

Mathematical Capabilities of ChatGPT

Simon Frieder^{*1,4}, Luca Pinchetti¹, Ryan-Rhys Griffiths³, Tommaso Salvatori², Thomas Lukasiewicz^{1,2}, Philipp Christian Petersen^{4,5}, Alexis Chevalier⁶, and Julius Berner⁴

¹Department of Computer Science, University of Oxford, Oxford, UK

²Institute of Logic and Computation, TU Wien, Vienna, Austria

³Department of Physics, University of Cambridge, Cambridge, UK

⁴Faculty of Mathematics, University of Vienna, Vienna, Austria

⁵Research Network Data Science, University of Vienna, Vienna, Austria

⁶School of Mathematics, Institute for Advanced Study, Princeton, US

February 1, 2023

Abstract

We investigate the mathematical capabilities of ChatGPT by testing it on publicly available datasets, as well as hand-crafted ones, and measuring its performance against other models trained on a mathematical corpus, such as Minerva. We also test whether ChatGPT can be a useful assistant to professional mathematicians by emulating various use cases that come up in the daily professional activities of mathematicians (question answering, theorem searching). In contrast to formal mathematics, where large databases of formal proofs are available (e.g., the Lean Mathematical Library), current datasets of natural-language mathematics, used to benchmark language models, only cover elementary mathematics. We address this issue by introducing a new dataset: GHOSTS. It is the first natural-language dataset made and curated by working researchers in mathematics that (1) **aims to cover graduate-level mathematics** and (2) **provides a holistic overview of the mathematical capabilities of language models**. We benchmark ChatGPT on GHOSTS and evaluate performance against fine-grained criteria. We make this new dataset publicly available¹ to assist a community-driven comparison of ChatGPT with (future) large language models in terms of advanced mathematical comprehension. We conclude that contrary to many positive reports in the media (a potential case of selection bias), **ChatGPT’s mathematical abilities are significantly below those of an average mathematics graduate student**. Our results show that ChatGPT often understands the question but fails to provide correct solutions. Hence, if your goal is to use it to pass a university exam, you would be better off copying from your average peer!

1 Introduction

Since its introduction, ChatGPT has rapidly become a widely known question-and-answer dialogue system. It has been mentioned in traditional media across the globe [33, 28, 22] and across all major internet platforms [40, 43]. According to Twitter data, it is by far the most talked about language model to date; cf. Figure 1.

The performance of ChatGPT has been analyzed in a large number of exam-related use cases, with varying degrees of scientific rigor, ranging from detailed studies to anecdotal evidence. Use cases include passing the United States Medical Licensing Examination [17], scoring highly on the *Psychology Today* Verbal-Linguistic Intelligence IQ Test [34], and answering (and generating) Operations Management exam questions that were

^{*}Corresponding author: simon.frieder@wolfson.ox.ac.uk. The subsequent author list is ordered randomly.

¹github.com/friederr/science-GHOSTS

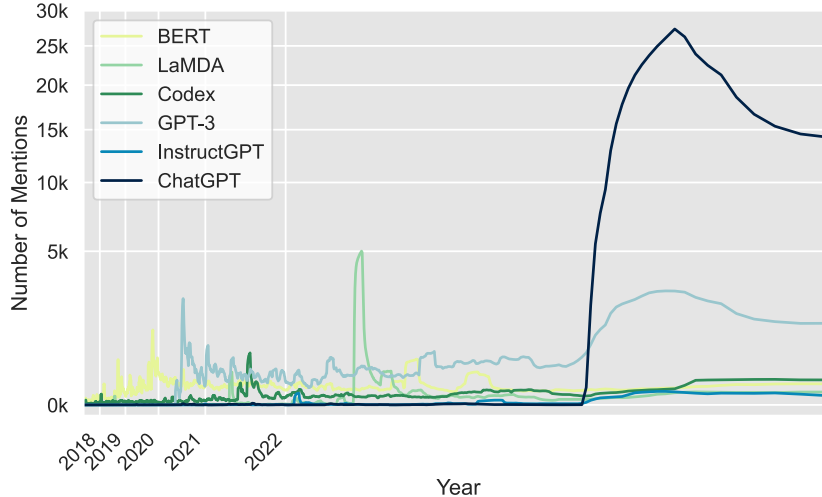


Figure 1: Twitter count data relating the counts of a selection of notable large language models from the beginning of the release date of GPT-3. The x -axis is a log-scaled. ChatGPT counts by far dominate those of all other language models. Vertical year-ticks denote the end of the mentioned year.

deemed to be within the scope of a typical MBA curriculum [41], all with a performance that elicited positive surprise of the authors. Due to this and other reasons, it is widely believed that large language models (LLMs) will impact a large number of areas and will be used as assistants by many professionals.

In this article, we will focus on performing a detailed analysis of the mathematical capabilities of ChatGPT. This includes, but is not limited to, answering exam-style mathematical questions and investigating how ChatGPT behaves in a number of mathematical contexts. Our analysis includes testing how many of the skills ChatGPT can emulate that are necessary to do professional mathematics. Examples of such skills are the ability to answer computational questions (“What is the value of $\int_0^{\frac{\pi}{2}} \arccos(\frac{\cos x}{1+2\cos x}) dx$?”), the ability to complete mathematical proofs that have gaps or missing steps, the ability to solve questions that are more focused on deep insights and original solutions, such as those of mathematical Olympiads, and the ability to survey the literature and think across domains (“which other theorems do we need to prove a given theorem?”).

To do this, we have designed a thorough testing methodology to evaluate the outputs of ChatGPT, including error codes that represent various possible failure modes of ChatGPT (see Section 3). We score ChatGPT’s responses, we report on the results using this methodology, and we compare ChatGPT to state-of-the-art models trained for mathematical comprehension.

Moreover, we have created new datasets of prompts that are aimed at testing specific aspects of ChatGPT related to mathematical comprehension. We evaluate ChatGPT by comparing it to random samples from existing datasets that were devised to test models that were specifically trained for mathematical comprehension [18, 16]. A number of the datasets are specifically designed so that the questions could not be answered if ChatGPT were memorizing the results. All of those datasets are created by the authors.

In summary, the contributions of this article are threefold:

- First, insight for mathematical use is provided. We show for which types of questions and which domains of mathematics, ChatGPT may be useful and how it could be integrated into the workflow of a mathematician.
- Second, the failure modes of ChatGPT are identified, as well as the limits of its capabilities. This can aid future efforts to develop LLMs that perform better in mathematics. Our analysis is akin to

a *mathematical model card*, where the mathematical strengths and weaknesses are summarized (see Section 4).

- Third, we provide benchmarks for testing the mathematical capabilities of future LLMs so that they can be compared to ChatGPT across a range of aspects regarding advanced mathematical comprehension. This is achieved by introducing new natural-language math datasets. Two of these benchmarks are derived from the most advanced datasets regarding mathematical queries for language models that exist today. Additionally, we devise four more datasets on which we benchmark ChatGPTs performance. We release the collection of these datasets publicly on GitHub², and we encourage community participation by allowing GitHub pull requests in order to grow the datasets beyond their current sizes.

2 Related Work

As a large language model, ChatGPT can be universally employed to perform mathematical reasoning and therefore has to compare with technologies that in this space are sometimes decades old. Performing mathematical reasoning in an automated way has a long history and can be traced back to 1959 [37], the most focus being devoted to proving theorems [11]. Presently, there is a realization that the classical approaches, using a symbolic encoding of mathematics, have reached a plateau [14].

There is now a growing body of literature on learning mathematical relationships directly in a supervised-learning manner [2, 10, 15] or by using LLMs to perform mathematical reasoning directly on mathematics encoded in natural language [20]. Sometimes, the distinction is blurred because Transformers can also be used in a supervised-learning setting and have been employed successfully in learning mathematical relationships [18, 6].

Most recently published large language models, such as PaLM [8], released in 2022, are tested only on elementary-level mathematical reasoning datasets, such as the GSM8K dataset [9]. We speculate this is because the obtained results already suggest that the models struggle on much simpler datasets than ours, such as the MathQA [1] dataset or the GSM8K dataset [9], respectively. For example, the version of PaLM with 540 billion parameters with chain-of-thought prompting and access to an external calculator solves only 58% on the GSM8K dataset [8, Table 10]. This model nonetheless outperforms GPT-3 [5] on the same dataset, which only solves at best 54%; this performance is consistent with the performance of older models. Variations of BERT [30] have been shown to only solve between 28% and 37% of the problems when fine-tuned and tested on the AQUA-RAT dataset [21], which is the direct predecessor of MathQA. In some cases, such as the LaMDA model [42] or BLOOM [19], both released also in 2022 by Google, an evaluation of mathematical reasoning capability is missing entirely.

Among the mentioned LLMs, Minerva [20], based on PaLM, stands out, being trained in equal parts on websites that contain MathJax elements and arXiv preprints (and on general natural language data on which PaLM was trained), achieving a score of roughly 50% on a significantly harder dataset, the MATH (Mathematics Aptitude Test of Heuristics) dataset [16] that was sourced from various mathematical competitions.

One distinguishing feature of the MATH dataset [16] is that its problems admit 1) a unique answer (no open-ended questions) 2) and the answer can be condensed within a few characters (a number, for example). This is beneficial in terms of automatic evaluation of a model on such a dataset since one can simply ask for the final answer, ignoring the step-by-step solution (moreover, one can train models, as [16] do, to fit this style of inquiry and output either the final solution only, or the step-by-step derivation leading to the solution).

Among the supervised approaches, we mention [18], where a Transformer architecture was used to generate symbolic solutions to integrating functions and finding closed-form solutions to first-order and second-order differential equations, which outperformed classical solvers, such as Mathematica, MATLAB, and Maple

²github.com/friederr/science-GHOSTS

by at least 14% on a test set of integration problems. On the task of solving differential equations, the Transformer-based approach still exceeds the classical approach, but by a smaller margin (at least 4% in the case of first-order differential equations and with more varied results for second-order equations). An up-to-date survey on mathematical datasets and performance of various LLMs can be found in [23].

For ChatGPT, most investigations related to mathematical reasoning consist to date of anecdotal evidence concerning its performance and its failure modes; see, e.g., [43, 32, 24, 40]. Unfortunately, a clear methodology is missing, as most of the results are scattered on various internet platforms and are not easily reproducible. To the best of our knowledge, the only mathematical investigation was undertaken in [4], which mainly investigated ChatGPT’s capability to compute irrational numbers to high accuracy.

On the other hand, we like to mention the case of *formalized* mathematics, where large databases that encode advanced mathematical concepts exist, e.g. the Lean Mathematical Library [25]. Some of the ideas that we have used in this article, such as prompting with missing proofs, are echoed in [31] for formal mathematics. Yet, for the purpose of doing mathematics with large language models, these formal datasets cannot be leveraged since no straightforward way exists to convert them to natural language (in addition to various issues, such as bias, that might occur in the context of an automatic conversion).

3 Datasets

3.1 Dataset creation

We assess the mathematical reasoning capabilities of ChatGPT by creating a collection of multiple datasets of prompts, totaling 728 prompts, for which ChatGPT’s output was manually rated by experts. Then, we record and rate each of the outputs provided by the model. The combined effort of devising mathematically insightful prompts, some of which are at graduate-level mathematics, and carefully rating the the output of ChatGPT amount to several hundreds of person-hours.

We divide our entire collection of prompts into six *subdatasets*³, called

- ***G**rad-**T**ext*
- ***H**oles-in-**P**roofs*
- ***O**lympiad-**P**roblem-**S**olving*
- ***S**ymbolic-**I**ntegration*
- ***MATH***
- ***S**earch-**E**ngine-**A**spects*

We summarize those in Table 1. The letters that are set in boldface make up the GHOSTS acronym.

Two of the subdatasets, the *MATH* subdataset and the *Symbolic-Integration* subdataset, use prompts taken from existing datasets, [16] and [18], respectively. This was done in order to be able to compare how ChatGPT performs against existing state-of-the-art models, one based on an LLM, Minvera [20], and one based on a supervised-learning approach [18]. Nonetheless, significant, additional annotation effort was involved since in both cases the authors, as experts in the field, rated the output. Furthermore, in the second case, a conversion from Polish notation was necessary.

The other subdatasets were hand-crafted by the authors. We note that it is neither possible to outsource the creation of these datasets to a crowdsourcing service, such as Amazon Mechanical Turk, nor is it possible to generate these datasets automatically from code because advanced mathematical insight is required for the creation of each prompt (though based on our work, it might be possible to extend the dataset by creating variations of our questions in a purely programmatic manner; see Section 5). Furthermore, unlike in the case

³In the GitHub repository, each subdataset corresponds to a folder, which in turn can consist of multiple files.

Dataset name	Comprised of the file(s)	Tags
<i>Grad-Text</i>	W. Rudin, Functional Analysis (ch. 1)	M3 Q4
	W. Rudin, Functional Analysis (ch. 2)	M3 Q4
	J. Munkres, Topology (ch. 1)	M3 Q4
	J. Munkres, Topology (ch. 2)	M3 Q4
	R. Durrett, Probability Theory	M3 Q4
<i>Holes-in-Proofs</i>	Proofs Collection A	M3 Q2 Q5
	Proofs Collection B Prealgebra	M1 Q5
	Proofs Collection B Precalculus	M1 Q5
<i>Olympiad-Problem-Solving</i>	Olympiad Problem Solving	M4 Q4 D2
<i>Symbolic-Integration</i>	Symbolic Integration	M2 Q3 D1
<i>MATH</i>	MATH Algebra	M1 M2 M3 Q3 Q4
	MATH Counting and Probability	M1 M2 M3 Q3 Q4
	MATH Prealgebra	M1 Q3 Q4
	MATH Precalculus	M1 Q3 Q4
<i>Search-Engine-Aspects</i>	Definition Retrieval	M3 Q1 Q2 D3
	Reverse Definition Retrieval	M3 Q2 D3
	Named Theorem Proof Completion	M3 Q1 Q2 D3

Table 1: A summary of all datasets, together with their associated tags. The tags M_i , Q_i , and D_i relate to the level of Mathematical difficulty, the Question type, and the Out-of-Distribution type from Section 3.3, respectively.

of the MATH dataset by [16] (see Section 2), the output of ChatGPT cannot be automatically evaluated and a professional opinion on its output needs to be given. This raises the difficulty of the creation of more data, since graduate-level mathematics (and in some cases PhD-level mathematics) is required. The mathematical skill level of the authors matches this requirement.

Our dataset goes beyond all the mentioned mathematical datasets for LLMs in Section 2 in terms of mathematical sophistication and in terms of the different aspects of mathematical reasoning that are being tested. It also surpasses datasets mentioned in the survey [23] along these dimensions. Furthermore, unlike for the MATH dataset by [16], a large number of our questions are not of the type of having an answer that can be condensed into a few tokens (e.g., when the answer is a mathematical proof), such as a number or a function, and therefore evaluation cannot be performed automatically.

3.2 Format

The format of each of our subdatasets that make up our GHOSTS dataset follows the same convention and each subdataset consists of (potentially multiple) JSON-formatted files with entries such as in the example below. Our format is similar to, e.g., the AQUA-RAT (Algebra Question Answering with Rationales) dataset [21]. A single datapoint in a JSON file has the following form:

```
{
  "prompt": "Can you quote a famous mathematical theorem?",
  "output": "Euclid's Elements, Proposition 47: In right-angled triangles, the
    square on the side opposite the right angle is equal to the sum of the squares
    on the other two sides.",
  "rating": "5",
  "errorcodes": [],
  "warningcodes": ["w1"],
  "comment": "This is actually Pythagoras' theorem, which was not mentioned, hence
    only a warning flag is raised.",
  "ref" : ""
}
```

```

"msc": "51M04",
"confidence": "high",
"timestamp": "2023-01-30"
}

```

An example where ChatGPT was less successful is the following:

```

{
  "prompt": "Suppose X is a vector space. Prove that  $0x = 0 = a0$  if  $x$  in  $X$  and  $a$  is a scalar.",
  "output": "To prove that  $0x = 0$  for any  $x$  in  $X$ , we use the fact that  $0$  is the additive identity of the vector space  $X$ , meaning that for any  $x$  in  $X$ ,  $0 + x = x$ . Thus, if we multiply  $0$  by any scalar  $a$ , we have  $a(0 + x) = ax = a0 + ax = a0$ , which shows that  $a0 = 0$ .  
Similarly, to prove that  $a0 = 0$  for any scalar  $a$ , we use the fact that  $0$  is the multiplicative identity of the scalar field, meaning that for any scalar  $a$ ,  $a \cdot 1 = a$ . Thus, if we multiply  $0$  by any scalar  $a$ , we have  $a \cdot 0 = a \cdot 1 \cdot 0 = a \cdot 0 = a0$ , which shows that  $a0 = 0$ .",
  "rating": "3",
  "errorcodes": ["e5_2", "e5_4"],
  "warningcodes": [],
  "comment": "part of the result is shown, but the argument for the second half is nonsense. Also makes wrong statements.",
  "msc": "46A03",
  "ref": "Rudin-Functional Analysis-Second-Ed.Part1-ex1/b-page38",
  "confidence": "high",
  "timestamp": "2023-01-18"
}

```

We explain each of these fields; unless explicitly mentioned below:

- **prompt** denotes the input that we supply to ChatGPT via its web interface. We always use a new session for each prompt to avoid biasing ChatGPT by using previous prompts.
- **output** denotes the raw output that ChatGPT supplies us with;
- **rating** is a number from 1 to 5 that shows how many points ChatGPT has scored, 5 being a perfect answer and 1 being the lowest. A detailed explanation regarding the rating policy that we followed is contained in the appendix, Section A.
- **errorcodes** highlight a list of error types that illustrate the failure mode of ChatGPT in a more fine-grained way. Not all types of errors apply to all (sub)datasets: For example, an error code for a missing proof step would not be applicable on a dataset that tests whether ChatGPT can multiply numbers or find prime divisors. This field can be empty if no error code applies. The detailed explanation of the error codes (and the warning codes; see below) that was provided to the annotators is contained in the appendix, Section B.
- **warningcodes** highlight any problematic aspects of ChatGPT; for example, ChatGPT might be rambling and providing the user with unrelated information or use a poor (but correct) way of solving problems. This field can be empty if no warning code applies.
- **comment** denotes any noteworthy commentary that an assessor of ChatGPTs may make. This can be related to giving a more detailed explanation of output (or its failure modes), providing reasoning behind awarding a certain error code, generally providing context, etc. For some subdatasets (see Section 3.3), this field was used to indicate the difficulty level of the prompt, as well as an official solution, if available. This field is not required to always have a value.

- **msc** denotes the math subject classification⁴ (MSC) that pertains to the output, not the prompt that one gives to ChatGPT. This is because the prompt, unlike the output, might not really have a classification, for example, if ChatGPT is asked what the most important theorem in all of math is⁵.
- **ref** indicates a reference to where the prompt was originally taken from (for some subdatasets, such as *Holes-in-Proofs*, we have changed proofs from various books or `math.stackexchange.com`; the original source was recorded in this field). This field can be empty if the question was formulated by the authors and no authoritative source was plausible.
- **confidence** indicates how confident we have perceived ChatGPT to be when presenting us with its output. We allow values of “high”, “medium”, and “low”.
- **timestamp** denotes when the prompt was entered into ChatGPT.

Each subdataset is made up of multiple such data points. In cases where we used prompt engineering (the *Olympiad-Problem-Solving* dataset), and asked variations of a single question, we enclosed multiple such data points in an array within the JSON file.

The fields within a single data point interact in nontrivial ways: If a rating of 5 is given, then it is expected that no error code is present—though there may be warning codes that are used. The error codes and warning codes are loosely in the spirit of a compiler throwing errors and warnings if it is given incorrect or sloppy code—although we have a role reversal, where the human is now the compiler and the machine produced the code. In this sense, for some prompts, we have used multiple error and/or warning codes, which is why these fields are arrays of strings. We use these codes to collect statistics on the behavior of ChatGPT; see Section 4.

The usage of MSC codes can be useful for mathematicians who want to integrate ChatGPT in their daily workflow, as it allows them to know in which areas the model performs better and can hence be trusted more. Our dataset is very diverse, as it has a total of 41 MSC codes. The top short version of these codes (first two digits) is 26 (“Real functions”, 125 occurrences) followed by 05 (“Combinatorics”, 124 occurrences) and 60 (“Probability theory and stochastic processes”, 100 occurrences). An exhaustive survey of ChatGPT’s performance across *every* MSC code would necessitate a large, community-driven effort to set up a very large database. Because of the high cost of rating each output, which requires specialized skills, this is something that no individual research group could reasonably do—but we hope that our approach is a starting point for such an effort.

For end-users of ChatGPT, it is desirable to avoid having a long-winded dialogue to arrive at a solution. Therefore, we require that ChatGPT gives us the correct solution by providing it only the input without any subsequent clarification. All chats with ChatGPT are thus “cold”. But we do allow the possibility of prompt engineering of the input, where more information is added beyond the core prompt content.

3.3 The subdatasets

For most of our subdatasets, we have used \LaTeX to encode mathematical input. Our experiments have shown that ChatGPT can process \LaTeX -encoded mathematics well. For example, on the *Holes-in-Proofs* dataset, except for one case, the output of the prompts by ChatGPT was valid \LaTeX source code and could be rendered.

The *Grad-Text* subdataset consists of a collection of books that are used widely in universities to teach upper undergraduate or first-year graduate courses in a degree in mathematics. We have used as prompt and rated the output of most of the exercises from the first and second chapters of these books (except for the book [12] of which we only used exercises from the (quite long) first chapter).

The *Olympiad-Problem-Solving* subdataset consists of a selection of exercises from the book *Problem-Solving Strategies*, [13], that is often used when preparing for mathematical competitions. We selected and graded

⁴A complete list of MSC codes can be accessed under the URL zbmath.org/static/msc2020.pdf.

⁵The answer is Pythagoras’ theorem, according to ChatGPT.

ChatGPT output on one hundred exercises drawn from all chapters.

The *Holes-in-Proofs* subdataset consists of a number of proofs drawn from `math.stackexchange.com`, a collection of books [3, 35], and the MATH dataset [16] where intentionally parts of the proofs were deleted and ChatGPT was prompted to fill the gaps: This was done either by (1) using a MISSING token, (2) finishing the proof early and prompting ChatGPT to complete it, or (3) explicitly asking for certain conditions or results.

The *MATH* subdataset consists of a random sample of prompts from the MATH dataset [16]. The MATH dataset by [16] comes with a level of difficulty for each problem. We focused our random samples on two domains, Algebra and Probability Theory, but sampled the same number of problems at each level of difficulty.

The *Symbolic-Integration* subdataset consists of random samples of integrals that were in the test set of [18]. We converted these from Polish notation to L^AT_EX before prompting ChatGPT. The assessment was done by comparing it with a classical symbolic integration system, Mathematica.

The *Search-Engine-Aspects* subdataset consists of problems that were not sampled from a particular source and generated by a human expert in the field. In the file *Named Theorem Proof Completion* we focused on prompting ChatGPT to provide proof outlines of various theorems that are sufficiently well-known within Functional Analysis to have names. In the *Definition Retrieval* file, we prompted ChatGPT to state correctly various definitions centered around Functional Analysis and Topology. In contrast, in the *Reverse Definition Retrieval* file, we verified whether ChatGPT was able to deduce the name of a mathematical object by describing its properties.

Our subdatasets can be categorized along multiple dimensions, which we state below:

Mathematical difficulty (ascending)

1. elementary arithmetic problems, as found in the MATH dataset [16] at lower levels of difficulty;
2. symbolic problems (integration of functions) that can be also solved via a supervised-learning, data-driven approach to mathematics [18];
3. (under)graduate-level exercises from well-known textbooks [35, 36, 26, 12, 3] as well as questions from `math.stackexchange.com`, spanning diverse domains of mathematics;
4. exercises that are in the style of mathematical olympiad problems, such as those taken from Engel’s *Problem-Solving Strategies* book [13];

Question type

1. Review questions, which ask to state or name correctly certain mathematical facts (the *Definition Retrieval* file of the *Search-Engine-Aspects* subdataset);
2. Overview-type review questions, which cut through an entire field of mathematics (the *Named Proof Completion* and *Reverse Definition Retrieval* parts of the *Search-Engine-Aspects* subdataset, as well the *Holes-in-Proofs* subdataset);
3. Computational questions (the *Symbolic Integration* subdataset as well as various prompts from the *MATH* dataset);
4. Proof-based questions, which ask for a theorem proof or for a puzzle solution (The *Olympiad-Problem-Solving* subdataset, the *Grad-Text* subdataset);
5. Proof-completion questions, which ask for a proof that has missing gaps, or is incomplete, to be completed (the *Holes-in-Proofs* subdataset as well as various prompts from the MATH dataset)

Types of high out-of-distribution likelihood

1. Nontrivial problem encoding: The data points from the *Symbolic Integration* subdataset come from [18] and are publicly available⁶. Since the online training set uses Polish notation, it is very unlikely that ChatGPT has seen these exact prompts before;
2. Succinct solution: The solutions for the *Olympiad-Problem-Solving* subdataset are included in the book by Engel [13]. But the solutions are extremely concise, and simply repeating them would not show an immediate understanding of the problem;
3. Spoken dialogue: The *Search-Engine-Aspects* subdataset is unlikely to be well represented in the data on which ChatGPT has been trained since its prompts resemble word fragments that might appear in a mathematical dialogue (e.g., an oral mathematical exam), rather than in a textbook;

One could, in theory, start to investigate every possible combination of these attributes (e.g., for elementary arithmetic problems, in a non-trivial encoding, one could generate data to cover every possible question type listed above). This would lead to 60 (sub)datasets, which, because of the manual, skilled curation effort, is by far too much for a single research group. Hence, we will allow pull requests in order to encourage the community to contribute and grow these datasets, so that they can be used as a useful benchmark for other LLMs. We have nonetheless striven in this work to cover each of these aspects individually, in some subdataset, as can be seen in Table 1. Investigating every possible combination of attributes with a separate (sub)dataset is not feasible.

Existing datasets do not cover all possible combinations of possibilities across all these dimensions. Devising further datasets to do so is not feasible. Instead, we will devise a specific, well-crafted dataset to cover a number of illustrative cases, which we describe below, so that we cover all these dimensions individually. We collect statistics for ChatGPT’s output, including output length, the stability of the answer under prompt engineering, as well as a personal rating of how close we perceived ChatGPT to be to the correct answer.

Because input to ChatGPT is purely textual, certain types of questions that have a strong geometrical flavor and might be stated and solved in non-text-based fashion (e.g., questions involving diagrams or small algorithms, as occasionally occur in [13]) have been excluded.

4 Results

Will ChatGPT get you through a university math class? No, you would be better off copying from your average peer! In this section, we analyze why by going through the common mistakes performed by ChatGPT one dataset at a time.

Grad-Text ChatGPT performed best on simple set-theory and logic questions (the first chapter from the book *Topology* by J. Munkres ([26])), which is reflected in its rating; see Figure 2. On the rest of the books, it performed substantially worse. We note that it never failed to understand a query; the lowest grade it received is a 2. Because of the confidence (**high**) with which it outputs the answer, the use of ChatGPT is particularly deceiving in this use-case, since it may be intensively used by students studying these subjects.

Olympiad-Problem-Solving On this subdataset, ChatGPT performed poorly. Extra points were awarded when the answer started to show promise, giving a score of 3, but most scores are 2 because the answer does not show promise. No rating of 5 was awarded, and only two ratings of 4 were achieved. ChatGPT had a tendency to try and solve many questions using induction arguments. While this is not obviously false, this was very far from the solutions given in the book. ChatGPT’s inductive proofs were easily seen to contain mistakes. ChatGPT often struggled to understand unusual puzzles and strange situations. For example, on the questions based on changing the colour of the squares on a chess board, the solution offered

⁶github.com/facebookresearch/SymbolicMathematics

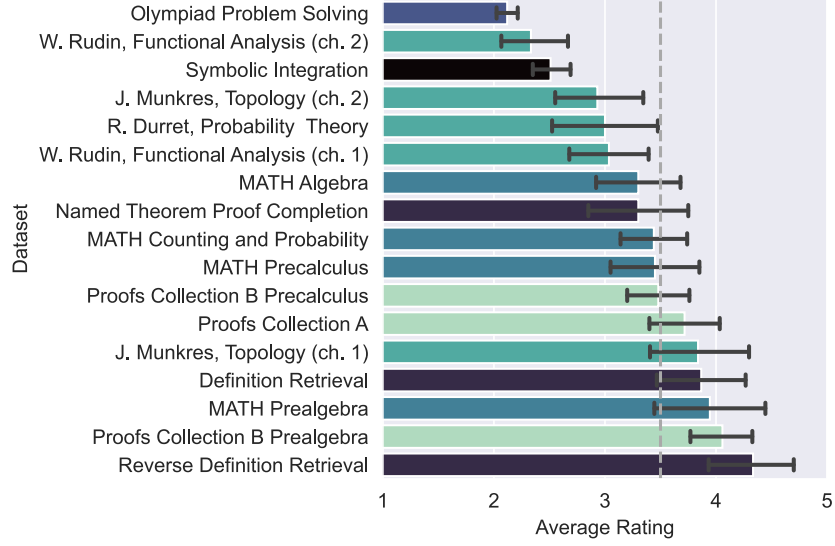


Figure 2: Average rating for each file in each subdataset (same colors indicate the files come from the same subdataset). Since the maximal ranking is 5, and the minimal ranking where the question was at least understood is 2, a passing grade (50% of points) would be 3.5, as indicated by the dotted line. The error bars represent 95% confidence intervals.

by ChatGPT obviously did not cover an 8×8 chessboard. Sometimes, ChatGPT would try to solve the problem by changing only 5 squares, far from the 32 required. Similarly, ChatGPT struggled to respect the unusual constraints given in the questions. This gave rise to lots of **e6** errors. The solutions offered by ChatGPT obviously did not respect the rules set out in the question and show that ChatGPT struggles to integrate surprising information into its answers. In some cases where the problem seemed to require complicated mathematics but was actually solvable by elementary techniques, ChatGPT did not spot this but instead referred to the general theory of, e.g., diophantine equations. ChatGPT would often say, e.g., that the question could be solved with these means but that this was hard, so the confidence score was downgraded to **medium** or **low**; this was the only dataset we gave such confidence score. From a mathematical point of view, these questions were also by far the hardest, as they can also pose difficulty to professional mathematicians.

Holes-in-Proofs ChatGPT correctly recognized most well-known results or concepts (e.g., filling the mean-value theorem, given a proof where a reference to this theorem was missing, and ChatGPT needed to fill it in). In only three cases within *Proofs Collection A*, the question was not understood, which resulted in the lowest possible rating of 1. We noted that ChatGPT was very strong at recognizing the context. The notation of the output has always been the one given in the prompt. The ability of ChatGPT to execute algebraic manipulations is surprisingly inconsistent. On many occasions, ChatGPT executes complicated symbolic tasks with ease, and on many occasions, ChatGPT fails on basic arithmetic or basic rearranging. The mistakes do not seem to correlate with the complexity of the algebraic expression. When ChatGPT makes an algebraic mistake, it carries over this mistake reliably to the rest of the computation. We did not spot any cases where the intermediate workings were wrong, but the final solution was correct.

MATH On the questions related to Algebra and Probability theory, ChatGPT got the reasoning often correctly. However, the most common type of error was **e4**: ChatGPT may struggle when confronted with standard operations, such as inverting fractions, least common multiples, changing the sign of numbers when moving them from one side of the equal sign to the other. Often, in these questions, a correct solution requires

performing multiple operations in sequence. In most cases, at least one operation was wrong, preventing the model to get a rating of 5 on the output.

Symbolic-Integration ChatGPT was dominated by systems that were trained specifically to solve integration problems [18]. In a number of instances, ChatGPT got the structure of terms right (for example, the number of summands in the output, as well as where factors had to be placed before summands), but it failed at concrete computations. Even very simple examples were not correct. For example, the antiderivative of $x \mapsto x^2/2$ is evaluated to $x \mapsto x^3/3 + C$, where C is a constant of integration (the correct answer is $x \mapsto x^3/6 + C$). For a number of prompts, ChatGPT claims there is no closed form solution for the integral with complete confidence when in fact there is a solution (see selection of worst-3 examples from Section F).

Named Theorem Proof Completion On this part of the *Search-Engine-Aspects* subdataset, ChatGPT knew almost all the theorems that it was asked at a basic level but made mistakes when stating them. When it came to listing other results required for the proofs, ChatGPT typically requested way more than the necessary theory—occasionally even results that only follow from the theorem which was asked for (error code e5_5).

Definition Retrieval On this part of the *Search-Engine-Aspects* subdataset, ChatGPT had a quite good performance: it recited most definitions correctly. It sometimes got confused when being asked about distributions in the sense of elements of the dual space of test functions. ChatGPT strongly favors the notion of distributions in the stochastic sense. Similarly, for the adjective “closed”, where it chose to pick the context of algebra (instead of topology) and interpreted it to mean “algebraically closed”.

Reverse Definition Retrieval On this part of the *Search-Engine-Aspects* subdataset, ChatGPT had the strongest performance, being able to recover most definitions from their descriptions, with an average rating of 4.3. This indicates the usefulness of ChatGPT as a general-purpose mathematical search engine. This subdataset is also the simplest from a mathematical point of view since no logical thinking is required, but only a name needs to be found.

4.1 Overall Performance

If we take a rating of 3.5 to be the threshold between success and failure, then Figure 2 shows that for most (in particular, harder) problems, ChatGPT will not pass. In particular, on problems that are within the style of mathematical olympiads, ChatGPT performs badly. Moreover, Figure 2 shows that the achieved ratings correspond closely to the ranking of mathematical difficulty of the exercises that a mathematician would assign. We analyze the results for different mathematical fields in Figure 3.

The prompt length has no clear effect on the rating; see Figure 4. This eliminates prompt length as a confounding variable, and in combination with the findings of Figure 2 strongly indicates that ChatGPT’s rating is indeed reflective of mathematical difficulty.

Comparing ChatGPT to the performance obtained by [18], who obtained a percentage of correctly solved integral that was close to 100% on a collection of 500 test equations [18, Table 3], our random sample of their dataset (after translation from Polish notation and conversion to L^AT_EX) achieves an average rating of 2.6, where a rating of 2 is dominating 60% of the time followed by a rating of 4 almost 20% of the time. These scores trail behind the performance that the model trained in [18] achieved.

The situation is similar when comparing ChatGPT to Minerva [20]. Their best model achieved an accuracy of 50% on the MATH dataset [16, Table 3]. On our random sample of the MATH dataset, ChatGPT achieves a perfect score only in 26% of cases.

One interesting finding is related to prompt engineering. On 20% of the questions from *Olympiad-Problem-Solving*, the question was asked a second time, where ChatGPT was additionally instructed to proceed

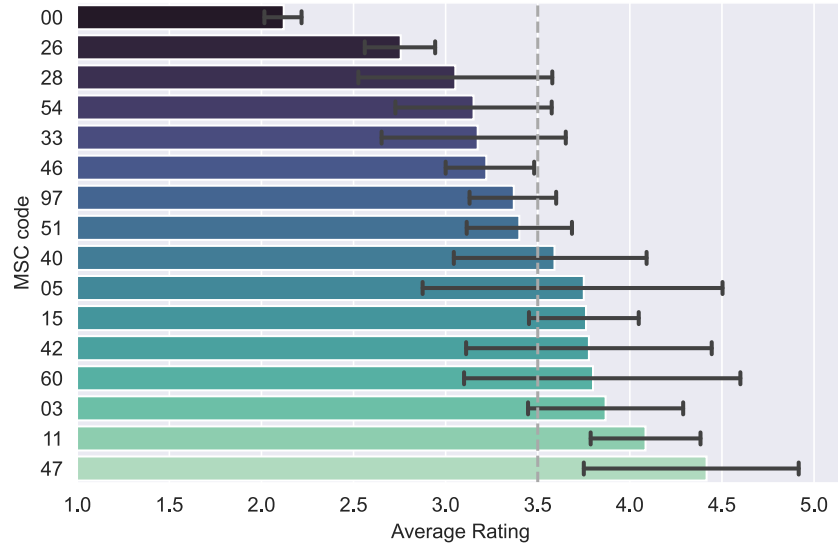


Figure 3: Average rating over mathematical fields (only those MSC codes that have at least 6 occurrences are shown). Note that the ranking is not indicative of the complexity of the fields since we do not use equally complicated exercises for all fields. The error bars represent 95% confidence intervals.

step-by-step. This type of engineering is recommended by OpenAI in their *cookbook*, to improve reliability⁷. This had an interesting effect: Less severe errors e2 went down, though more severe errors (e5-type errors) stayed about the same, as did the average rating, see Figure 5.

ChatGPT is usually very confident, unlike other GPT-like models, such as Codex, that we have experimented with. Consider in this sense the following statement (of statements):

Here are two statements: "Let (X, A, μ) be a measure space. Show that the signed measures $\lambda: A \rightarrow \mathbb{R}$ form a Banach space."; "Let (X, A, μ) be a measure space. Show that the signed measures $\lambda: A \rightarrow \mathbb{R}$ form a Banach space." Are they equivalent?

The response by ChatGPT is not phrased in order to show any nuance in terms of confidence (this is typical, even if ChatGPT is wrong):

Yes, the two statements are equivalent. They both express the same mathematical idea, but use different notation and formatting conventions. The first statement uses parenthesis to enclose the elements of the measure space and regular letters to represent the functions, while the second statement uses subscripted parentheses and uses the LaTeX formatting to represent the elements of the measure space and the function.

The response by Codex [7], another model that we briefly tested (but whose scope would have exceeded that of a single conference article) gives a cautions response and, unlike ChatGPT, is capable of voicing doubt:

I think they are equivalent, but I am not sure.

⁷github.com/openai/openai-cookbook/blob/main/techniques_to_improve_reliability.md

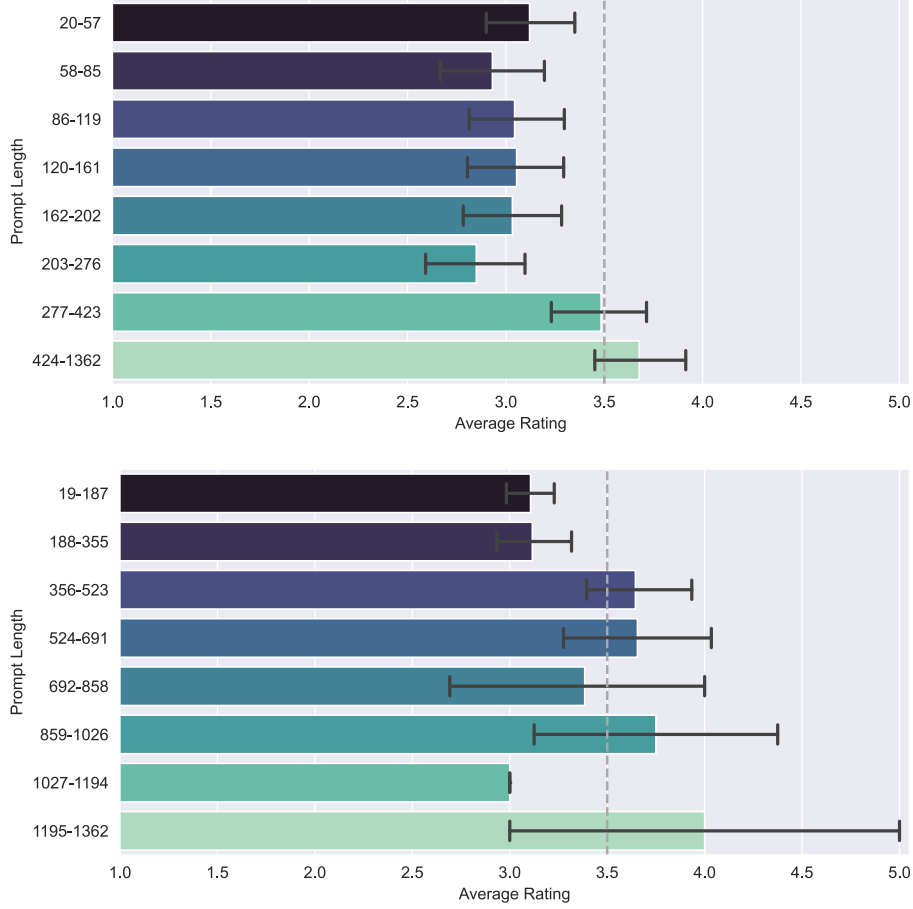


Figure 4: Prompt length (for two binnings) has no significant effect on the rating. Over all prompts, average is 3.2, with a standard deviation of 1.23.

5 Conclusion

We have examined the behavior of ChatGPT across various datasets that test multiple aspects of mathematical skill. Contrary to the media sensation that ChatGPT has caused (see the Twitter counts from Figure 1), ChatGPT is not yet ready to deliver high-quality proofs or calculations consistently. At the same time, the quality of the answer can be positively surprising. In Section F in the appendix, we collect the best and the worst results for a number of selected datasets. The best responses can be seen to justify the media sensation. It seems fair to say that ChatGPT is *inconsistently* bad at advanced mathematics: While its ratings drop with the mathematical difficulty of a prompt, it does give insightful proofs in a few cases.

However, ChatGPT falls short of achieving the same performance as models that are specifically trained for one single task. These models, in contrast, lack the flexibility of ChatGPT, which is a *universal* tool suitable for any area of mathematics.

Nonetheless, its ability to search for mathematical objects, given information about them, is where ChatGPT shines. It received its highest scores on the *Reverse Definition Retrieval* files from the *Search-Engine-Aspects* subdataset.

Because of prohibitive annotation effort, our dataset is not large enough to be used to fine-tune LLMs in

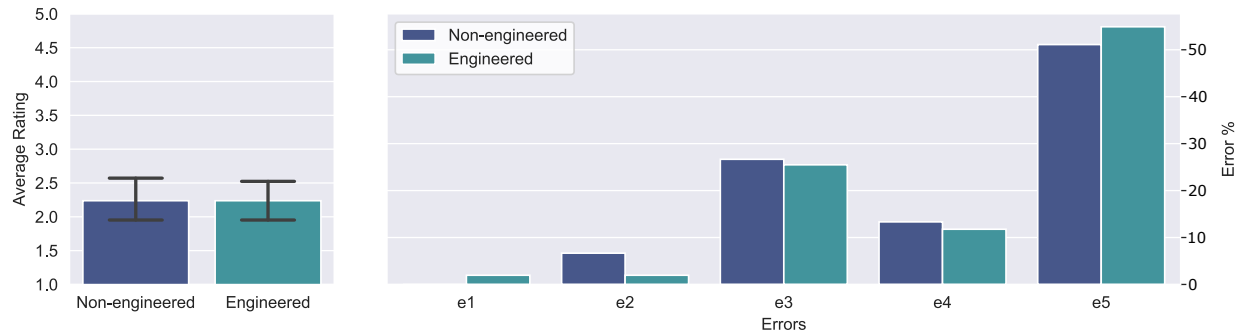


Figure 5: Effect of prompt engineering on the rating (left-most bars). Prompt engineering seems to only reduce error codes with smaller indexes (**e2**, **e3**, **e4**) that are not rooted in faulty logic. On the other hand, error codes related to logical mistakes (**e5**) even increase. Moreover, prompt engineering does not affect the average rating at all.

order to increase their mathematical ability; though we believe it is sufficiently comprehensive to allow an evaluation of existing LLMs. We also note as a recommendation for future LLM design, that incorporating some form automatic evaluation capabilities, as done by [16] is essential for lowering the cost of rating the output.

We hope that the dataset that we release with this publication will motivate other professional mathematicians to contribute in order to establish a thorough benchmark for assessing the mathematical abilities of LLMs. We will allow pull requests on our GitHub repository and encourage public participation. We encourage other researchers to mine our dataset beyond the descriptive statistics we have computed, in order to gain a deeper understanding of the behavior of ChatGPT (and other LLMs) on mathematics.

References

- [1] Aida Amini, Saadia Gabriel, Shanchuan Lin, and Rik Koncel-Kedziorski et al. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 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 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245.
- [2] Malik Amir, Yang-Hui He, Kyu-Hwan Lee, Thomas Oliver, and Eldar Sultanow. Machine Learning Class Numbers of Real Quadratic Fields. *arXiv preprint arXiv:2209.09283*, 2022.
- [3] Sheldon Axler. *Linear algebra done right*. Springer, 2015.
- [4] Amos Azaria. ChatGPT Usage and Limitations, Retrieved 2023-01-15. URL <https://hal.science/hal-03913837>.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and Jared D Kaplan et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [6] Francois Charton, Amaury Hayat, and Guillaume Lample. Learning advanced mathematical computations from examples. In *International Conference on Learning Representations*, 2021.
- [7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, and Henrique Ponde de Oliveira Pinto et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

- [8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, and Gaurav Mishra et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, and Heewoo Jun et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [10] Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, and Daniel Zheng et al. Advancing mathematics by guiding human intuition with AI. *Nature*, 600(7887):70–74, 2021.
- [11] Jörg Denzinger, Matthias Fuchs, Christoph Goller, and Stephan Schulz. Learning from previous proof experience: A survey. Technical report, TU München, 1999.
- [12] R. Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.
- [13] Arthur Engel. *Problem-Solving Strategies*. Springer, New York, 1998.
- [14] John Harrison, Josef Urban, and Freek Wiedijk. History of interactive theorem proving. In *Computational Logic*, volume 9, pages 135–214, 2014.
- [15] Yang-Hui He. Machine-learning the string landscape. *Physics Letters B*, 774:564–568, 2017. doi: <https://doi.org/10.1016/j.physletb.2017.10.024>.
- [16] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, and Steven Basart et al. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- [17] Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, and Lorie De Leon et al. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. *medRxiv*, 2022. doi: 10.1101/2022.12.19.22283643.
- [18] Guillaume Lample and François Charton. Deep learning for symbolic mathematics. *arXiv preprint arXiv:1912.01412*, 2019.
- [19] Teven Le Scao and Angela Fan et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- [20] Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, and Henryk Michalewski et al. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, 2022.
- [21] 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)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015.
- [22] Sascha Lobo. Das Ende von Google, wie wir es kannten, Retrieved 2023-01-10. URL <https://www.spiegel.de/netzwelt/netzpolitik/bessere-treffer-durch-chatgpt-das-ende-von-google-wie-wir-es-kannten-kolumne-a-77820af6-51d7-4c03-b822-cf93094fd709>.
- [23] Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. A survey of deep learning for mathematical reasoning. *arXiv preprint arXiv:2212.10535*, 2022.
- [24] Tranquil Sea Of Math. Does ChatGPT code LaTeX and write proofs?, Retrieved 2023-01-12. URL https://www.youtube.com/watch?v=ge2N7VI_8P0.
- [25] The mathlib Community. The lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*. ACM, jan 2020. doi: 10.1145/3372885.3373824.

- [26] James R. Munkres. *Topology*. Prentice-Hall, Upper Saddle River, NJ, 2000.
- [27] Natalie. ChatGPT – Release Notes, Retrieved 2023-01-31. URL <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>.
- [28] John Naughton. The ChatGPT bot is causing panic now – but it’ll soon be as mundane a tool as Excel, Retrieved 2023-01-14. URL <https://www.theguardian.com/commentisfree/2023/jan/07/chatgpt-bot-excel-ai-chatbot-tec>.
- [29] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, and Carroll L. Wainwright et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.
- [30] Piotr Piękos, Mateusz Malinowski, and Henryk Michalewski. Measuring and improving BERT’s mathematical abilities by predicting the order of reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 383–394. Association for Computational Linguistics, August 2021. doi: 10.18653/v1/2021.acl-short.49.
- [31] Markus N. Rabe, Dennis Lee, Kshitij Bansal, and Christian Szegedy. Language modeling for formal mathematics. *arXiv preprint arXiv:2006.04757*, 2020.
- [32] Richard Van Noorden @richvn@mastodon.social [@Richvn]. *Huh. ChatGPT confidently gives the right kind of reasoning to solve this math problem, but whiffs on the algebra in the middle and gets the answer wrong.* [Tweet]. Twitter, Retrieved 2023-01-09. URL <https://twitter.com/Richvn/status/1598714487711756288>.
- [33] Kevon Roose. The Brilliance and Weirdness of ChatGPT, Retrieved 2023-01-24. URL <https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html>.
- [34] David Rozado. What is the IQ of ChatGPT?, Retrieved 2023-01-09. URL <https://davidrozado.substack.com/p/what-is-the-iq-of-chatgpt>.
- [35] W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976.
- [36] Walter Rudin. *Functional analysis*. McgGraw-Hill, Inc, New York, 1991.
- [37] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. doi: 10.1147/rd.33.0210.
- [38] Sarah Wiegrefe (sigmoid.social/@sarah) [@sarahwiegrefe]. *If text-davinci-001 is a rough approximate to the model reported in the NeurIPS 2020 paper, and text-davinci-002 is InstructGPT in the 2022 preprint, then what is just "davinci"? Trying to reproduce results from a time before this naming existed.* [Tweet]. Twitter, 2023-01-15. URL <https://twitter.com/BlancheMinerva/status/1584788418751406080>.
- [39] John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, and Jacob Menick et al. ChatGPT: Optimizing Language Models for Dialogue, Retrieved 2023-01-24. URL <https://openai.com/blog/chatgpt>.
- [40] teddy [@teddynpc]. *I made ChatGPT take a full SAT test. Here’s how it did:* [Tweet]. Twitter, 2023-01-13. URL <https://twitter.com/teddynpc/status/1598767389390573569>.
- [41] Christian Terwiesch. Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course, Retrieved 2023-01-04. URL <https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP.pdf>.
- [42] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, and Apoorv Kulshreshtha et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

- [43] Timothy Gowers [@wtgowers]. *It's amusing when ChatGPT makes ridiculous mathematical mistakes. But of course, it's more interesting to find out what it can do well. Here's one example that wasn't bad: I gave it a very rough outline of a proof and asked it to fill in the details.* [Tweet]. Twitter, 2023-01-13. URL <https://twitter.com/wtgowers/status/1611750773607604224>.
- [44] Carroll Wainwright and Ryan Lowe. InstructGPT: Training Language Models to Follow Instructions with Human Feedback , Retrieved 2023-01-09. URL <https://github.com/openai/following-instructions-human-feedback>.

Appendix

A Rating policy

Below is the policy that was followed by each assessor of ChatGPT’s output regarding the rating score:

- 0 → ChatGPT cannot process the query (due to consistently occurring timeouts or errors);
- 1 → failure to understand the query (e.g., the user asks it something about number theory and it responds with information about differential equations);
- 2 → query was understood but the answer was entirely wrong (e.g., the user asks what the prime divisors of 111 are⁸ and it responds with 8 and 6);
- 3 → query was understood but the answer was only partially correct (e.g., the user asks it what the prime divisors of 111 are and it responds with 3 and 6);
- 4 → query was understood and the answer was mostly correct (e.g., the user asks it what the prime divisors of 222 are⁹ and it responds with 3 and 37);
- 5 → query was understood and answer was completely correct.

B Error and warning code policy

Below is the policy that was followed by each assessor of ChatGPT’s output regarding the error codes and warning codes:

Error codes:

- **e1** → missing examples (e.g., the user asks it what the prime divisors of 111 are and it responds with 3, missing 37);
- **e2** → a few wrong statements (e.g., the user asks it what the prime divisors of 30030 are¹⁰ and it responds with 2, 3, 5, 7, 13);
- **e3** → a lot of wrong statements (e.g., the user asks it what the prime divisors of 30030 are and it responds with 2, 5, 8, 12, 13, 15);
- **e4** → wrong computations (an additional error flag to disambiguate between statements that are of computational nature or not);
- **e5** → denotes wrong logic or wrong flow of arguments, which we further subdivide into specific flags, as we prohibit the use of **e5** on its own, as it would be uninformative:
 - **e5_1** → ChatGPT claims that to complete a proof, statements need to be shown that are unrelated to the claim;
 - **e5_2** → a proof step is missing;
 - **e5_3** → an edge case has not been considered ;
 - **e5_4** → an inference step is not supported (e.g., ChatGPT claims that from A follows B, but this claim is not true);
 - **e5_5** → circular logical argument (using the hypothesis to prove the hypothesis);

⁸They are 37 and 3.

⁹They are 2, 37 and 3.

¹⁰They are 2, 3, 5, 7, 11.

- **e6** → the general set-up is understood but the legal operations are not respected (e.g., we are given a puzzle where we’re only allowed to add even integers but it changes the rules and motivates the solution by allowing addition of odd integers).

Warning codes:

- **w1** → ChatGPT is withholding essential information related to the prompt (e.g., the user asked it something about the integral $\int_{-\infty}^{\infty} e^{-x^2} dx$ and it answers correctly but doesn’t tell the user that the integral was actually a famous, named integral, the Gaussian integral)
- **w2** → ChatGPT is rambling (after answering (correctly or incorrectly) ChatGPT tells the user much more than the user wanted to know)
- **w3** → ChatGPT is hallucinating (after answering (correctly or incorrectly) ChatGPT tells the user unrelated stuff)
- **w4** → weirdness (ChatGPT is being weird, for example, by using a weird proof structure (where applicable), using strange mathematical formulations, or by adopting a strange tone of the conversation or making opinionated statements)
- **w5** → it takes a number of tries to get ChatGPT to answer a prompt (because occasional timeouts or errors occur that are related to this particular prompt)
- **w6** → ChatGPT changes the notation from the prompt without being instructed to do so (e.g., the prompt contains a vector space \mathbf{X} , but ChatGPT calls it \mathbb{F})

C ChatGPT version

We focus on the 9th-January-2023 version of ChatGPT [27], as made available through web access at chat.openai.com/chat. This version was online for the majority of the writing process of this article. Since the 30th of January 2023, a new version has been online which we will comment on at the end of this section.

Focusing on one version is necessary because precise details of the model architecture and, in particular, ChatGPT’s training methodology have not been released in the introductory statement [39] by its creator, OpenAI. ChatGPT is the latest model of the GPT lineage [27], being based on InstructGPT, which in turn is based on a trained GPT-3 [5], and fine-tuned using reinforcement learning with human feedback [29]. We note that even for models that predate ChatGPT, such as InstructGPT, where research articles and model cards [44] have been released, full reproducibility is not possible since the code and exact datasets have not been released. Furthermore, it was confirmed by OpenAI employees that a slight mismatch exists between the trained model that is accessible via OpenAI web interface and the model referred to in the official paper [38]. This indicates how essential it is to document carefully which model our analysis pertains to; in our dataset, we have included time stamps for each prompt in order to be able to track any changes in ChatGPT’s version that have occurred.

In contrast to the version that was studied in this manuscript, the latest version from 30th-January-2023 is claimed to have “improved factuality and mathematical capabilities.” [27]. We compared the outputs computed by the latest version to the old version on two of our subdatasets and could not find any substantial differences in the resulting average rating: On the *W. Rudin Functional Analysis (ch. 1)* file of the *Grad-Text* subdataset (29 prompts) the average rating of the 9th January version was 2.96 and on the 30th January version it was 2.90. Additionally, on the 20 prompts of the *Proofs Collection A* file in the *Holes-in-Proofs* subdataset, where missing equations or conditions need to be filled in, the average rating slightly dropped from 3.5 to 3.3. It is planned to add a full comparison between the 09th and 30th January versions in the near future.

D Further figures of ChatGPT's performance

In this section, we collect further figures, which are not directly related to our discussion in the main body of the article, but which provide other views on the data and other descriptive statistics.

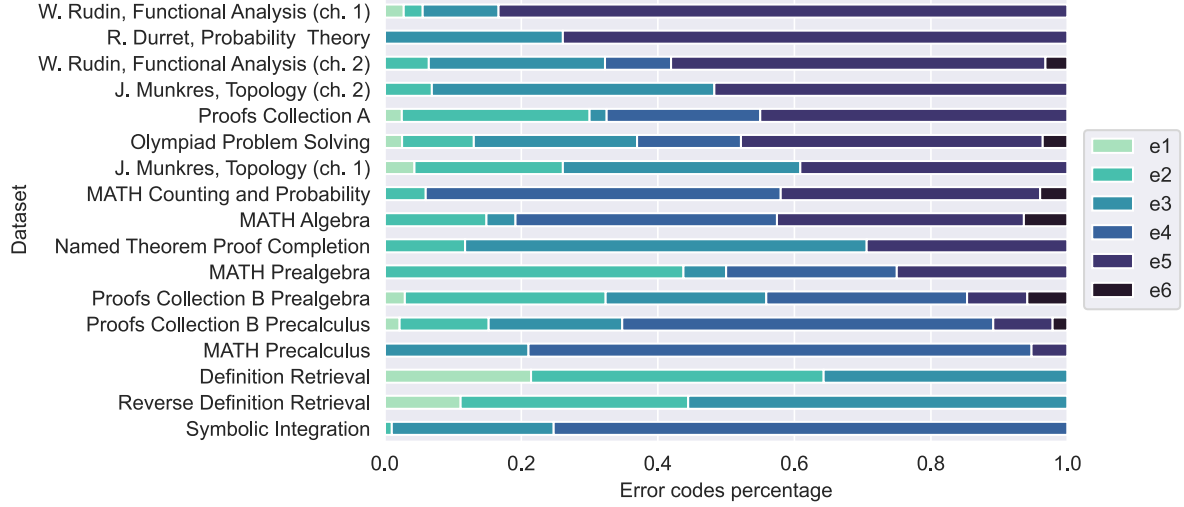


Figure 6: Error types per dataset sorted by relative frequency of e5 (the most high-level error type).

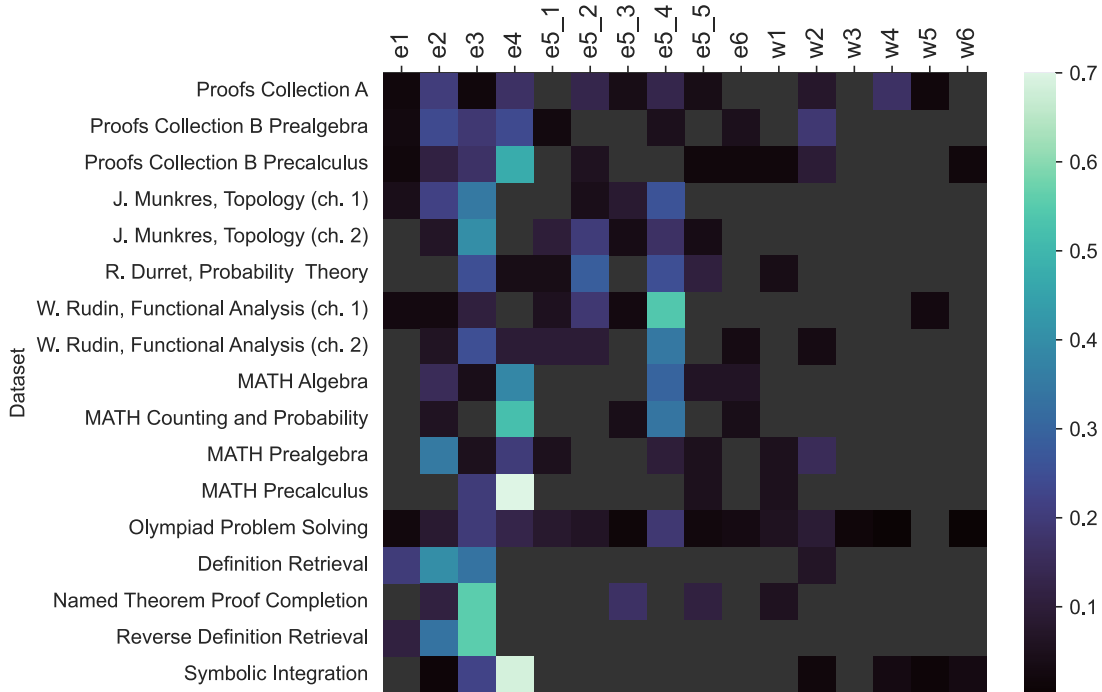


Figure 7: Relative frequencies of error codes by subdatasets and files.

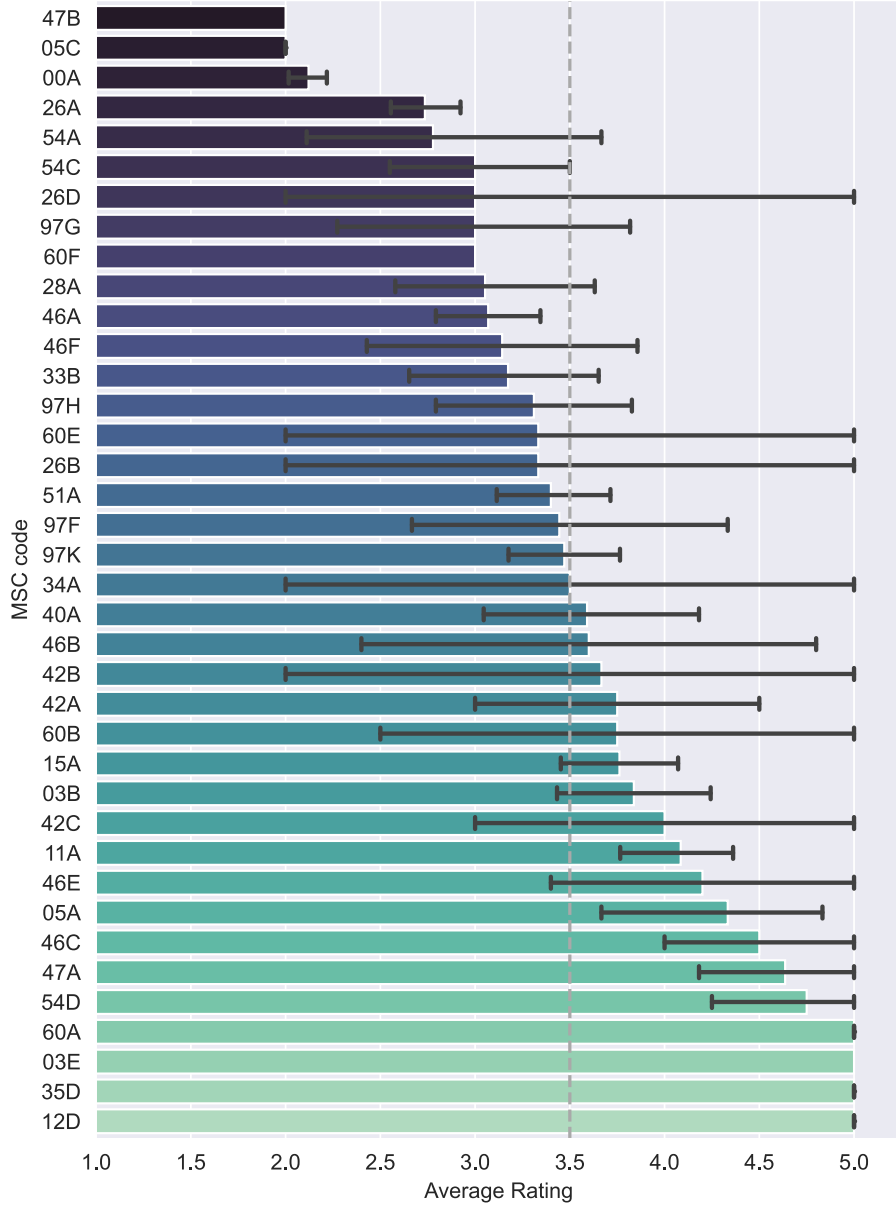


Figure 8: Rating by MSC codes (MSC code longer version).

E Copyright

Some of the subdatasets contain prompts that may be protected under copyright (e.g., exercises from textbooks). In these cases, we have used the SHA256 hash function, and the publicly released datasets will only contain the hashed value of the prompt that we used, together with a detailed reference that points to the exact text page and theorem or exercise that was used. The non-hashed version of these datasets is kept privately.

F Best-3 and worst-3 across selected subdatasets and files

We list below the best and worst answers of Chat-GPT over a selection of subdatasets. For readability, the prompts and answers are lightly modified so that the L^AT_EX-based formulas are correctly displayed and whitespace is removed.

F.1 Grad-Text

Examples from *Grad-Text* subdataset, comprising the books [36, 26, 12].

Best-3

"prompt": "Let μ be a measure, $\|f\|_\infty = \inf\{M : \mu(\{x : |f(x)| > M\}) = 0\}$ and $\|g\|_1 = \int |g|d\mu$. Prove that $\int |fg|d\mu \leq \|g\|_1 \|f\|_\infty$."

"output": "Let $M = \|f\|_\infty$ and $\epsilon > 0$. Then, by the definition of $\|f\|_\infty$, we have $\mu(\{x : |f(x)| > M + \epsilon\}) = 0$. Therefore,

$$\int |fg|d\mu = \int_{|f| \leq M+\epsilon} |fg|d\mu + \int_{|f| > M+\epsilon} |fg|d\mu \leq (M+\epsilon) \int |g|d\mu + \int_{|f| > M+\epsilon} |f||g|d\mu.$$

The second term on the right side is zero because $\mu(\{x : |f(x)| > M + \epsilon\}) = 0$. So we have $\int |fg|d\mu \leq (M+\epsilon) \int |g|d\mu = (M+\epsilon) \|g\|_1$. As ϵ can be arbitrarily small, we can let it go to zero and we get $\int |fg|d\mu \leq M \|g\|_1 = \|f\|_\infty \|g\|_1$. Hence proved."

"prompt": "If $\{T_\alpha\}$ is a family of topologies on X , show that $\bigcap_\alpha T_\alpha$ is a topology on X ."

"output": "To show that $\bigcap_\alpha T_\alpha$ is a topology on X , we need to show that it satisfies the three axioms of a topology: \emptyset and X are both in $\bigcap_\alpha T_\alpha$ since \emptyset and X are in every topology T_α . If $\{U_i\}_{i \in I}$ is a collection of sets in $\bigcap_\alpha T_\alpha$, then $\bigcup_{i \in I} U_i$ is also in $\bigcap_\alpha T_\alpha$. Since each T_α is a topology, $\bigcup_{i \in I} U_i$ is in every T_α , and so it is also in $\bigcap_\alpha T_\alpha$. If $\{U_i\}_{i \in I}$ is a collection of sets in $\bigcap_\alpha T_\alpha$, then $\bigcap_{i \in I} U_i$ is also in $\bigcap_\alpha T_\alpha$. Since each T_α is a topology, $\bigcap_{i \in I} U_i$ is in every T_α , and so it is also in $\bigcap_\alpha T_\alpha$. Therefore, $\bigcap_\alpha T_\alpha$ is a topology on X ."

"prompt": "Let X be a vector space and A and B subsets of X . If A and B are balanced, so is $A+B$."

"output": "A subset S of a vector space X is called balanced if for every vector x in X and every scalar λ with $|\lambda| \leq 1$, λx is in S if x is in S . To prove that $A+B$ is balanced, we must show that for every vector $x = a+b$ in $A+B$, and every scalar λ with $|\lambda| \leq 1$, $\lambda x = \lambda(a+b) = \lambda a + \lambda b$ is in $A+B$ if x is in $A+B$. Since A and B are balanced, we know that for every a in A and every scalar λ with $|\lambda| \leq 1$, λa is in A if a is in A . Similarly, for every b in B and every scalar λ with $|\lambda| \leq 1$, λb is in B if b is in B . So, if $x = a+b$ is in $A+B$, then a is in A and b is in B . Therefore, λa is in A and λb is in B , and thus $\lambda x = \lambda a + \lambda b$ is in $A+B$. Therefore, $A+B$ is balanced."

Worst-3

"prompt": "Find a function $f: \mathbb{R} \rightarrow \mathbb{R}$ that is continuous at precisely one point."

"output": "One example of such a function is

$$f(x) = \begin{cases} x^2 & \text{if } x \neq 0 \\ 1 & \text{if } x = 0. \end{cases}$$

This function is continuous at $x=0$ but not at any other point."

"prompt": "Let X be a topological vector space. All sets mentioned below are understood to be the subsets of X . Prove the following statement: If A is compact and B is closed, then $A+B$ is closed.",

"output": "A topological vector space is a vector space with a topology such that the vector addition and scalar multiplication are continuous. To prove that $A+B$ is closed, we must show that the complement of $A+B$ is open. Let $y \in X \setminus (A+B)$. Then there exists some $x_1 \in A$ and $x_2 \in B$ such that $y = x_1 + x_2$. Since A is compact and B is closed, there exist neighborhoods U_1 of x_1 and U_2 of x_2 such that U_1 is relatively compact and $X \setminus U_2$ is open. Now consider the set $V = U_1 + U_2 = \{x_1 + x_2 : x_1 \in U_1, x_2 \in U_2\}$. This set is an open neighborhood of y and it is contained in $X \setminus (A+B)$, which shows that $X \setminus (A+B)$ is open and hence $A+B$ is closed."

"prompt": "Let X be a topological vector space. All sets mentioned below are understood to be the subsets of X . Prove the following statement: The sum of two closed sets may fail to be closed.",

"output": "A counterexample can be used to prove that the sum of two closed sets may fail to be closed. For example, let $X = \mathbb{R}$ and let $A = [0, 1]$ and $B = (1, 2]$. Both A and B are closed sets in \mathbb{R} , but $A+B = [0, 2]$ is not a closed set in \mathbb{R} ."

F.2 Holes-in-Proofs (Proofs Collection A)

Examples from the *Holes-in-Proofs* subdataset, *Proofs Collection A* file, based on the books [3, 35] and questions from math.stackexchange.com

Best-3

"prompt": "We want to show that any open subset of \mathbb{R} is a countable union of disjoint open intervals. Let U be a non-empty open subset of \mathbb{R} . For $x, y \in U$ define $x \sim y$ iff [MISSING]. It's easily checked that \sim is an equivalence relation on U whose equivalence classes are pairwise disjoint open intervals in \mathbb{R} . (The term \code{interval} here includes unbounded intervals, i.e., rays.) Let \mathcal{I} be the set of \sim -classes. Clearly $U = \bigcup_{I \in \mathcal{I}} I$. For each $I \in \mathcal{I}$ choose a rational $q_I \in I$; the map $\mathcal{I} \rightarrow \mathbb{Q} : I \mapsto q_I$ is injective, so \mathcal{I} is countable. Can you fill in the missing condition?",

"output": "The missing condition is \"there exists an open interval containing x and y that is contained in U ." "

"prompt": "The p -norm on \mathbb{R}^n is given by $\|x\|_p = \left(\sum_{k=1}^n |x_k|^p\right)^{1/p}$. We want to show that $\|x\|_p \leq C\|x\|_q$ for arbitrary $1 \leq p < q$. We compute

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \leq \left(\left(\sum_{i=1}^n |x_i|^q\right)^{\frac{p}{q}} n^{1-\frac{p}{q}}\right)^{1/p} = \left(\sum_{i=1}^n |x_i|^q\right)^{\frac{1}{q}} n^{\frac{1}{p}-\frac{1}{q}} = n^{1/p-1/q} \|x\|_q.$$

Which result has been used in the inequality above?",

"output": "The inequality used in the above calculation is Hölder's inequality. It states that for any $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, and for any vectors $x \in \mathbb{R}^n$, $y \in \mathbb{R}^n$,

$$\left(\sum_{i=1}^n |x_i y_i|\right) \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |y_i|^q\right)^{1/q}."$$

"prompt": "Let f be a real-valued function continuous on $[a, b]$ and differentiable on (a, b) . Suppose that $\lim_{x \rightarrow a} f'(x)$ exists. Then, we want to prove that f is differentiable at a and $f'(a) = \lim_{x \rightarrow a} f'(x)$. Using the definition of the derivative, we need to show that

$$\lim_{h \rightarrow 0^+} \frac{f(a+h) - f(a)}{h}$$

exists and is equal to $\lim_{x \rightarrow a^+} f'(x)$. Note that for $h > 0$ the Mean Value Theorem provides a point c_h with $a < c_h < a+h$ such that [MISSING]. Finally, note that $c_h \rightarrow a^+$ as $h \rightarrow 0^+$. Can you fill in the missing equation?",

"output": "The missing equation from the Mean Value Theorem is:

$$f'(c_h) = \frac{f(a+h) - f(a)}{h}."$$

Worst-3

"prompt": "We want to compute

$$\sum_{n=1}^{\infty} \frac{2n}{3^{n+1}}.$$

Let's first find a general formula for the following sum: $S_m = \sum_{n=1}^m nr^n$. Notice that

$$S_m - rS_m = -mr^{m+1} + \sum_{n=1}^m r^n = -mr^{m+1} + \frac{r - r^{m+1}}{1 - r} = \frac{mr^{m+2} - (m+1)r^{m+1} + r}{1 - r}.$$

Hence

$$S_m = \frac{mr^{m+2} - (m+1)r^{m+1} + r}{(1-r)^2}.$$

Can you finish the calculation?",

"output": "Yes, to finish the calculation we can substitute the values of $m = \infty$, $r = 2/3$ into the formula we found for S_m :

$$S_{\infty} = \frac{\infty \left(\frac{2}{3}\right)^{\infty+2} - (\infty+1) \left(\frac{2}{3}\right)^{\infty+1} + \frac{2}{3}}{\left(1 - \frac{2}{3}\right)^2}$$

This simplifies to

$$S_{\infty} = \frac{\frac{2}{9}}{\frac{1}{9}} = 2,"$$

"prompt": "Suppose V is an inner product space and $T \in \mathcal{L}(V)$. We want to prove that

$$|\det(T)| = \det \sqrt{T^*T}.$$

By the Polar Decomposition there is an isometry $S \in \mathcal{L}(V)$ such that

$$T = S\sqrt{T^*T}.$$

Can you finish the proof?",

"output": "Yes, once we have that $T = S\sqrt{T^*T}$, we can take the determinant of both sides to get

$$\det(T) = \det(S) \det(\sqrt{T^*T}).$$

Since S is an isometry, it preserves inner products, so $\det(S) = \pm 1$. Therefore,

$$|\det(T)| = |\det(S)| |\det(\sqrt{T^*T})| = |\det(\sqrt{T^*T})|.$$

Since the determinant of a matrix is always non-negative, we have that

$$|\det(T)| = \det(\sqrt{T^*T})"$$

"prompt": " F is a finite field of order q . We want to compute the size of $GL_n(F)$. In order for an $n \times n$ matrix to be invertible, we need the rows to be linearly independent. As you note, we have $q^n - 1$ choices for the first row; now, there are q vectors in the span of the first row, so we have $q^n - q$ choices for the second row. Now, let v_1, v_2 be the first two rows. Then the set of vectors in the span of v_1, v_2 is of the form $\{c_1 v_1 + c_2 v_2 | c_1, c_2 \in F\}$. This set is of size q^2 , as we have q choices for c_1 and q choices for c_2 . Thus, we have $q^n - q^2$ choices for the third row. Continuing this gives the desired formula. Can you state the formula?",

"output": "The size of $GL_n(F)$ is given by the formula

$$(q^n - 1)(q^{n-1} - q)(q^{n-2} - q^2) \cdots (q^2 - q)(q - 1)."$$

F.3 Holes-in-Proofs (Proofs Collection B Prealgebra and Precalculus)

Examples from the *Holes-in-Proofs* subdataset, *Proofs Collection B Prealgebra* and *Proofs Collection B Precalculus* files, based on [16].

Best-3

"prompt": "Find the smallest positive integer that is greater than 1 and relatively prime to the product of the first 20 positive integers. Reminder: two numbers are relatively prime if their greatest common divisor is 1. Solution: <MISSING> the smallest possible integer is 23. Fill in the missing steps.",

"output": "The first 20 positive integers are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, and 20. The product of these numbers is 2432902008176640000. To find the smallest positive integer greater than 1 that is relatively prime to this product, we can start at 2 and incrementally check each number until we find one that has a greatest common divisor (GCD) of 1 with the product. We can check 2, 3, 4, ..., 22, and see that the GCD of each of these numbers with the product is not 1. However, when we check 23, we find that the GCD of 23 and 2432902008176640000 is 1. Therefore, 23 is the smallest positive integer that is greater than 1 and relatively prime to the product of the first 20 positive integers.",

"prompt": "Find the number of solutions to $\cos \frac{x}{4} = \cos x$ in the interval $0 < x < 24\pi$. Solution: From the equation $\cos \frac{x}{4} = \cos x$, $\cos x - \cos \frac{x}{4} = 0$. From the sum-to-product formula, we can write this as

$$-2 \sin \frac{5x}{8} \sin \frac{3x}{8} = 0.$$

<MISSING> Fill in the missing steps.",

"output": "Since the product of sines is equal to zero, either $\sin \frac{5x}{8} = 0$ or $\sin \frac{3x}{8} = 0$. Solving for each case: $\sin \frac{5x}{8} = 0 \implies \frac{5x}{8} = n\pi$, where n is an integer. This gives us the solutions $x = \frac{8n}{5}\pi$. $\sin \frac{3x}{8} = 0 \implies \frac{3x}{8} = n\pi$, where n is an integer. This gives us the solutions $x = \frac{8n}{3}\pi$. Now we need to find the solutions that are in the interval $0 < x < 24\pi$. $x = \frac{8n}{5}\pi$ gives solutions $x = \frac{16}{5}\pi, \frac{24}{5}\pi, \frac{32}{5}\pi, \frac{40}{5}\pi, \frac{48}{5}\pi$ which are in the given interval. $x = \frac{8n}{3}\pi$ gives solutions $x = \frac{8}{3}\pi, \frac{16}{3}\pi, \frac{24}{3}\pi$ which are in the given interval. Therefore, there are 5 solutions for $x = \frac{8n}{5}\pi$ and 3 solutions for $x = \frac{8n}{3}\pi$ in the given interval. The total number of solutions is $5+3=8$."

"prompt": "Let \mathbf{a} , \mathbf{b} , \mathbf{c} be three vectors such that $\|\mathbf{a}\| = \frac{1}{\sqrt{2}}$, $\|\mathbf{b}\| = \frac{1}{\sqrt{3}}$, $\|\mathbf{c}\| = \frac{1}{\sqrt{6}}$, $\|\mathbf{a} + \mathbf{b} + \mathbf{c}\| = 1$, and

$$\mathbf{a} = k(\mathbf{b} \times \mathbf{c})$$

for some scalar k . Find the angle between \mathbf{b} and \mathbf{c} , in degrees. Solution: Since $\mathbf{b} \times \mathbf{c}$ is orthogonal to both \mathbf{b} and \mathbf{c} , $\mathbf{a} \cdot \mathbf{b} = 0$ and $\mathbf{a} \cdot \mathbf{c} = 0$. Since $\|\mathbf{a} + \mathbf{b} + \mathbf{c}\| = 1$,

$$(\mathbf{a} + \mathbf{b} + \mathbf{c}) \cdot (\mathbf{a} + \mathbf{b} + \mathbf{c}) = 1.$$

<MISSING> Fill in the missing steps.",

"output": "Expanding the dot product on the left side gives:

$\mathbf{a} \cdot \mathbf{a} + \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{c} + \mathbf{c} \cdot \mathbf{a} + \mathbf{c} \cdot \mathbf{b} + \mathbf{c} \cdot \mathbf{c} = 1$. Since $\mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{c} = \mathbf{b} \cdot \mathbf{a} = \mathbf{c} \cdot \mathbf{a} = 0$, this simplifies to: $|\mathbf{a}|^2 + |\mathbf{b}|^2 + |\mathbf{c}|^2 + 2\mathbf{b} \cdot \mathbf{c} = 1$. We know that $|\mathbf{a}|^2 = \frac{1}{2}$, $|\mathbf{b}|^2 = \frac{1}{3}$, $|\mathbf{c}|^2 = \frac{1}{6}$, and substituting these values in gives: $\frac{1}{2} + \frac{1}{3} + \frac{1}{6} + 2\mathbf{b} \cdot \mathbf{c} = 1$. Solving for $\mathbf{b} \cdot \mathbf{c}$ gives: $\mathbf{b} \cdot \mathbf{c} = \frac{1}{6} - \frac{1}{6} = 0$. Which means the vectors are orthogonal, hence the angle between them is 90 degrees."

Worst-3

"prompt": "The line described by

$$\begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix} + t \begin{pmatrix} k \\ 2 \\ 1 \end{pmatrix}$$

is perpendicular to the line described by

$$\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + u \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$$

and passes through the point $(4, a, b)$. Find $a + b + k$. Solution: The direction vector of the first line is $\begin{pmatrix} k \\ 2 \\ 1 \end{pmatrix}$, and the direction vector of the second line is $\begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$.

<MISSING> Then $t = -1$, so $a = -3$ and $b = 2$, so $a + b + k = \boxed{-3}$. Fill in the missing steps.",

"output": "The direction vector of the first line is $\begin{pmatrix} k \\ 2 \\ 1 \end{pmatrix}$, and the direction

vector of the second line is $\begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$. To find the point of intersection, we can set the two parametric equations equal to each other:

$$\begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix} + t \begin{pmatrix} k \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + u \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$$

Solving for t, u we get $t = -1, u = -3$. So the point of intersection is

$$\begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix} - \begin{pmatrix} k \\ 2 \\ 1 \end{pmatrix} = (4, a, b). \text{ Given that, we get } a = -3 \text{ and } b = 2, \text{ then } a + b + k = \boxed{-3}."$$

"prompt": "How many positive integers less than 100 and divisible by 3 are also divisible by 4? Solution: If we start counting from 1, every third consecutive integer is divisible by 3 and every fourth consecutive integer is divisible by 4. <MISSING> so there are $\boxed{8}$ multiples of 12 between 1 and 100. Fill in the missing steps.",

"output": "The missing steps are: We need to find the multiples of 12, which are divisible by both 3 and 4. To find the first multiple of 12 less than 100, we divide 100 by 12 and round down to the nearest integer, which is

8. To find the last multiple of 12 less than 100, we divide 12 by 100 and round up to the nearest integer, which is 8. The number of multiples of 12 between 1 and 100 is the difference between the last and first multiple, which is $8 - 8 = 0$. So there are no positive integers less than 100 that are divisible by both 3 and 4."

"prompt": "The complex numbers $5 + 10i$, $7 + 2i$, and $11 + 3i$ are three vertices of a rectangle. Find the fourth vertex. Solution: Let $a = 5 + 10i$, $b = 7 + 2i$, and $c = 11 + 3i$. Let d be the fourth vertex. Note that

$$c - b = 4 + i$$

and

$$a - b = -2 + 8i = 2i(4 + i) = 2i(c - b).$$

Thus, the angle between the segment joining a and b and the segment joining b and c is 90° . <MISSING> Fill in the missing steps."

"output": "Since the angle between the segment joining a and b and the segment joining b and c is 90° , it follows that the segment joining c and d is perpendicular to the segment joining a and b . Since $c - b = 4 + i$, we have $c - d = -(b - d) = -(-2 + 8i) = 2 - 8i$. Therefore, $d = c - (2 - 8i) = 11 + 3i - (2 - 8i) = 9 - 5i$. So the fourth vertex is $(9, -5i)$."

F.4 Olympiad-Problem-Solving

Examples from the *Olympiad-Problem-Solving* subdataset based on the book [13].

Best-3

"prompt": "n numbers are written on a blackboard. In one step, you may erase any two of the numbers, say a and b, and write, instead $(a+b)/4$. Repeating this step n-1 times, there is one number left. Prove that, initially, if there were n ones on the board, at the end, a number which is not less than $1/n$ will remain."

"output": "This can be proven by induction. Base case: $n = 2$. If there are two ones on the board, after one step, the remaining number is $(1+1)/4 = 1/2$, which is not less than $1/2$. Inductive case: Assume that the statement is true for $n = k$. Now consider $n = k+1$. If there are $k+1$ ones on the board, we can erase any two of them and write their average, which is not less than $1/k+1$ by the induction hypothesis. Therefore, we can conclude that the statement holds for all n."

"prompt": "The integers $1, \dots, n$ are arranged in any order. In one step you may switch any two neighboring integers. Prove that you can never reach the initial order after an odd number of steps."

"output": "This problem can be proven using the concept of parity. Specifically, the parity of a permutation is the number of inversions (i.e. pairs of integers (i, j) such that $i < j$ but the i th integer appears after the j th integer in the permutation) modulo 2. It can be shown that the parity of a permutation remains unchanged after any single swap of neighboring integers. Since the initial ordering has zero inversions (i.e. it has even parity), it follows that an odd number of swaps will always yield a permutation with odd parity, and thus it cannot be the same as the initial ordering."

"prompt": "Let's answer this question step by step. Derive divisibility criteria for 9 and 11.",

"output": "Divisibility Criteria for 9: If the sum of digits of a number is divisible by 9, then the number is also divisible by 9. If the difference between the sum of digits in the odd positions and even positions is divisible by 9, then the number is also divisible by 9. Divisibility Criteria for 11: If the difference between the sum of digits in the odd positions and even positions is divisible by 11, then the number is also divisible by 11. If the alternating sum of digits (first minus second plus third minus fourth, etc.) is divisible by 11, then the number is also divisible by 11."

Worst-3

"prompt": "Assume an 8×8 chessboard with the usual coloring. You may repaint all squares (a) of a row or column (b) of a 2×2 square. The goal is to attain just one black square. Can you reach the goal?",

"output": "Yes, it is possible to attain just one black square on an 8×8 chessboard. One way to do this is to repaint all the squares of the top-left 2×2 square, which will leave one black square in the bottom-right corner of the chessboard. Another way to do this is to repaint all the squares of the first column, which will leave one black square in the bottom-right corner of the chessboard."

"prompt": "Start with the positive integers $1, \dots, 4n-1$. In one move you may replace any two integers by their difference. Prove that an even integer will be left after $4n-2$ steps.",

"output": "We can prove this by mathematical induction. Base case: When $n = 1$, we have the integers 1, 2, 3, 4. We can perform the following moves: $(1, 2) \rightarrow 1, 3, 4$; $(3, 4) \rightarrow 1, 3$; $(1, 3) \rightarrow 2$, an even integer. Inductive case: Assume that the statement is true for $n = k$. We will show that it is also true for $n = k+1$. We have the integers 1, 2, 3, ..., $4(k+1)-1$. We can perform the following moves: $(1, 2) \rightarrow 1, 3, 4, \dots, 4(k+1)-1$; $(3, 4) \rightarrow 1, 3, \dots, 4(k+1)-3$; ...; $(4k-1, 4k) \rightarrow 1, 3, \dots, 4(k+1)-5$. Now we have $4k-2$ integers left, 1, 3, ..., $4(k+1)-5$, which are $4k-2$ consecutive odd integers. By the induction hypothesis, after $4k-2$ steps, we will be left with an even integer. Therefore, by mathematical induction, we can prove that an even integer will be left after $4n-2$ steps."

"prompt": "Let's answer this question step by step.. Assume an 8×8 chessboard with the usual coloring. You may repaint all squares (a) of a row or column (b) of a 2×2 square. The goal is to attain just one black square. Can you reach the goal?",

"output": "Yes, it is possible to reach the goal of having just one black square on an 8×8 chessboard. One way to do this is to repaint all of the squares in the first row and first column black, and then repaint the top-left square of the 2×2 square white. This results in one black square and all other squares white."

F.5 Symbolic-Integration

Examples from our *Symbolic-Integration* subdataset based on [18].

Best-3

"prompt": "What is the integral of $4x^7 + \sin(1 + x)$ ",
"output": "The integral of $4x^7$ is $(4/8)x^8 + C$, where C is the constant of integration. The integral of $\sin(1 + x)$ is $-\cos(1 + x) + C$. So the integral of $4x^7 + \sin(1 + x)$ is $(4/8)x^8 - \cos(1 + x) + C$."

"prompt": "What is the integral of $2048 + 32x^3$ ",
"output": "The antiderivative or indefinite integral of $2048 + 32x^3$ with respect to x is $\int (2048 + 32x^3)dx = 2048x + (32/4)x^4 + C$ where C is the constant of integration."

"prompt": "What is the integral of $12 + 6\cosh(x)$ ",
"output": "The integral of $12 + 6\cosh(x)$ is $12x + 6 \sinh(x) + C$, where C is the constant of integration."

Worst-3

"prompt": "What is the integral of $(4 + x)[4 + 3x][e^{\{x*(4+x)^2\}}$ ",
"output": "The integral of $(4 + x)[4 + 3x][e^{\{x*(4+x)^2\}}$ is not expressible in terms of elementary functions."

"prompt": "What is the integral of $xe^{\{-3x - \cosh(x)\}}$ ",
"output": "The integral of $xe^{\{-3x - \cosh(x)\}}$ with respect to x does not have a closed-form solution. It must be evaluated numerically or approximated using numerical integration methods such as Simpson's rule or the trapezoidal rule."

"prompt": "What is the integral of $1/(2 + \ln(3/2))$ ",
"output": "The integral of $1/(2 + \ln(3/2))$ cannot be expressed in terms of elementary functions and must be approximated numerically or solved using special functions."
