

COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

EXPLORING ADVANCED REASONING ABILITIES
OF LARGE LANGUAGE MODELS IN SLOVAK

BACHELOR THESIS

2024
ADAM ZAHRADNÍK

DRAFT

COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

EXPLORING ADVANCED REASONING ABILITIES
OF LARGE LANGUAGE MODELS IN SLOVAK
BACHELOR THESIS

Study Programme: Applied Computer Science
Field of Study: Computer Science
Department: Department of Applied Informatics
Supervisor: Mgr. Marek Šuppa

DRAFT



ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Adam Zahradník

Študijný program: aplikovaná informatika (Jednoodborové štúdium, bakalársky I. st., denná forma)

Študijný odbor: informatika

Typ záverečnej práce: bakalárska

Jazyk záverečnej práce: anglický

Sekundárny jazyk: slovenský

Názov: Exploring Advanced Reasoning Abilities of Large Language Models in Slovak
Skúmanie pokročilých schopností uvažovania veľkých jazykových modelov v slovenčine

Anotácia: V nedávnej minulosti veľké jazykové modely (LLM) preukázali pôsobivé schopnosti pri riešení rôznych kvantitatívnych úloh a úloh v oblasti vedomostí v oblastiach, ako je matematika, fyzika a informatika. Tieto úspechy sa zvyčajne hodnotia pomocou metodológií na základnych pre základné a stredné školy, často odvodených zo štandardizovaných testov. Tento prístup však predstavuje problém: štandardizované testy sa často dostanú do tréningových datasetov LLM, čo vede k skresleniu hodnoteniam ich výkonu. Okrem toho hodnotenie rozumových schopností (či schopnosti uvažovania) LLM prebieha primárne v angličtine, čo vyvoláva obavy z použiteľnosti a zovšeobecneniteľnosti výsledkov týchto hodnotení.

Ciel: Ciele bakalárskej práce zahŕňajú (ale nie sú obmedzené na)

- analýza súčasného stavu v hodnotení schopnosti uvažovania
- vytváranie a/alebo zhromažďovanie logických hodnotiacich datasetov v slovenskom jazyku na základe úloh/cvičení z matematických, fyzikálnych alebo informačných olympiád, ako aj rôznych korešpondenčných seminárov z rovnakej oblasti
- vyhodnotenie najmodernejších LLM na pripravených datasetoch
- analýza výstupov (a chybových režimov) vytvorených modelmi s najlepšími výsledkami

Literatúra: Henkuycks, Dan, et al. "Measuring mathematical problem solving with the maln dataset." arXiv preprint arXiv:2103.03874 (2021). (<https://arxiv.org/pdf/2103.03874.pdf>)

Cobbe, Karl, et al. "Training verifiers to solve math word problems." arXiv preprint arXiv:2110.14168 (2021). (<https://arxiv.org/pdf/2110.14168.pdf>)

Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in Neural Information Processing Systems 35 (2022): 24824-24837. (<https://arxiv.org/abs/2201.11903>)

Sawada, Tomohiro, et al. "Arb: Advanced reasoning benchmark for large language models." arXiv preprint arXiv:2307.13692 (2023). (<https://arxiv.org/pdf/2307.13692.pdf>)

Vedúci: Mgr. Marek Šuppa

Katedra: FMFI.KAI - Katedra aplikovanej informatiky



Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 15.10.2023

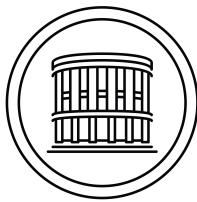
Dátum schválenia: 16.10.2023

doc. RNDr. Damas Gruska, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

DRAFT



THESIS ASSIGNMENT

Name and Surname: Adam Zahradník
Study programme: Applied Computer Science (Single degree study, bachelor I. deg., full time form)
Field of Study: Computer Science
Type of Thesis: Bachelor's thesis
Language of Thesis: English
Secondary language: Slovak

Title: Exploring Advanced Reasoning Abilities of Large Language Models in Slovak

Annotation: In the recent past, Large Language Models (LLMs) have shown impressive capabilities in tackling various quantitative reasoning and knowledge challenges in fields like mathematics, physics and computer science. These achievements are typically assessed using benchmarks designed for primary and secondary school levels, often derived from standardized tests. However, this approach poses a problem: standardized tests frequently find their way into the training data of LLMs, leading to skewed performance evaluations. Additionally, the evaluation of LLMs' reasoning abilities primarily occurs in English, raising concerns about the applicability and generalizability of the results.

Aim: The goals of the bachelor's thesis include (but are not limited to)
- analysis of the current state-of-the-art in reasoning capability evaluation
- creation and/or collection of reasoning evaluation datasets in Slovak language based on tasks/exercises (Olympiads in Mathematics, Physics or Informatics, as well as various correspondence seminars in the same areas)
- evaluation of state-of-the-art LLMs on the prepared datasets
- analysis of the outputs (and error modes) produced by the best performing models

Literature:
Hendrycks, Dan, et al. "Measuring mathematical problem solving with the math dataset." arXiv preprint arXiv:2103.03874 (2021). (<https://arxiv.org/pdf/2103.03874.pdf>)
Cordts, Karl, et al. "Training verifiers to solve math word problems." arXiv preprint arXiv:2110.14168 (2021). (<https://arxiv.org/pdf/2110.14168.pdf>)
Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in Neural Information Processing Systems 35 (2022): 24824-24837. (<https://arxiv.org/abs/2201.11903>)
Sawada, Tomohiro, et al. "Arb: Advanced reasoning benchmark for large language models." arXiv preprint arXiv:2307.13692 (2023). (<https://arxiv.org/pdf/2307.13692.pdf>)

Supervisor: Mgr. Marek Šuppa
Department: FMFI.KAI - Department of Applied Informatics
Head of department: doc. RNDr. Tatiana Jajcayová, PhD.



48302124

Comenius University Bratislava
Faculty of Mathematics, Physics and Informatics

Assigned: 15.10.2023

Approved: 16.10.2023 doc. RNDr. Damas Gruska, PhD.
Guarantor of Study Programme

.....
Student

.....
Supervisor

DRAFT

DRAFT

Acknowledgments: Tu môžete podakovať školiteľovi, prípadne ďalším osobám, ktoré vám s prácou nejako pomohli, poradili, poskytli dátá a podobne.

Abstrakt

Slovenský abstrakt v rozsahu 100-500 slov, jeden odstavec. Abstrakt stručne sumarizuje výsledky práce. Mal by byť pochopiteľný pre bežného informatika. Nemal by teda využívať skratky, termíny alebo označenie zavedené v práci, okrem tých, ktoré sú všeobecne známe.

Kľúčové slová: jedno, druhé, tretie (prípadne štvrté, piate)

DRAFT

Abstract

Abstract in the English language (translation of the abstract in the Slovak language).

Keywords:

DRAFT

Contents

1	Introduction	1
1.1	Large Language Models	1
1.2	Prompting Techniques	2
1.2.1	Zero-, One- and Few-shot Prompting	2
1.2.2	Chain-of-Thought	3
1.2.3	Zero-shot Chain-of-Thought	3
1.2.4	Generated Knowledge	3
1.2.5	Dual Prompt Generated Knowledge	3
1.2.6	Least-to-Most Prompting	4
1.3	Existing Datasets	4
1.4	Evaluating LLM Answers	5
1.5	Prior Research	6
1.6	Related Benchmarks	7
2	Our Goals	9

DRAFT

Chapter 1

Introduction

Latest developments in the field of artificial intelligence led to great improvements in the abilities of large language models to solve many different types of tasks. Prior work demonstrates ambitious results on tasks that challenge the models' ability to reason about maths, physics and informatics. Researchers introduced many datasets and methods to evaluate the large language models' abilities to solve these problems.

These datasets and methods are usually based off standardized elementary and high school tests in these fields. Standardized tests that are publicly available can end up included in the models' training data, which can bias the models' evaluation results. Currently, most available datasets and evaluation methods are in English, which raises the question whether observed large language models' abilities can be generalized to other languages as well.

As organizers of Slovak high school competitions ourselves, we want to find out how current large language models deal with the problems to gain insights into their capabilities. A fellow organizers already experimented with using large language models to solve our problems, but we want to try it out on a much larger dataset. This will allow us to better understand the models, their pitfalls and explore different techniques of working with them.

1.1 Large Language Models

Large language models are a recent advancement in artificial intelligence, particularly in the realm of generative models. They operate by first receiving an initial input text, which is also called a prompt. Then, the model uses its neural network to predict the next word or token. This token is then appended to the prompt, creating an extended text input. The process is repeated until a specified length of text or another stopping criterion is reached. The result is a coherent and contextually relevant piece of text, generated entirely by the model's learned patterns and associations within its training

data.

The prompt that was given to the large language model together with the generated text is called a context window. The model can recall information from its context window. This allows the user to provide additional context to the model. Additionally, the model can in some circumstances use its context window as a kind of scratch pad and thus better prepare its output. This is taken advantage of by some of the prompting techniques.

It was observed that large language models exhibit common-sense reasoning capabilities [16]. More importantly, these models can perform advanced reasoning needed to solve mathematical problems [14]. Even though large language models are often less capable than humans in solving such problems, it still allows them to be used in a wide range of applications.

In our research, we will utilize language models provided by OpenAI, specifically the GPT-3.5-Turbo [13] and GPT-4 [14]. These models will be used to benchmark our dataset, enabling us to assess and compare the performance of each model under various scenarios.

1.2 Prompting Techniques

Prior research has shown that manipulating the way in which the model’s prompt is constructed can have a significant impact on the quality of the resulting output. In this section, we introduce successful techniques that we use to compare large language model performance in the Slovak language.

1.2.1 Zero-, One- and Few-shot Prompting

Even though large language models are trained on generic text corpus datasets, prior research has indicated that LLMs do not require fine-tuning the model on the desired task [17][3].

Brown et al. [3] shows that this fine-tuning step can be replaced by a technique called few-shot prompting. Few-shot prompting provides the model with few examples of the desired task, along with sample solutions right in the model’s input. This measurably improves the models’ capability to solve the desired task, even though the model was not trained to solve that particular task in the first place. Few-shot prompting is typically done with 10 to 100 task examples, depending on the size of the model’s context window.

There are two related techniques to few-shot prompting, introduced by Brown et al. [3]: one-shot and zero-shot prompting. One-shot prompting is done in the same way as few-shot, but the model is provided with only one example of the task. Zero-shot

prompting is a similar, but generally less effective, technique in which the model is provided with a natural language description of the task instead of any examples.

1.2.2 Chain-of-Thought

Another promising prompting technique is chain-of-thought introduced by Wei et al. [23]. This technique mimics one’s own thought process when solving tasks. The goal of chain-of-thought prompting is to make the model generate a series of intermediate steps that lead to the final answer of a problem. Wei et al. [23] shows that large language models are capable of generating such chain-of-thoughts when provided with such chain-of-thoughts in the examples used for few-shot prompting. The model is provided with example solutions that walk the reader through the different steps leading to the final answer.

1.2.3 Zero-shot Chain-of-Thought

Classical chain-of-thought prompting as introduced by Wei et al. [23] has the disadvantage of needing to provide examples of task solutions including chain-of-thoughts, which are usually not readily available.

To address such problems, Kojima et al. [7] introduced a simple, yet effective technique, called zero-shot chain-of-thought. The idea is that when the model is prompted to ”think step by step”, it can generate a chain-of-thought without needing any examples beforehand. So, the model is prompted with the question and a simple prompt like ”Let’s think step by step” to force it to generate a chain-of-thought.

1.2.4 Generated Knowledge

Another similar approach to zero-shot chain-of-thought was demonstrated by Liu et al. [11]. The generated knowledge prompting technique leverages the fact that large language models can use their context windows for short-term memory. This is very similar to some teaching techniques employed when teaching humans new concepts.

The model is prompted to first describe all concepts relevant to solving the problem and then attempt to solve the problem. This way, the model will introduce a lot of new information into its context window. It can later retrieve information from the context window to help itself to solve the problem.

1.2.5 Dual Prompt Generated Knowledge

The generated knowledge prompting method has a slight disadvantage - the model has a limited number of words or tokens it can produce. Because of this, it can spend a lot

of its available space on preparing the relevant context and end up not having enough tokens left for the solution itself.

Dual prompt generated knowledge improves on this by prompting the model to only generate the relevant context. After it generates the context, the model is prompted again with the original question and the context it generated. This allows the model to generate longer answers.

1.2.6 Least-to-Most Prompting

Least-to-most prompting takes advantage of the model’s context window combined with multiple prompts. The idea introduced by Zhou et al. [24] is that we break the problem into smaller sub-problems, which are then solved sequentially.

We start by prompting the model with the problem and ask it to list out the sub-problems that are required to solve the whole problem. We then take the first sub-problem it generates and ask it to solve it. This is usually done by prompting it with the original problem and a question to solve a given sub-problem. The solution of the sub-problem is then appended to the prompt along with another sub-problem. This process is repeated until the model solves all sub-problems. At that point, we should have the whole solution.

1.3 Existing Datasets

Large language models have already been evaluated on mathematical reasoning tasks by researchers using numerous datasets. Most of these datasets were created by scraping problems from the internet. We provide a comparison of few selected datasets related to our work to better understand the types and problems involving creating such a dataset.

MultiArith released by Roy and Roth [18] contains multistep arithmetic problems without irrelevant quantities. That means that those problems require a combination of different arithmetic operations to get the answer and don’t have any information that isn’t needed to solve the problem. The dataset contains symbolic solutions.

Math23K by Wang, Liu, and Shi [22] consists of Chinese elementary school maths problems scraped from the internet. This dataset contains only problems with single linear unknown variable. This dataset contains only symbolic solutions.

AQuA introduced by Ling et al. [10] consists of multi-choice word problems covering a broad range of topics and difficulty levels. This dataset also contains descriptions of the rationale to reach the correct answer.

MATH is a dataset consisting of challenging competition mathematic problems with step-by-step natural language solutions introduced by Hendrycks et al. [6]. The

problems were retrieved from United States' mathematics competitions. These problems are designed to be challenging for humans and often require more than just basic application of mathematic tools.

GSM8K released by Cobbe et al. [5] consists of multistep middle school word problems with natural language solutions. These problems take between 2 and 8 steps to solve. A bright student should be able to solve all of them.

ASDiv is a mathematical word problem dataset with a strong emphasis on great diversity. This dataset of arithmetic and algebraic problems was introduced by Miao, Liang, and Su [12].

SVAMP introduced by Patel, Bhattacharya, and Goyal [15] contains many variations of elementary school mathematical word problems.

MGSM is a multilingual dataset introduced by Shi et al. [21] containing 250 manually translated grade-school problems from the *GSM8K*.

Dataset	Size	Answer	Difficulty	Language
MultiArith	600	symbolic	elementary school	English
Math23K	23 161	symbolic	elementary school	Chinese
AQuA	100 949	multi-choice	diverse	English
MATH	12 500	natural language	competitions	English
GSM8K	8 500	natural language	elementary school	English
ASDiv	2 305	symbolic	elementary school	English
SVAMP	1 000	symbolic	elementary school	English
MGSM	250	natural language	elementary school	multilingual

Figure 1.1: Comparison of existing datasets

1.4 Evaluating LLM Answers

There are few ways to evaluate answers generated by LLMs. The method used varies depending on the type of question. We will only focus on problems that have known solutions that can be used to verify the model's answers.

For questions with concrete numerical answers, the most straightforward approach is to compare the LLM's numerical answer to the correct one. However, LLMs usually generate natural language output. This usually involves extracting the number or equation from the model's output. Such an approach is very precise, as shown by Hendrycks et al. [6], Wang, Liu, and Shi [22] and Sawada et al. [19].

LLMs can also be evaluated on multiple-choice problems. This is done by providing the model with the options and prompting it to select one of the provided options [21]. With the right prompting, models can output answers that can be extracted in more than 97% of the time [19]. Alternatively, the model is not provided with the options,

but its answer is extracted as a number and then compared to the available options as in Amini et al. [2].

The most problematic questions are those which have open answers that cannot be easily extracted from the model’s output. Unfortunately, these are the types of questions that we are most interested in. One of the approaches is to try and convert these questions into ones that allow automatic answer extraction. This is done by extracting a number or equation from the reference solution and trying to match it with the model’s output. Other methods involve changing the question into a multiple-choice one. Both of those approaches, while valid, do not fully evaluate the model’s capabilities to correctly solve problems that require natural language solutions.

The most straightforward approach to evaluating such problems is using a human evaluator. Nonetheless, this approach is highly labor-intensive and ineffective. We can leverage the existing model’s capabilities and use it instead of a human to evaluate the answers. Such an approach is called model-based evaluation [19].

In its simplest form, the model is provided by the reference solution and the output that should be evaluated. It is then asked to grade the provided output. Prior research indicates that such an approach is possible, but the evaluations are not reliable enough to be used alone [9] [20]. Some researchers went so far as to avoid providing the model with the reference solution. Those experiments provided promising results, but they still failed to be reliable enough [8].

An improved approach was introduced by asking the model to generate evaluation rubrics, and then using those rubrics to evaluate the solutions [19]. The model is provided with the reference solution and generates rubrics and allocates points to them. It was shown by Sawada et al. [19] that GPT-4 designs rubrics that cover most of the solution steps correctly, but sometimes fail to properly allocate points based on their importance. The model is quite reliable on assigning the correct number of points to solutions based on the generated rubrics. However, the model cannot score solutions that do not follow the generated rubrics, but are otherwise correct. Another issue with this approach is that the model attempts to assign points to attempted solutions that are outside the generated rubrics. A human evaluator would score these solutions with zero points [19]. Knowing its limitations, we will base our approach on this method to evaluate models on our dataset.

1.5 Prior Research

The MGSM paper by Shi et al. [21] tries to evaluate models’ reasoning abilities in multiple languages. They achieve this by manually translating 250 problems from GSM8K [5] into ten typologically diverse languages, which they then used to benchmark

GPT-3 on their dataset. Their research reveals that results are very similar, with insignificant differences between the various languages. It has been demonstrated that using an intermediate English chain-of-thought provides results that are on par with or better than answers written in the question’s original language.

Another multilingual research by Ahuja et al. [1] comprehensively evaluates the models on various multilingual datasets. Even though this research does not evaluate the advanced reasoning abilities, it demonstrates the overall capabilities of large language models to reason in languages other than English. Their results show no significant differences between the results achieved in the different languages.

1.6 Related Benchmarks

In this section, we examine existing benchmarks and findings related to advanced reasoning abilities. We aim to establish a baseline against which we can compare our own findings. This comparison will not only help in highlighting any unique contributions from our research, but also in identifying potential gaps and opportunities for further investigation.

Experiments conducted by OpenAI et al. [14] indicate that when utilizing a few-shot prompting, GPT-3.5-Turbo achieves a score of 57.1% on the GSM8K benchmark. In contrast, the more advanced GPT-4 demonstrates significantly improved performance, reaching a score of 92.0% on the same benchmark. Similar results for GPT-4 were also published by Bubeck et al. [4], who scored 87.1%. The GSM8K benchmark shares similarities with our dataset in that it features answers composed in natural language. However, the benchmark’s difficulty is at an elementary school level, while our dataset is targeted for high school students.

In their research, Sawada et al. [19] employed a model-based evaluation technique to determine the proficiency of GPT-4 in solving academic challenges, specifically within the realms of mathematics and physics. The dataset used is similar to ours, featuring problems typically encountered in mathematical competitions. The results show that GPT-4 achieved a score of 50.5% in physics problems with symbolic answers and 33.7% in similar mathematics problems. Additionally, when faced with problems resembling formal proofs, GPT-4 managed to secure a score of 38%.

Experiments conducted by Bubeck et al. [4] using the GPT-4 have demonstrated proficiency in solving mathematical problems, scoring 42.5% on the MATH benchmark. GPT-3.5 This benchmark is particularly significant as it consists of natural language solutions and competition problems that are similar to those found in our dataset.

DRAFT

Chapter 2

Our Goals

Our research aims to explore the extent to which advanced reasoning capabilities exhibited by large language models can be observed in the Slovak language. To achieve this goal, we intend to compare our findings against existing research conducted in the English language. By using the techniques established in prior research, we seek to investigate whether similar patterns of reasoning can be seen when large language models are working with Slovak problems.

We will also experiment with multiple prompting techniques to compare their effectiveness in the Slovak language. By doing this, our research aims to provide insights into the ability to generalize results of large language models across different languages.

To accomplish this, we will firstly introduce a new dataset consisting of various tasks aimed at advanced reasoning in maths, physics and informatics in the Slovak language from local high school competitions.

We then establish a framework for evaluating large language models using our dataset. We proceed to benchmark multiple large language models on our dataset to evaluate their performance in tasks requiring advanced reasoning abilities. Through experimentation, we test different prompting techniques, aiming to replicate findings observed in prior research conducted in English.

We aim to answer the following questions:

- What is the overall performance of large language models on our dataset?
- Which prompting techniques are effective in the Slovak language, and to what extent?
- How do our results compare to those published by English researchers?
- How well do models perform when given English problems compared to Slovak problems?

DRAFT

Bibliography

- [1] Kabir Ahuja et al. *MEGA: Multilingual Evaluation of Generative AI*. 2023. arXiv: 2303.12528 [cs.CL].
- [2] Aida Amini et al. “MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms”. In: *CoRR* abs/1905.13319 (2019). arXiv: 1905.13319. URL: <http://arxiv.org/abs/1905.13319>.
- [3] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].
- [4] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712 [cs.CL].
- [5] Karl Cobbe et al. *Training Verifiers to Solve Math Word Problems*. 2021. arXiv: 2110.14168 [cs.LG].
- [6] Dan Hendrycks et al. *Measuring Mathematical Problem Solving With the MATH Dataset*. 2021. arXiv: 2103.03874 [cs.LG].
- [7] Takeshi Kojima et al. *Large Language Models are Zero-Shot Reasoners*. 2023. arXiv: 2205.11916 [cs.CL].
- [8] Gerd Kortemeyer. *Performance of the Pre-Trained Large Language Model GPT-4 on Automated Short Answer Grading*. 2023. arXiv: 2309.09338 [cs.CL].
- [9] Gerd Kortemeyer. “Toward AI grading of student problem solutions in introductory physics: A feasibility study”. In: *Physical Review Physics Education Research* 19.2 (Nov. 2023). ISSN: 2469-9896. DOI: 10.1103/physrevphyseducres.19.020163. URL: <http://dx.doi.org/10.1103/PhysRevPhysEducRes.19.020163>.
- [10] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. “Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 158–167. DOI: 10.18653/v1/P17-1015. URL: <https://aclanthology.org/P17-1015>.

- [11] Jiacheng Liu et al. *Generated Knowledge Prompting for Commonsense Reasoning*. 2022. arXiv: 2110.08387 [cs.CL].
- [12] Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. *A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers*. 2021. arXiv: 2106.15772 [cs.AI].
- [13] OpenAI. *Models*. <https://platform.openai.com/docs/models/gpt-3-5-turbo>. [Accessed 09-05-2024].
- [14] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL].
- [15] Arkil Patel, Satwik Bhattacharya, and Navin Goyal. *Are NLP Models really able to Solve Simple Math Word Problems?* 2021. arXiv: 2103.07191 [cs.CL].
- [16] Chengwei Qin et al. *Is ChatGPT a General-Purpose Natural Language Processing Task Solver?* 2023. arXiv: 2302.06476 [cs.CL].
- [17] Alec Radford et al. “Language Models are Unsupervised Multitask Learners”. In: 2019. URL: <https://api.semanticscholar.org/CorpusID:160025533>.
- [18] Subhro Roy and Dan Roth. *Solving General Arithmetic Word Problems*. 2016. arXiv: 1608.01413 [cs.CL].
- [19] Tomohiro Sawada et al. *ARB: Advanced Reasoning Benchmark for Large Language Models*. 2023. arXiv: 2307.13692 [cs.CL].
- [20] Johannes Schneider, Bernd Schenk, Christina Niklaus, and Michaelis Vlachos. *Towards LLM-based Autograding for Short Textual Answers*. 2023. arXiv: 2309.11508 [cs.CL].
- [21] Freda Shi et al. *Language Models are Multilingual Chain-of-Thought Reasoners*. 2022. arXiv: 2210.03057 [cs.CL].
- [22] Yan Wang, Xiaojiang Liu, and Shuming Shi. “Deep Neural Solver for Math Word Problems”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 845–854. DOI: 10.18653/v1/D17-1088. URL: <https://aclanthology.org/D17-1088>.
- [23] Jason Wei et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL].
- [24] Denny Zhou et al. *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*. 2023. arXiv: 2205.10625 [cs.AI].