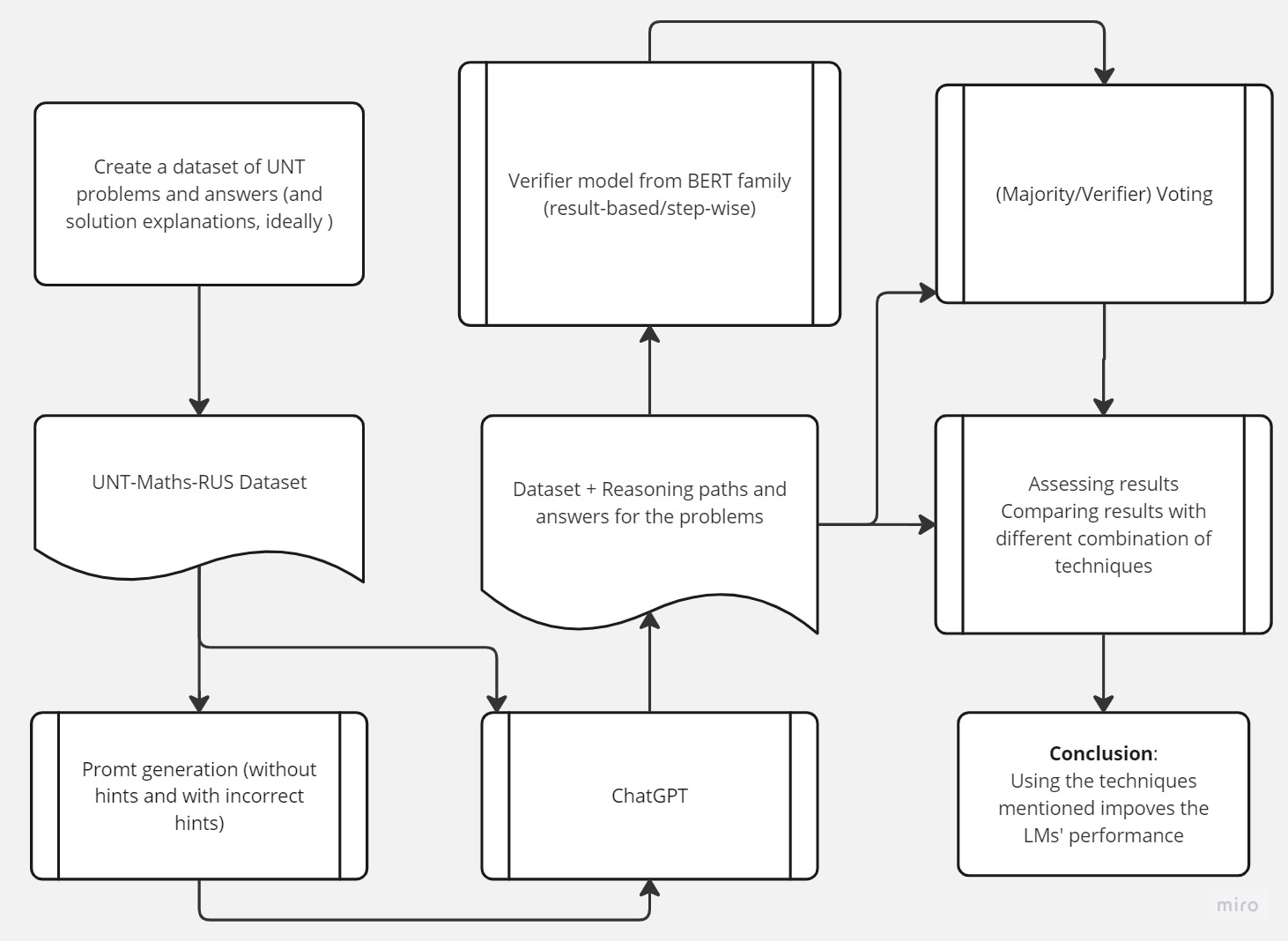


Figure 1 Experiment pipeline

***Dataset description.*** Collected dataset is referred to as the UNT dataset. It consists of 969 math problems in Russian language that are similar to the problems encountered in the standardized test. The dataset consists of the following columns:

* *Q\_num*: number of the question;
* *Question*: text of the problem;
* *Options*: five options (from “A” to “E”) among which one is the correct answer to the problem, presented in the “letter) answer” format;
* *Answer*: option letter that refers to the correct answer (A/B/C/D/E).

Table 1 displays samples from the UNT dataset.

***Exploratory data analysis.*** *Distribution of answers****.*** First question that arises upon inspecting the dataset is whether the dataset is balanced, as imbalanced datasets impose problems in both training and evaluation stages. Such datasets require a more careful selection of training strategy and evaluation metrics.

Table 1 – Samples of the problems from the UNT dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Q\_num | Question | Options | Answer |
| 1 | Заяц соревновался с черепахой в беге на 100 метров. Когда заяц прибежал к финишу, черепахе оставалось до него еще 90 метров. На сколько метров надо отодвинуть назад стартовую линию для зайца, чтобы при новой попытке оба бегуна пришли к финишу одновременно? | A) 90  B) 100  C) 10  D) 900  E) 1000 | D |
| 2 | Среднее арифметическое шести чисел равно 70, а среднее других четырех чисел равно 100. Все десять чисел сложили. Чему равно их среднее арифметическое? | A) 85  B) 82  C) 17  D) 14  E) 36 | B |
| 3 | В коробке есть белые, зеленые и синие шары. Количество синих шаров не меньше белых и равно одной трети количества зеленых шаров. Сумма синих и белых шаров равняется 55. В одноразовой попытке, несмотря, сколько нужно взять шаров, чтобы был 1 белый? | A) 100  B) 106  C) 109  D) 113  E) 119 | D |
| 4 | Большой куб, окрашенный в зеленый цвет, распилили на 27 маленьких одинаковых кубиков. Сколько маленьких кубиков имеют только одну окрашенную грань? | A) 12  B) 8  C) 9  D) 18  E) 6 | E |
| 5 | Известно, что из 1000 произвольно выбранных деталей примерно двенадцать деталей бракованы. Сколько приблизительно бракованных деталей окажется среди 5500 деталей, отгруженных в мастерскую? | A) 72  B) 52  C) 66  D) 68  E) 63 | C |
| 6 | Найдите среднее арифметическое и медиану ряда. 16; 19; 42; 35; 37; 13. | A) 27; 26  B) 26;27  C) 25;23  D) 27;19  E) 27;27 | E |

Distribution of answers presented in Figure 2 shows that the dataset is balanced and approximately equally distributed, and that there are no options that have a lot less presence compared to other options: option “A” – 14.3%, option “B” – 23.1%, option “C” – 26.0%, option “D” – 20.4%, and option “E” – 16.1%. This suggests that a random guess would on average result in accuracy of about 20%, and the models are first compared against the random guess when judging their ability to discriminate.

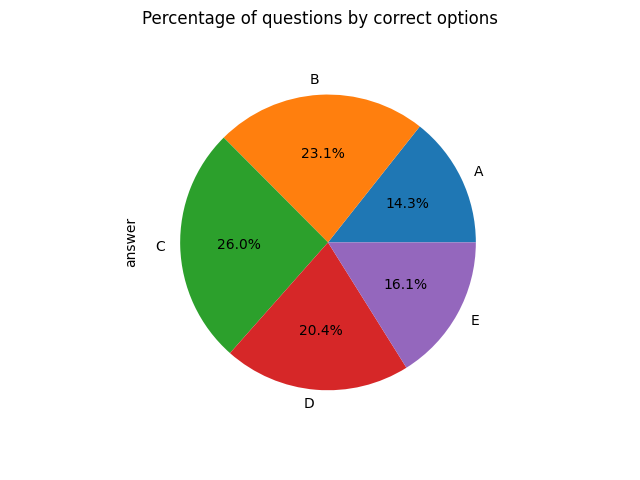


Figure 2 Distribution of answers in the UNT dataset

Another important part of the exploratory data analysis is examining the input features of the dataset. In textual data, it is usually looking at the length of the text, identifying the most frequent words (key words), measuring number of characters, etc.

*Distribution of the length of questions.* Examining the length of the questions helps us to understand the nature of the questions: how long the questions are, how the length of the questions varies in the dataset, etc. According to the statistical descriptions of length of questions provided in table 2, the average length of the questions is approximately 25.9 words, and the median is quite close to the average – 24.0 words. Minimum number of words in the questions from the dataset is 3.0, while the maximum number of words in the questions is 175.0. It is also to note that 75% of the questions have length of greater than or equal to 32.0. Figure 3 illustrates the histogram of the length of questions, providing the detailed description of the distribution of the questions in the dataset by their lengths.

Table 2 – Statistical description of the length of questions (in words)

|  |  |
| --- | --- |
| Parameter | Value |
| count | 969.0 |
| mean | 25.88 |
| std | 14.09 |
| min | 3.0 |
| 25% | 16.0 |
| 50% | 24.0 |
| 75% | 32.0 |
| max | 175.0 |

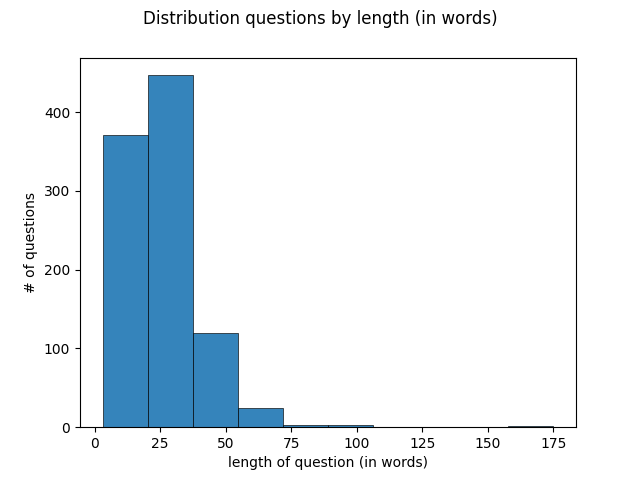


Figure 3 Distribution of questions by length in words

## **RESULTS**

This section presents the main results obtained over the course of the experimental procedure designed to answer the objective questions of the study.

## **Baseline performance**

A baseline performance evaluation was conducted by directly comparing the output of the Large Language Model (LLM) to the correct labels present in the dataset. To account for the multiple sampled answers from the LLM for each question, the average accuracy over 100 answers was computed to establish the baseline accuracy value. Consequently, the obtained baseline accuracy is approximately 48.90%. The baseline accuracy is also referred to as a single answer accuracy.

## **Single answer vs Majority Voting**

As the next step, the output generated by the Large Language Model (LLM) underwent the majority voting process to determine the final answers. The number of voters, denoted as *n*, was varied in the experiment, ranging from 2, 5 to 100, with a step size of 5. Figure 7 illustrates the results of the majority voting. The figure depicts the relationship between the accuracy and the number of voters employed in the majority voting. The results demonstrate a significant improvement in the accuracy as the number of voters *n* increases. Notably, the maximum accuracy of 63.67% was attained when number of voters is equal to 100. It represents an accuracy gain of approximately 14% compared to the baseline accuracy. However, it is worth noting that the rate of accuracy improvement starts to gradually diminish from *n = 20.*

## **DISCUSSION**

***Commonsense reasoning in LLMs.*** The inquiry into the capability of Large Language Models (LLMs) to engage in commonsense reasoning for solving tasks requiring such reasoning stands as a pivotal research question. By analyzing the outcomes of employing LLMs to solve math problems from the Unified National Test (UNT) dataset, we obtained an average single-answer accuracy of approximately 48.9%. Comparatively, this performance significantly surpasses that of random guessing, which has a 20% chance of being correct.

# **CONCLUSION**

In conclusion, this study aimed to enhance the ability of Large Language Models (LLMs) to engage in Commonsense Reasoning for solving Unified National Test (UNT) math problems. By employing techniques such as diverse prompts, zero-shot and few-shot learning, chain-of-thought methods, and the introduction of a voting verifier, the performance of LLMs on the UNT dataset was optimized.

The results showed that LLMs were capable of solving a substantial portion of the math problems, with an average single-answer accuracy of approximately 48.9%. This performance significantly surpassed random guessing, indicating the LLMs' aptitude for engaging in commonsense reasoning. The implementation of majority voting further improved the accuracy by 14%, reaching around 63.7%. These findings demonstrate the potential of LLMs in solving complex problems and highlight the effectiveness of employing diverse prompts and voting mechanisms.  
The utilization of confidence scores as a measure of stability provided valuable insights into the behavior of LLMs. Higher confidence scores correlated with a higher likelihood of the answer being correct, with scores above 0.9 achieving a 91% accuracy rate. This understanding of confidence scores helps identify challenging examples and improves the overall reliability of LLM responses.

The introduction of a verifier model demonstrated higher precision in negative predictions, indicating its ability to identify incorrect answers accurately. While the verifier model did not lead to significant improvements in overall accuracy, it showed potential in scenarios with a smaller number of voters. Further refinement and training data could enhance the verifier model's performance and contribute to decision-making processes.

However, the study also revealed the vulnerability of LLMs to adversarial attacks, resulting in a substantial decline in accuracy. This emphasizes the need for robust mechanisms to enhance the resilience of LLMs against such attacks without compromising the reliability of their responses.  
Future research directions include conducting investigations on a larger scale with more extensive datasets, incorporating additional human-annotated data, and exploring few-shot learning approaches. These efforts aim to improve the performance and generalization of LLMs in solving math problems, enhance the voting verifier model, and expand the practical applicability of LLMs.

In summary, this study has contributed to advancing the capabilities of LLMs in commonsense reasoning for math problem-solving. Further research and development are necessary to enhance their performance, improve resilience against adversarial attacks, and expand their capabilities for solving a wide range of math problems in real-world applications.

# **REFERENCES**

1. *Speer, R., Chin, J., & Havasi, C.* (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). <https://doi.org/10.1609/aaai.v31i1.11164>
2. *Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., … Schulman, J.* (2021). Training Verifiers to Solve Math Word Problems. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2110.14168
3. OpenAI. (2023). GPT-4 Technical Report. *arXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/2303.08774
4. *Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y.* (2023). Large Language Models are Zero-Shot Reasoners. *arXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/2205.11916
5. *Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D.* (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/2201.11903
6. *Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W.* (2021). What Makes Good In-Context Examples for GPT-3? *arXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/2101.06804
7. *Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., … Zhou, D.* (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2203.11171>
8. *Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J.* (2022). Large Language Models Can Self-Improve. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2210.11610>
9. *Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., … Fiedel, N.* (2022). PaLM: Scaling Language Modeling with Pathways. *arXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/2204.02311
10. *Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J.-G., & Chen, W.* (2023). Making Large Language Models Better Reasoners with Step-Aware Verifier. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2206.02336>
11. *Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Matthewson, K., Tessler, M. H., Creswell, A., … Hill, F.* (2022). Can language models learn from explanations in context? *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2204.02329>
12. *Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., … Chi, E.* (2023). Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv [Cs.AI]*. Retrieved from <http://arxiv.org/abs/2205.10625>
13. *Xie, Y., Kawaguchi, K., Zhao, Y., Zhao, X., Kan, M.-Y., He, J., & Xie, Q.* (2023). Decomposition Enhances Reasoning via Self-Evaluation Guided Decoding. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2305.00633>
14. *Ni, A., Iyer, S., Radev, D., Stoyanov, V., Yih, W.-T., Wang, S. I., & Lin, X. V.* (2023). LEVER: Learning to Verify Language-to-Code Generation with Execution. *arXiv [Cs.LG]*. Retrieved from <http://arxiv.org/abs/2302.08468>
15. *Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., … Misra, V.* (2022). Solving Quantitative Reasoning Problems with Language Models. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2206.14858>
16. *Zheng, C., Liu, Z., Xie, E., Li, Z., & Li, Y.* (2023). Progressive-Hint Prompting Improves Reasoning in Large Language Models. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2304.09797>
17. *Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T.* (2023). Complexity-Based Prompting for Multi-Step Reasoning. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2210.00720>
18. *Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., … Clark, P.* (2023). Self-Refine: Iterative Refinement with Self-Feedback. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2303.17651>
19. *Fu, J., Ng, S.-K., Jiang, Z., & Liu, P.* (2023). GPTScore: Evaluate as You Desire. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2302.04166>
20. *Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., … Gao, J.* (2023). Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2302.12813>
21. *Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., … Higgins, I.* (2022). Solving math word problems with process- and outcome-based feedback. *arXiv [Cs.LG]*. Retrieved from <http://arxiv.org/abs/2211.14275>
22. *Zelikman, E., Wu, Y., Mu, J., & Goodman, N.* (2022). Star: Bootstrapping reasoning with reasoning. Advances in Neural Information Processing Systems, 35, 15476-15488.

**APPENDIX A**