

# Mathematics, word problems, common sense, and artificial intelligence

Ernest Davis  
 Dept. of Computer Science  
 New York University  
 New York, NY 10012  
 davis@cs.nyu.edu

January 26, 2023

## Abstract

The paper discusses the capacities and limitations of current artificial intelligence (AI) technology to solve word problems that combine elementary knowledge with commonsense reasoning. No existing AI systems can solve these reliably. We review three approaches that have been developed, using AI natural language technology: outputting the answer directly, outputting a computer program that solves the problem, and outputting a formalized representation that can be input to an automated theorem verifier. We review some benchmarks that have been developed to evaluate these systems and some experimental studies. We discuss the limitations of the existing technology at solving these kinds of problems. We argue that it is not clear whether these kinds of limitations will be important in developing AI technology for pure mathematical research, but that they will be important in applications of mathematics, and may well be important in developing programs capable of reading and understanding mathematical content written by humans.

## 1 Mathematical problems and world knowledge

A central aspect of understanding mathematics is understanding how math applies in real-world situations and how mathematical analysis can be used to answer questions about real things. Consider, for instance, the following two elementary problems:

**Problem 1:** George has seven pennies, a dime, and three quarters. Harriet has four pennies and four quarters. First, George gives Harriet thirty-one cents in exact change; then Harriet gives him back exactly half of her pennies. How much money does George now have?

**Problem 2:** You have an empty cylindrical open container whose inside has a diameter of 8 centimeters and a height of 20 centimeters. and a pitcher with 200 ccs of water. You first empty the pitcher into the cylinder, then put a solid rock cube, 4 cm on a side, into the container so that it is sitting flush against the bottom of the container. What is the height of the water in the container?

Answering these questions requires knowledge of three kinds. **First**, and most obviously, there is the elementary math involved: elementary arithmetic and solid geometry, respectively. **Second**, one has

to know the **language**; if these questions, as printed above in English, are presented to a speaker who knows only Polish, then they will not be able to answer, no matter how well they know the math. **Third**, and most easily overlooked, each of these requires **basic knowledge about the world**, and an understanding of how the world is characterized mathematically. In problem 1, you have to know the value of U.S. coins, but more fundamentally, you have to understand the dynamics of ownership and giving: If  $A$  gives  $X$  to  $B$ , then the result is that  $B$  now owns  $X$ . In problem 2, you have to understand the basic physics of liquids and solids: solids maintain their shape over time, liquids maintain their volume but not their shape, solids and liquids do not overlap and so on.

In many simple problems, a significant part of the real-world understanding required is purely “**common sense**” — **the basic understanding of the realities of daily human existence** that is shared by all people past early childhood.<sup>1</sup> When people work on problems of this kind, the commonsense knowledge and reasoning involved is often so obvious that it goes entirely unnoticed; what is difficult is finding the mapping to mathematics and carrying out the mathematics.

Other forms of world knowledge also enter into mathematical problem solving. **Common knowledge**, such as the value of U.S. coins in problem 1, can be culturally dependent and is sometimes taught explicitly in schools, but is universal among adults in the society under discussion. **Encyclopedic knowledge** is specific facts about particular entities, such as the knowledge needed to answer the question, “How much older, in days, was George Washington than Abraham Lincoln?” **Expert knowledge** is knowledge held by experts but not by lay people. Obviously these categories are vague and there is no point in trying to delimit them with any precision.

A problem that calls on the combination of mathematics with world knowledge will generally involve other cognitive capacities as well. If it is posed in language, then it requires comprehension of the language. If the problem statement includes mathematical notation or other technical notation (musical, chemical, etc.). then that notation must be understood. If an embodied agent must solve a problem in the physical world, then understanding it requires perception, most commonly vision. If the problem involves a diagram or graph, then it requires vision plus an understanding of the graphical conventions. If an embodied agent must carry out a task in the real world, then that requires perception and manipulation.

It will be convenient to distinguish a number of categories of math problems. **A symbolic mathematics problem** is one posed in mathematical notation with minimal use of natural language, e.g. “Solve  $x^3 - 6x^2 + 11x - 6 = 0$ ”. **A word problem** is a mathematical problem with more than minimal use of natural language. **A purely mathematical word problem** is a word problem that makes minimal reference to any non-mathematical concepts, e.g. “Find a prime number  $p$  such that  $p+36$  is a square number.” “What is the volume of a regular icosahedron with diameter 5?” **A real-world word problem** is a word problem whose solution requires the use of non-mathematical knowledge of some kind. **A commonsense word problem (CSW)** is a word problem that involves significant use of commonsense knowledge and perhaps also common knowledge, but not encyclopedic or expert knowledge; this is the **most important category** for our purposes. Finally, an **elementary commonsense word problem** is a CSW that requires only elementary math, however one chooses to draw that line (generally elementary school math or high school math). Again, obviously, the lines delimiting these categories are vague.

Real-world word problems generally and commonsense word problems in particular are common in grade school math classes, rarer in introductory college courses, and extremely rare in advanced college math courses. They remain common in advanced courses in other topics that use mathematics for applications; there, generally, the knowledge involved is primarily expert knowledge, though commonsense knowledge often forms an underlying and largely unobserved foundation.

---

<sup>1</sup>This characterization of common sense has obvious significant problems and limitations, but it will suffice for our discussion here. An in-depth discussion will be found in [8].

## 2 Artificial intelligence for word problems

In recent years, artificial intelligence (AI) has made extraordinary progress in a wide range of domains, such as game playing, image tagging, speech transcription, and machine translation. In particular, in recent years AI has achieved extraordinary success at certain kinds of language tasks, and, as discussed throughout this volume, noteworthy successes at a range of mathematical tasks.

One might reasonably suppose, therefore, that problems as simple as problems 1 and 2 above must be well within the capacities of the current AI technology. However, that is not at all the case. *As of the end of 2022, there does not exist any AI that can reliably solve elementary CSWs.* The goal of this article is to elaborate on that observation; to describe at a very high level how the current technology works and what it can and cannot do, and to discuss the challenges that lie ahead. That is, this paper addresses the intersection of four issues: The use of artificial intelligence (i) to solve problems requiring math (ii) and commonsense reasoning (iii) posed in language (iv).

The challenge of developing AI systems that solve mathematical word problems has been studied since Daniel Brobrow’s STUDENT system of 1964 [2]. Skipping over the history of the problem over the subsequent 50 years, we will focus here on recent systems that apply AI technology to math word problems.

Progress in AI over the last twenty years has been almost entirely driven by *corpus-based machine learning* (ML). Broadly speaking, in corpus-based machine learning, a general-purpose computational architecture, with little or no built-in knowledge of the domain or the task, is trained on a corpus of relevant data, and it finds underlying patterns in the data that it can use<sup>2</sup> to carry out the task with some degree of accuracy. Almost all current AI programs are built around a core of modules created using corpus-based learning, often combined with additional, hand-crafted modules for particular aspects of the task.

### 2.1 Large Language Models

As discussed above, solving CWP requires combining commonsense reasoning, language abilities, and mathematics. For the last five or six years, both natural language processing (NLP) and (more surprisingly) commonsense reasoning in AI has been entirely dominated by one particular technology known as large language models (LLMs) (in current AI parlance “model” means “computer program.”). *Almost all recent work on math world problems has an LLM at the core.*<sup>3</sup>

Large language models are the most recent development of thirty years of study of how neural network AI technology (now often called “deep learning”) can be applied to language-oriented tasks. We will not attempt a technical description of their workings here (see [11]), but we will describe the characteristics most important for our discussion here.

The task that an LLM carries out is text generation in response to a prompt. It generates its response word by word. That is: the user provides the AI with an input text viz. a string of tokens (words, punctuation marks, and special symbols)  $w_1 \dots w_n$ . The AI computes the most probable (or, depending on the settings, a reasonably probable) next word  $w_{n+1}$ . It then computes word  $w_{n+2}$  that is most likely to follow  $w_1 \dots w_{n+1}$ . And so on; it repeatedly computes the most probable next word, until it reaches a halting state. (Chatbots such as ChatGPT have some additional mechanisms to produce reasonable interactions.) This is known as the “language modeling” task.

<sup>2</sup>I am including supervised learning, unsupervised learning, and reinforcement learning in this general category.

<sup>3</sup>An exception is Wolfram Alpha <https://www.wolframalpha.com/>, a manually constructed system with a large body of mathematical algorithms and encyclopedic and expert knowledge from many scientific and sociological. It takes input posed in natural language; however, its natural language abilities are comparatively limited and it has little commonsense reasoning ability.

The generation of words is thus driven by the conditional probability of producing  $w_n$  following  $w_1 \dots w_{n-1}$ ,  $P(w_n | w_1 \dots w_{n-1})$ . The AI computes this in terms of a function  $f_{\vec{\alpha}}(w_1 \dots w_n)$  which is implicitly encoded in the network structure of the AI. This function  $f_{\vec{\alpha}}$  is controlled by a parameter  $\vec{\alpha}$ , a real-valued vector. In current systems, the dimension of  $\vec{\alpha}$  ranges between about  $10^9$  and  $10^{12}$

In current system, during generation, the system only keeps track of the most recent  $k$  words where  $k$  is several hundreds or thousands. These systems therefore cannot read a long text of tens or hundreds of thousands of words and remember those as a whole.

The training set for the LLM is a vast corpus of text ( $10^{10} - 10^{13}$  tokens) downloaded from the Web. Detailed accounts of the content of these corpora have not, in general, been published; but it is safe to say that the majority consists of English language documents of various kinds, but that the training copora also include substantial quantities of texts in other languages, of software written in popular computer languages, of mathematics in mathematical notation, of images, and other kinds of data.

Training the LLM consists in finding the value of the parameter  $\vec{\alpha}$  such that the probability function  $P(w_n | w_1 \dots w_{n-1}) = f_{\vec{\alpha}}(w_1 \dots w_n)$  matches the training set  $T$  as closely as possible. That is, there is an error function  $E_T(\vec{\alpha})$  that measures the discrepancy between the system’s predictions and the actual sequences in  $T$  when the AI uses  $f_{\vec{\alpha}}$  to generate predictions. The value of  $\vec{\alpha}$  is chosen to (approximately) minimize  $E_T(\vec{\alpha})$ . This minimization uses a gradient descent algorithm. The training process requires enormous amounts of computation — months of computation time on large networks of computers — as well as a substantial amount of expert human labor. It is feasible only for large AI labs, not for individual academic researchers or small companies.

This training procedure for an LLM is carried out only once, when the LLM is created. However, it is possible to further “fine-tune” an LLM to a particular task by further training on texts specifically relevant to that task.

The underlying architecture of the LLM, which determines the function  $f_{\vec{\alpha}}$ , and the training procedure are both very general in structure and are built to carry out prediction for essentially any kind of input string with some kinds of regularities. They do not reflect *any* knowledge of the characteristics of natural language generally, of any particular language, of the external world that language describes, of the various uses of language, or of any task other than string prediction. *All* the knowledge of language, its meaning, and its content that the system possesses is in terms of how best to carry out prediction over the training set.

As with all AI systems based on neural networks/deep learning, the function  $f_{\vec{\alpha}}$  is *opaque*, in the sense that it is extremely difficult to find a relation between the vector  $\vec{\alpha}$  and the behavior of the system, either on a particular example or in general. It is therefore generally impossible to incorporate any of what is known about language, the world, or mathematics into the system other than through training. It is also impossible to debug errors as is done in “conventional” computer programming. All that can be done is to retrain the system from scratch (or, in rare cases, to add a hand-crafted patch at the back-end to deal with the problem.)

In earlier NLP research, different tasks — question answering, chatbots, summarization, information extraction, translation and so on — were each handled separately by systems built specifically for that purpose. However, in the last three or four years, it has turned out that language modeling tasks can serve as a basis for all of these; indeed for any task where the input and output are in language [3]. For some tasks, like chatbots, LLMs are currently the most effective technology. For others, like translation, they are not competitive with specialized technology but they are remarkably good.

It has also turned out, quite surprisingly, that the quality of answers generated by the very large LLMs can often be significantly improved by including a handful of examples of the kind of output desired in the prompt, or even by including general directives such as “Let’s think about this step by

step”. Thus, if you want it to translate from German to English, it helps to include a few examples of translation in the prompt; if you want it to answer mathematical questions of a particular kind, it helps to include a few examples of that kind of problem in the prompt. This is known as “few-shot prompting”; it was first observed in the LLM GPT-3, released in 2020 [4]. It has given rise to a new area of study: “prompt engineering”, the creation of prompts that guide LLMs to correct outputs of the proper form.

A final general remark is that LLMs are prone to so-called *hallucinations*: since they have no sense of underlying reality, they often generate text that reads smoothly but is actually incoherent. Depending on circumstances, a more savvy human reader can find this amusing, frightening, or misleading. Table 1 shows a mathematical example generated by the LLM ChatGPT (experiment by the author, 1/23/2023).

% clearpage

## 2.2 LLMs’ abilities at common sense, math, and language

Outside a mathematical context, LLMs achieve significant levels of success at commonsense reasoning, though they are certainly not reliable, and they have been steadily improving over time. Unquestionably LLM-based AIs are currently the most powerful and general, publicly available<sup>4</sup> technology for commonsense reasoning that have been built.

LLMs also have displayed some mathematical abilities. GPT-3, when originally released in 2020 [4], was tested on simple integer arithmetic problems. When prompted with a few correct examples, it achieved 100% accuracy on two-digit addition (e.g. “What is 35 plus 72?”) with gradually diminishing accuracy as the number of digits increases, though only 21.3% accuracy on problems involving two arithmetic operations and one-digit numbers (e.g. “What is  $(2 + 4) * 6$ ?”). More recent systems have achieved significantly higher levels of success [18] but the general pattern remains.

LLMs also can often, though unreliably, generate correct mathematics at a higher level. The errors in the “proof” of Fermat’s little theorem table 1 are amusing, but one should note what ChatGPT got right here. The theorem was correctly stated. The algebraic manipulations are correct, on the unjustified assumption that  $a-1$  is divisible by  $p$ . The general structure of an inductive proof — prove the base case  $a=1$ , prove that if the theorem holds for  $a$  then it proves for  $a+1$  — is followed, though ChatGPT fails to notice that the statement that it “proved” was not the statement it wanted to prove.

Finally, LLMs are fundamentally language technology, and in many respects their purely linguistic abilities are extraordinary and flawless, though not generally very creative. The current generation of LLMs can produce essays several paragraphs long that, in their language use, are often indistinguishable from capable human writers. I have never seen them make a grammatical error or malapropism in writing modern English.<sup>5</sup>

LLMs thus have at least some facility at the three components of CSW: common sense, math, and language. However, putting these together is a different matter.

<sup>4</sup>The CYC project [12]. begun in 1985 and still ongoing, aims at encoding commonsense knowledge in a quasi-logical symbolic system and using automated inference to do commonsense reasoning. It is certainly large: it contains some millions of “concepts” and tens of millions of propositions, and it represents several hundred person-years of expert human labor. However, the contents are proprietary: no systematic description, evaluation, or benchmark test has ever been published, and the scientific literature on its workings is very small. Consequently, it is impossible for outsiders to know what its capabilities are.

<sup>5</sup>I did once catch ChatGPT using “doth” as a plural verb in trying to write a Shakespearean pastiche.

**User query:** Please sketch the proof of Fermat's little theorem.

**ChatGPT output:** Fermat's Little Theorem states that if  $a$  is an integer and  $p$  is a prime, then the following equation holds:

$$a^p = a \pmod{p}$$

This means that if you raise  $a$  to the power of  $p$  and divide the result by  $p$ , the remainder will always be equal to  $a$ .

One way to prove Fermat's Little Theorem is by using mathematical induction.

First, we prove the base case, where  $a = 1$ . In this case, the equation holds true because  $1^p = 1$ , and 1 divided by  $p$  will always have a remainder of 1.

Next, we assume that the equation holds true for some arbitrary value of  $a$ , and we will prove that it holds true for  $a+1$ . Using the assumption that the equation holds true for  $a$ , we know that:

$$a^p = a \pmod{p}$$

We can now multiply both sides of this equation by  $a$ :

$$a^{p+1} = a^2 \pmod{p}$$

Now, we can use the fact that  $a^2$  can be written as  $(a+1)(a-1) + 1$ .

So,

$$a^{p+1} = (a+1)(a-1) + 1 \pmod{p}$$

Since we know that  $a-1$  is divisible by  $p$ , and that  $p$  is prime, we can say that  $(a+1)(a-1)$  is also divisible by  $p$ .

Therefore,

$$a^{p+1} = 1 \pmod{p}$$

which is what we wanted to prove.

By mathematical induction, we have proved that Fermat's Little Theorem holds true for all integers  $a$  where  $p$  is a prime.

It's important to note that Fermat's Little theorem is a fundamental theorem in number theory and it has many useful applications in cryptography, coding theory and some other fields.

Table 1: Hallucination from ChatGPT

## 2.3 Using an LLM to solve a word problem

There are three major approaches to using an LLM to solve mathematical word problems. They all take the problem, possibly supplemented with a prompt as input to the LLM; they differ in the nature of the output.

The most straightforward approach is simply to ask the LLM to produce the answer to the problem. This can work well with simple mathematics, though it breaks down when more complicated math is required.

### 2.3.1 Translating a word problem into executable code

A second approach is to have the LLM produce computer code that can then be executed to compute the answer. Generating code is, in fact, the most successful practical applications of LLMs to date. AI programs like OpenAI's Codex [5] and GitHub CoPilot [16] are used by professional programmers as assistants to help write code; some experts have enthusiastically reported that they save significant time and effort. These AIs are particularly effective at generating code of standard but non-obvious form, such as finding the names and arguments to library functions based on a verbal description; they thus save the programmer a tedious search through documentation. They are much less reliable in generating programs of significant length of complex structure.

Drori et al. [9] used Codex to generate Python code that would compute the answers to a collection of problems drawn from undergraduate math courses at MIT and Columbia. Their system involved the following steps. First, the problem was modified by a hand-crafted automated front end into a more standardized natural language form; generally this required only adding a few stock phrases such as "Use sympy". Second, the modified problem was then input to Codex, which output Python code. Finally, if the Python code did not give the correct answer, the system searched for similar examples, with solutions, in the training corpus. These were used to create a few-shot prompt, and the problem was attempted again with the new prompt.

Drori et al. claimed a success rate of 81%; however, for a number of reasons, that figure is highly misleading [7]. Two successful examples, chosen as illustrations in [9], are shown in table 2.

A few points about these examples should be noted. In both examples, the system ended up altering the original specification: in the first, it did not use the definition of the derivative; in the second, presumably the author of the problem did not intend that simulation should be used. In the first example, Codex is carrying out an almost trivial transformation from the language expression to a hand-constructed sympy library function. In the second example, the transformation from language to code is not at all trivial, but it is safe to say that Codex is here closely imitating some code that was in its training set.

An inherent limitation of this approach is that it carries out the translation in strict order: It first generates the code, then executes it. Presumably all the numerical calculations are done in the execution part. If so, then the code must work for all numerical values of the parameters. This can end up turning an easy problem into a well-nigh impossible one (or even an absolutely impossible one). Many commonsense problems are easy for some of the numerical parameters involved but extremely difficult for other values. If the problem is stated with values for which the problem is easy, then generating the code for the difficult cases is pointless, but may well be impossible to avoid in this approach. As one example of many, consider the problem of determining whether an object of complex shape fits through a cylindrical hole. If the object is small enough (e.g. its diameter is less than the diameter of the cross section of the hole), then trivially, it can fit through. If it is too large (e.g. it contains a sphere whose diameter is larger than the diameter of the cross section of the hole) then, trivially, it cannot fit through. Otherwise the calculation is more difficult, and for tight

**Question:** Find the derivative of the function using the definition of a derivative.

$f(x) = Y (x^{**2}-1)/(2*x-3)$ .

**Input:** Using sympy find the derivative of the function using the definition of a derivative.

$f(x) = Y (x^{**2}-1)/(2*x-3)$ .

**Program:**

```
import sympy as sp

x = sp.Symbols('x')

f = (x**2-1)/(2*x-3)

print(sp.diff(f,x))
```

**Answer:**  $2*x/(2*x-3)-2*(x^{**2}-1)/(2*x-3)^{**2}$

**Problem:** One generates a number  $x$  from a uniform distribution on the interval  $[0, \theta]$ . One decides to test  $H_0: \theta = 2$  against  $H_A: \theta \neq 2$  by rejecting  $H_0$  if  $x \leq 0.1$  or  $x \geq 1.9$ . Compute the probability of a type 1 error.

**Input:** One generates a number  $x$  from a uniform distribution on the interval  $[0, \theta]$ . One decides to test  $H_0: \theta = 2$  against  $H_A: \theta \neq 2$  by rejecting  $H_0$  if  $x \leq 0.1$  or  $x \geq 1.9$ . Using simulation compute the probability of a type 1 error.

**Program**

```
import numpy as np

def generate_x(theta)
    return np.random.uniform(0,theta)

def test_hypothesis(x)
    if x <= 0.1 or x >= 1.9:
        return TRUE
    else:
        return FALSE

def simulate_type_1_error(theta,n)
    count = 0
    for i in range(n)
        x = generate_x(theta)
        if test_hypothesis(x):
            count += 1
    return count / n

print(simulate_type_1_error(2,100000))
```

**Answer:** 0.1

Table 2: Translating word problems into Python code. From [9].



**Question** Natural language version: “When all the girls at Madeline’s school line up in rows of eight, there are seven left over. If instead they line up in rows of four, how many are left over? The final answer is 3.” Translate the natural language version to an Isabelle version

**PaLM output:**

```
theorem
  fixes n::nat
  assumes "n mod 8 = 7"
  shows "n mod 4 = 3"
```

**Question** Natural language version: "A map of the town that Annie, Barbara, and Charlie live in can be represented by the Cartesian plane. Annie is located at (6, -20) and Barbara is located at (1, 14). They agree to meet at the closest point that is equidistant from their current locations and walk upwards together to get to Charlie’s location at  $(\frac{7}{2}, 2)$ . How many units upward do Annie and Barbara walk together to get to Charlie? The final answer is 5." Translate the natural language version to an Isabelle version:

**Codex Output:**

```
theorem
  fixes a b c :: "real \ real"
  assumes h0 : "a = (6, -20)"
    and h1 : "b = (1, 14)"
    and h2 : "c = (7 / 2, 2)"
  shows "dist a c + dist b c = 5"
```

Table 3: **Examples of autoformalization** from [19]. The translation in the second example is incorrect.

fits, it can be *much* more complicated. Certainly the output code can be written so that, in simple cases, the code for difficult cases is not *executed*, but the point is, if the problem at hand is an easy case, it should not be necessary to even *generate* the code.

### 2.3.2 Translating word problems into logical specifications

A third approach to solving word problems is built around **autoformalization**, the process of automatically translating mathematical specifications given in natural language into a formal specification that then can be given to a formal verification system. Wu et al [19] used two LLMs, PaLM and Codex, to translate word problems into the formal notation used by the interactive proof assistant Isabelle. Since the training set for the LLMs presumably contains only a comparatively small quantity of Isabelle, they provided long few-shot prompts (different prompts for different areas of mathematics) with ten examples. They tested their systems on examples taken from benchmark collections **MATH** and **MiniF2F**, both of which contain mathematical competition problems. They found that the system created a perfect translation 25% of the time. Problems that require finding a specific value as an answer were reworded by adding, “The final answer is  $A$ ” where  $A$  is the correct answer at the end of the natural language formulation. (This is reasonable for testing the adequacy of the translation of  $\alpha$  into a formal notation; it would not, of course, be reasonable for testing the capacity of an AI to *solve* the problem.) Table 3 shows two examples of two problems translated into Isabelle. Note that the second example is in fact incorrect, though it is marked in [19] (p. 23) as “correct”.

Allan brought two balloons and Jake brought four balloons to the park. How many balloons did Allan and Jake have in the park?

Allan brought two balloons and Jake brought four balloons to the park. How many more balloons did Jake have than Allan in the park?

Jack had 142 pencils. Jack gave 31 pencils to Dorothy. How many pencils does Jack have now?

Dorothy had 142 pencils. Jack gave 31 pencils to Dorothy. How many pencils does Dorothy have now?

Jack had 142 pencils. Dorothy had 50 pencils. Jack gave 31 pencils to Dorothy. How many pencils does Jack have now?

Table 4: Examples from SVAMP dataset

## 2.4 Benchmarks

Standard collection of elementary math problems serve as benchmarks to compare the capacities of different AI systems with one another and with humans, to measure progress in the technology, and to serve as targets for research.

Numerous benchmark collections of word problems have been assembled [14]. The mathematical difficulty ranges from elementary school to International Mathematical Olympiad and introductory college math courses. The language use ranges from non-linguistic to moderately complex. Some require no world knowledge; others requires some combination of commonsense, common, expert, and encyclopedic knowledge. Question formats include open form question answering, fill-in-the-blank, multiple choice, and proof construction.

One high-quality collection of elementary CSPs is SVAMP (Simple Variations on Arithmetic Math Problem) dataset [17]. The problems in SVAMP involve only commonsense and common knowledge; the language is simple; the math required is either one or two arithmetic operations. The numbers involved are positive integers between 1 and 999. Each example is given in a number of paraphrases and variants, in order to avoid the kinds of artifacts discussed below in section 2.7. (AIs presented with multiple examples do not, ordinarily, carry information from one to the next; each is considered from scratch.<sup>6</sup>) Thus, the fact that there are multiple similar problems given in sequence does not affect behavior as it would with human experimental subjects.) Table 4 shows some example problems in SVAMP. There are 1000 such problems in SVAMP.

Another notable benchmark for math word problems, named LILA, is a large, structured collection, covering a range of linguistic complexity, mathematical difficulty, and requisite background knowledge [14]. The collection was assembled by combining twenty-three previously existing benchmarks collections and standardizing their formats. Table 5 shows the different categories of math problems included, with examples.

It is certainly disconcerting that three of the questions shown in table 5 are flawed. The answer to question 1 is in fact indeterminate; it can depend on the species of beetle (beetles vary in length from a fraction of a millimeter to 20 cm), the angle from which the insect is being viewed, whether the grasshopper has its wings spread or shut, and whether perceived area or perceived length is the measure in question. It is also not at all clear to me why the beetle/grasshopper problem is considered “basic arithmetic” whereas the much simpler pizza problem is put into the presumably more advanced category “Muldiv”. The answer “15” given for question 4 does not follow the format specified in the question, “What are the four consecutive integers?” As worded, Question 6 is both meaningless and ungrammatical; presumably “three” is a typo for “five”. Regrettably, problems

---

<sup>6</sup>Chatbots such as ChatGPT are exceptions to this general rule.

	Category	Example
1.	Basic math: addition, subtraction, fact-based QA etc.	<b>Question</b> If Jimbo is 484 feet away from a beetle and quarter of 827 feet from a grasshopper, which insect will seem bigger to him? "Option 1": beetle, "Option 2": grasshopper. <b>Answer:</b> Option 2
2.	Muldiv: multiplication, division along with addition, subtraction etc.	<b>Question:</b> Mrs. Hilt bought 2 pizzas. Each pizza had 8 slices. So, she had ____ total slices of pizza. <b>Answer:</b> 16
3.	Number theory: prime, power, negation, modulus and other operators etc.	<b>Question:</b> How many numbers are divisible by both 2 and 3 up to 300? <b>Answer:</b> 50
4.	Algebra: equations, functions, polynomials, series etc.	<b>Question:</b> The sum of the three smallest of four consecutive integers is 30 more than the largest integer. What are the four consecutive integers ? <b>Answer:</b> 15.0
5.	Geometry: triangles, polygons, 3D structures etc.	<b>Question:</b> A hall is 6 meters long and 6 meters wide. If the sum of the areas of the floor and the ceiling is equal to the sum of the areas of four walls, what is the volume of the hall (in cubic meters)? <b>Answer:</b> 108
6.	Statistics: binomial, divergence, mean, median, mode, variance etc.	<b>Question:</b> There are 11 boys and 10 girls in a class. If three students are selected at random, in how many ways that 3 girl and 2 boys are selected? [sic] <b>Answer:</b> 6600
7.	Calculus: differentiation, integration, gradient, series expansion etc	<b>Question:</b> Let $g(y) = 9*y^{**4} + 25*y^{**2} + 6$ . Let $s(d) = 1 - d^{**4}$ . Let $x(t) = -g(t) + 6*s(t)$ . What is the third derivative of $x(f)$ wrt $f$ ? <b>Answer:</b> -360*f
8.	Linear algebra: vectors, dot products, Eigen vectors, matrices etc.	<b>Question:</b> Problem: Convert the following matrix to reduced row echelon form $\begin{pmatrix} 7 & -2 & -10 & 4 \\ 5 & -10 & 2 & -7 \end{pmatrix}$ <b>Answer:</b> $\begin{pmatrix} 1 & 0 & \frac{-13}{10} & \frac{-13}{40} \\ 0 & 1 & \frac{9}{20} & \frac{69}{80} \end{pmatrix}$
9.	Advanced math: heuristics required along with probability, statistics, or algebra, Olympiad level problems	<b>Question:</b> Let $f(x) = 2^x$ . Find $\sqrt{f(f(f(f(1))))}$ . <b>Answer:</b> 256.

Table 5: Categories of math difficulty in LILA, together with examples [14]

[User] What is the fourth word in the sentence, “I am not what I am”?

[ChatGPT] The fourth word in the sentence, “I am not what I am” is “not.”

[User] The children arrived one by one to the birthday party: First Tome, then Mary, then Barbara, then Harry, then Ida. Who was the fourth child to arrive?

[ChatGPT] Barbara was the fourth child to arrive to the birthday party.

[User] How many words are in the sentence, “I am not what I am”?

[ChatGPT] There are 7 words in the sentence “I am not what I am.”

[User:] Barbara, Mary, and Harry had red balloons. Ted had a blue balloon. Sam also had a red balloon. How many children had red balloons?

[ChatGPT:] Barbara, Mary, Harry, and Sam had red balloons. In total, there were 5 children with red balloons.

Table 6: Some tests carried out by the author on ChatGPT, December 30, 2022. ChatGPT did sometimes answer these kinds of questions, but it often answered them incorrectly.

with flaws of these kinds, and much worse, are all too common in AI benchmarks [8]. One cannot have much confidence in measurements made with such flawed instruments.

#### 2.4.1 Untested abilities

Many of the benchmark collections used to test AI math abilities are taken from tests developed for human students — math course problems, IQ tests, SAT tests, Math Olympiad tests, and so on. However, these generally take for granted that the test-takers have certain basic abilities; this is a safe assumption for humans but not for AIs, whose abilities are very different [6]. Other benchmarks have been developed specifically for AIs, but these tend to focus on some abilities and omit others. Consequently, many categories of problems can go untested.

For example, apparently none of the benchmark sets for mathematical ability test an AI’s ability to count. Surely, an AI system that can solve (some) advanced math problems can count up to four? Not necessarily, as determined in informal test (table 6). Likewise, many forms of commonsense mathematical inference that draw on a basic of time, space, and physical reasoning are not tested in any existing AI benchmark.

## 2.5 State of the art with ML-based systems

In experiments carried out in 2021, the creators of SVAMP found that, for the three state-of-the-art AI programs they tested, the best performance on problems with one arithmetic operation was 51.9%; the best performance on problems that required combining two arithmetic operations was 33.9%. (These were achieved by different programs.)

A recent (October 2022) extensive study of the state of the art in LLM-based models for math word problems is reported in [14], using the LĪLA benchmark described above. A new AI for math word problems, named BHĀSKARA was created, by fine tuning GPT-Neo (a variant of GPT-3), trained on code as well as text) on a training set drawn from LĪLA. More precisely, two variants of BHĀSKARA were created. BHĀSKARA-A outputs the answer directly; BHĀSKARA-P outputs Python code that computes the answer.

Category	Best result	Best program	Worst result	Worst program
Basic math	0.79	Codex	0.61	Neo-P
Muldiv	0.69	Codex	0.16	BHĀSKARA-A
Number theory	0.47	Codex	0.12	BHĀSKARA-A
Algebra	0.60	Codex	0.20	BHĀSKARA-A
Geometry	0.30	BHĀSKARA-A	0	Codex
Statistics	0.95	BHĀSKARA-P	0.02	BHĀSKARA-A
Calculus	0.93	Codex	0.10	BHĀSKARA-A
Linear algebra	0.81	Neo-P	0.12	GPT-3
Advanced math	0.47	Codex	0.02	BHĀSKARA-A
Commonsense	0.79	GPT-3	0.62	Neo-P

Table 7: Results from tests on LĪLA. From [14], table 3. Results are for the “IID” regime. NEO-A, which never does well, is excluded from the listing of worst results

Six AI programs were then tested on a test set<sup>7</sup> drawn from LĪLA and disjoint from the training set: (i) GPT-3, with few-shot prompting; (ii) Codex, with few-shot prompting; (iii) Neo-A, i.e. GPT-Neo configured to output the answers; (iv) Neo-P, i.e. GPT-Neo configured to output Python code; (v) BHĀSKARA-A; (vi) BHĀSKARA-P. Table 7 shows the best and worst results on each category of math problem, as well as the results on problems considered “commonsense”.

Conspicuous in table 7 is that both absolute performance and relative performance between programs varies very much between categories. Each of the five systems is represented in the “Best” column; four of them are also in the “Worst” column and BHĀSKARA-P misses that by a hair, having scored 0.079 in the Geometry category. The wide divergence between program-based systems and answer-based systems in the Statistics category is easily explained: these problems are mostly easy if you can call library functions and very difficult otherwise.

## 2.6 General caveat

It cannot be assumed that, because an AI does well on a class of problems, it can do well on a seemingly easier problem. As discussed in section 2.4.1, the powerful AI ChatGPT, which does reasonably well on various simple math tasks, cannot reliably count to four. Likewise, it cannot be assumed that an AI that can carry out each of two components of a given task separately; as we will discuss below, LLMs in general do much worse on problems that involve two arithmetic operations than those that require one, both in word problems and in purely mathematical problems. The capacities and incapacities of these kinds of programs do not at all resemble those of human beings.

## 2.7 Artifacts

A persistent problem in corpus-based ML, in all its applications, is that the patterns that the AI finds do not actually reflect the fundamental characteristics of the problem, but rather superficial regularities in the training data, known as “artifacts”. For example, AIs trained over corpora of

<sup>7</sup>More precisely, they were each tested on two different test sets drawn LĪLA: one test set where the problems were drawn from the same original benchmark as those in the training set (the “independent identically distributed (IID) state) and a different test set where the problem of the same category were drawn from a different original benchmark (the “out of distribution” OOD state). Table 7 gives only the results for IID; these were generally, but not always, higher than OOD.

medical images have “learned” to associate diagnoses with features of the different imaging devices used at different hospitals rather than the actual content of the image. Artifacts are particularly likely to arise if a large data corpus is partitioned into a training set, used to train the ML system, and a test set, used to evaluate it; any superficial regularity in the corpus as a whole can be learned from the training set and then applied to the test set. In general, **avoiding artifacts in AI systems demands great care in constructing corpora and carrying out tests.**

These kinds of problems have, in fact, arisen in AI for math word problems. For example, in one **experiment [17]**, an AI trained on the **benchmark ASDiv-A** seems to have learned<sup>8</sup> that if the word “every” appears in a problem, it should multiply the two numbers involved, whereas if the word “each” appears, it should divide them. The AI therefore got the right answer on questions like “John delivered 3 letters at every house. If he delivered for 8 houses, how many letters did John deliver?” and “Sam made 8 dollars mowing lawns over the Summer. He charged 2 bucks for each lawn. How many lawns did he mow?”, but the wrong answer on questions like, “John delivered 3 letters at every house. He delivered 24 letters in all. How many houses did John visit to deliver letters?” and “Sam mowed 4 lawns over the Summer. If he charged 2 bucks for each lawn, how much did he earn?.” Apparently problems conforming to the rule were more common in the dataset than those violating the rule. Thus, if an AI is trained on a training set collected from this corpus, and then tested on a separate group of problems from the same corpus, it may well be able to use these kinds of regularities to achieve a fairly high success rate without at all understanding the actual meaning of the problem or its relation to the mathematical operations.

### 3 Further forms of common sense

In addition to non-mathematical common sense about the world, there are various forms of commonsense knowledge about mathematics itself and its relation to the world. To date, as far as I know, these have not been addressed in any AI systems. However, ultimately, in AIs that truly understand what mathematics is about they will have to be addressed.

One issue is the degree of precision that is reliable or meaningful in characterizing a real-world situation. Specifying numbers to meaningless degrees of precision, of course, is a common error in human writers, but one may hope that AIs can do better. It happens to be a pet peeve of mine. One example that come to hand: The accuracy numbers that I have given in table 7 are given to three-digit precision in [14]. Now, three-digit precision is not even warranted statistically; that would require a sample of size approximately one million. In table 7 I truncated these figures to two digits. However, as any kind of replicable and meaningful measure of the quality of current AI programs on math problems of various categories, probably one digit is too many; my guess is that the precision is more like a half a digit (two bits). Probably a program that scores 0.7 on a category is better than one that scores 0.2 and possibly it is better than one that scores 0.5; but my hunch is that would be risky to claim more than that. (The wide range between IID tests and OOD tests for many systems on many categories and especially the fact that OOD measures are sometimes substantially larger than IID values is a further indication that these numbers are not very meaningful.)

A second, very difficult, issue is deciding, in trying to answer a question about a real-world situation, how best to model it mathematically and how much detail to include. Consider, for instance, a pendulum swinging on a string. The simplest model, from a high school physics class, is the harmonic oscillator  $\ddot{\theta}(t) = -g\theta(t)$ . A more sophisticated model changes this to  $\ddot{\theta}(t) = -g\sin(\theta(t))$ . But under other circumstances, you might need to take into account: the three dimensional angle

---

<sup>8</sup>AI systems that used “deep learning” and similar learning techniques are opaque; characterizing what they have learned in terms of these kinds of rules is always approximate and relies on indirect evidence.

and rotation around a vertical axis (if it is swinging freely in 3D), the three dimensional shape of the bob (if it is not a sphere), the physical characteristics of the string (e.g. for a yo-yo, or if you plan to cut the string), the frictional forces (if you are using it for a grandfather clock and need to supply energy), the rotation of the earth and the latitude (for Foucault’s pendulum), the presence of other nearby massive objects (if, like Pierre Bouguer in 1735, you are camped near the base of a mountain and using a plumb line to determine the vertical) and so on [1]. The choice depends on the physical circumstances, the information you have available, the kind of question you are asking, and the precision and certainty that you require in the answer.

## 4 Relevance to mathematical research

As we have seen, current AI technology has serious limitations in its ability to engage with even very simple real-world mathematical problems posed in natural language. However, it is far from clear how relevant those limitations to an AI’s ability to do advanced mathematics. **One might naturally suppose that a technology that cannot reliably count up to four in a simple three-sentence narrative is not about to have much impact on groundbreaking mathematical research, but that is not a safe assumption.**

I doubt, for instance, that the limitations discussed here are very relevant to the prospects of building AIs that can construct formal proof of statements in pure mathematics given formal specifications. Human mathematicians may well draw on their commonsense knowledge in understanding advanced mathematics and searching for proofs, though the degree to which they do this is unknown. (It probably varies widely across individuals.) But that does not at all imply that AIs would need to do likewise. There are no AI programs that can reliably recognize and name concepts like a “fork” or a piece being “pinned” in chess, or that can read and discuss an article in a chess journal, but nonetheless they play very good chess.

However, there are important aspects of research mathematics where I expect that these limitations will be important. One is applications of all kinds. The mathematization of real-world domains and tasks generally conceals the commonsense reasoning involved and gives the illusion of a rigorously formal process, but it does not actually eliminate the need for commonsense understanding. This is true even in physics, at least on the experimental side. Understanding how the LIGO gravitational wave detector works involves a commonsense understanding of the components in addition to a lot of technical understanding; you cannot prove the correctness of the experimental device from first principles. In areas such as biology, medicine, and the social sciences and cognitive sciences, commonsense reasoning is even more unavoidable.

I would also conjecture, with much less confidence, that the limitations discussed here would raise challenges to building an AI that can read and understand (human-written) mathematical articles. I do not know of any systematic analysis of the cognitive processes required to read an extended proof, but it seems likely that they draw significantly on basic commonsense understanding. Again, no doubt it varies by field and by proof style. It seems likely that proofs that draw strongly on one’s visual sense and spatial, such as (at the elementary level) those in [15] or the beautiful proof of Fermat’s little theorem in [10] will be particularly difficult. *Writing* mathematical proofs that are human readers find intelligible and well-written may well be easier than reading them; it is often found, in AI, that generation is easier than comprehension.

## Acknowledgements

Thanks to Scott Aaronson, Yuling Gu, and Doug Hofstadter for helpful feedback.

## References

- [1] Beech, Martin. *The Pendulum Paradigm: Variations on a Theme and the Measure of Heaven and Earth*. Brown Walker Press, 2014.
- [2] Bobrow, Daniel G. “A question-answering system for high school algebra word problems.” *AFIPS*, 1964. 591-614.
- [3] Bommasani, Rishi, et al. “On the Opportunities and Risks of Foundation Models.” arXiv preprint arXiv:2018.07258. <https://arxiv.org/abs/2108.07258>
- [4] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. “Language models are few-shot learners.” *Advances in Neural Information Processing Systems* 33 (2020): 1877-1901. <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- [5] Chen, Mark et al. (2021). “Evaluating large language models trained on code.” arXiv preprint arXiv:2107.03374 <https://arxiv.org/abs/2107.03374>
- [6] Davis, Ernest. “Using human skills taxonomies and tests as measures of AI.” In Stuart Elliott (ed.) *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, 2021.
- [7] Davis, Ernest. “Limits of an AI program for solving college math problems.” arXiv preprint arXiv:2208.06906 <https://arxiv.org/abs/2107.03374>
- [8] Davis, Ernest. “Benchmarks for Automated Commonsense Reasoning: A Survey”. In preparation.
- [9] Drori, Iddo et al. (2022) “A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level.” *Proceedings of the National Academy of Sciences (PNAS)* 119(32), p.e2123433119. <https://www.pnas.org/doi/abs/10.1073/pnas.2123433119>
- [10] Ellenberg, Jordan. *Shape: The Hidden Geometry of Information, Biology, Strategy, Democracy, and Everything Else*. Penguin Press, 2021.
- [11] Gillioz, Anthony, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. “Overview of the Transformer-based Models for NLP Tasks.” In 2020 15th Conference on Computer Science and Information Systems (FedCSIS), pp. 179-183. IEEE, 2020. <https://ieeexplore.ieee.org/abstract/document/9222960>
- [12] Lenat, Douglas B., Mayank Prakash, and Mary Shepherd. “CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks.” *AI Magazine* 6, no. 4 (1985): 65-65.
- [13] Miao, Shen-yun, Chao-Chun Liang, and Keh-Yih Su. 2020. “A diverse corpus for evaluating and developing English math word problem solvers.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, <https://doi.org/10.18653/v1/2020.acl-main.92>
- [14] Mishra, Swaroop, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit et al. “LILA: A unified benchmark for mathematical reasoning.” arXiv preprint arXiv:2210.17517 (2022). <https://arxiv.org/abs/2210.17517>
- [15] Nelsen, Roger B., *Proofs Without Words: Exercises in Visual Thinking*. Mathematical Association of America, 1993.



- [16] Nguyen, Nhan, and Sarah Nadi. “An empirical evaluation of GitHub copilot’s code suggestions.” In *Proceedings of the 19th International Conference on Mining Software Repositories*, pp. 1-5. 2022.
- [17] Patel, Arkil, Satwik Bhattamishra, and Navin Goyal. “Are NLP Models really able to Solve Simple Math Word Problems?.” arXiv preprint arXiv:2103.07191 (2021). <https://arxiv.org/abs/2103.07191>
- [18] Qian, Jing, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. “Limitations of Language Models in Arithmetic and Symbolic Induction.” arXiv preprint arXiv:2208.05051 (2022). <https://arxiv.org/abs/2208.05051>
- [19] Wu, Yuhuai, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. “Autoformalization with large language models.” arXiv preprint arXiv:2205.12615 (2022). <https://arxiv.org/abs/2205.12615>