Math Natural Language Inference: This Should Be Easy!

Valeria de Paiva
Topos Institute
valeria@topos.institute

Qiyue Gao
UC San Diego
q3gao@ucsd.edu

Hai Hu
Shanghai Jiao Tong Univ.
hu.hai@outlook.com

Pavel Kovalev
Carnegie Mellon Univ.
pkovalev@andrew.cmu.edu

Yikang Liu
Shanghai Jiao Tong Univ.
yikangliu@sjtu.edu.cn

Lawrence S. Moss
Indiana Univ.
lmoss@iu.edu

Zhiheng Qian
Shanghai Jiao Tong Univ.
n1vnhil@sjtu.edu.cn

Abstract

We ask whether contemporary LLMs are able to perform natural language inference (NLI) tasks on mathematical texts. We call this the Math NLI problem. We construct a corpus of Math NLI pairs whose premises are from extant mathematical text and whose hypotheses and gold labels were provided by people with experience in both research-level mathematics and also in the NLI field. We also investigate the quality of corpora using the same premises but whose hypotheses are provided by LLMs themselves. We not only investigate the performance but also the inter-group consistency of the diverse group of LLMs. We have both positive and negative findings. Among our positive findings: in some settings, using a majority vote of LLMs is approximately equivalent to using human-labeled data in the Math NLI area. On the negative side: LLMs still struggle with mathematical language. They occasionally fail at even basic inferences. Current models are not as prone to hypothesis-only "inference" in our data the way the previous generation had been. In addition to our findings, we also provide our corpora as data to support future work on Math NLI. Our data can be found at https://github.com/MathNLI/MathNLI.

1 Introduction

We study natural language inference (NLI) tasks in the area of research-level mathematics. One might think that LLMs would do extremely well on this task. After all, what counts as an entailment or contradiction in everyday-language texts is often taken as a complicated version of what happens with mathematics. So we might expect purported mathematical inferences to be *easier* to evaluate than those in everyday language. And unlike language in the wild, the domain of mathematics is fairly well-defined. Facts, definitions, and logical reasoning play a large role in mathematical writing. Sentences ought to be precise and unambiguous.

However, there are complications with mathematical text from the start. The vocabulary may be unfamiliar to a generic audience: mathematical parlance can use daily words with new, unfamiliar meanings, e.g. 'ring', 'field', or even 'folklore'. On top of this, the use of visual elements such as symbols, equations, and diagrams, almost changes the very language of the text from plain text to a richer, multimodal language. The field lacks open-source resources such as dictionaries and glossaries for mathematical concepts. It is much harder to find a "person on the street" annotator of mathematics than of more common forms of text.

When confronted with the incredible solutions to mathematical-like problems that deep learning systems can offer nowadays (e.g., AlphaGeometry (Trinh et al., 2024)), it is difficult to believe that these systems cannot understand the basics of causality or of propositional reasoning used throughout mathematics. Nonetheless, when tested on these basics, the LLM-based systems still make very surprising (to humans) mistakes. Further, the fact that LLMs do not have a notion of self-consistency has been documented in many recent papers (Sedova et al., 2024; Kıcıman et al., 2024; Xu et al., 2024). But mathematics, as usually practiced, needs self-consistency. In a sense, it seems that sometimes the deep learning systems deserve an A+ in advanced problem solving but a B in the basics.

For all of these reasons, we could conclude, perhaps surprisingly, that the NLI task is not much easier when using LLMs to deal with mathematical text after all. In this paper, we shall see how precisely correct Math NLI using LLMs can be. We decided to experiment and build a corpus of NLI inference pairs, comparing the output of several LLMs on mathematical text.

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 179–188
November 8-9, 2025 ©2025 Association for Computational Linguistics

P (Premise)	H (Hypothesis)	Label
A notion of central importance in categorical topology is that of topological functor.	Topological functor is a notion of categorical topology.	E
The problem of relating a factorization system to a pointed endofunctor is considered.	The problem of relating a factorization system to a pointed endofunctor is not discussed.	C
A notion of central importance in categorical topology is that of topological functor.	There are many notions of central importance in categorical topology.	N

Table 1: Examples in human-created seed Math NLI corpus.

1.1 Research questions

Our big question: Can LLMs be reliable constructors and annotators of Math NLI corpora? We address this by asking and answering some secondary questions: (a) How well do LLMs perform on a Math NLI corpus annotated by mathematicians? (b) Are there common features to the errors which they make? (c) How good is a Math NLI corpus annotated entirely by LLMs? (d) Are LLMs more unanimous on human-written corpora or on corpora generated by LLMs themselves?

1.2 Goal, plan and structure of the paper

The "deliverables" of this paper are two corpora for Math NLI: one written by humans and the other by GPT. These are not benchmarks. But we believe that they will help others who work on this topic.

Equally important, this paper details what we have learned about Math NLI from several years of work, including work that did not turn out as well as we had hoped. Overall, our goal is to make some points about Math NLI which we believe have not been made elsewhere, based on data and examples which we have collected. The plan of the paper is to tell the story of this work.

2 Math NLI seed corpus

2.1 Creation of a seed set of pairs

Our first experiment used a corpus of abstracts of articles in the journal *Theory and Applications of Categories* (TAC) developed in (Collard et al., 2022)[1]. This corpus has some 3K sentences, but 432 were singled out as 'Goldilocks-like sentences': not too short, not too long, and with little or no LaTeX markup. Then we chose 31 of these sentences, and for each sentence S in this set, three of our team members were asked to write a sentence entailed by S, a sentence contradicting S, and a sentence neutral with respect to S. (So we had the "gold labels" by construction. But as we found repeatedly, getting consistent data from humans is difficult, even about mathematical texts.) The team members were told to produce grammatical sentences that did not depend on factual knowledge about the mathematics in the original TAC sentence and that tried to introduce as few new facts as possible. It is impossible to do this perfectly, but the team members strove to do so. We had three people, three labels, and 31 starting sentences. Hence we had $3 \times 3 \times 31 = 279$ pairs, equally divided with E, C, and N labels.

We aimed to fulfill the following conditions as much as possible:

1. Inferences should be uncontroversial. We want inferences which most mathematicians would take to be "immediate."

2. We treat mathematical concepts as black boxes. (Inference should depend as little as possible on the background mathematical knowledge of the assessor.)

3. We avoid "dangling references", pronouns (it, they) or demonstratives (this, that, here, there) without clear antecedents. In general, we tried to avoid all of the problematic issues in natural language semantics.

Table 1 shows some examples of human-created hypotheses and their labels.

Having constructed our seed set of 279 pairs we used a collection of LLMs to evaluate it, as shown in Table 2. This led to the realization that not only did human creators disagree with each other, also the rate of unanimity between machines was not very stable. In particular, we discovered some 20 pairs with contradictory evaluations between machines and humans. We called these the *red pairs*, as they deserved further attention. We explain our process of evaluation, the LLMs used, and our set

[1] Available at https://github.com/ToposInstitute/tac-corpus.

Abbr.	Model
GPT4	GPT-4[2]
L2	Llama 2 (Touvron et al., 2023) 70B
L3	Llama 3 (Grattafiori et al., 2024) 70B
C3	Claude 3 `https://claude.ai/` unknown
Mistral	Mistral-large
L3.1	Llama-3.1-70B-Instruct (Grattafiori et al., 2024)
Q2	Qwen2-72B-Instruct (Bai et al., 2023)
Mixtral	Mixtral-8x22B-Instruct-v0.1 (Jiang et al., 2024)
DS	deepseek-llm-67b-chat (Bi et al., 2024)
Ge2	gemma-2-27b-it (Team et al., 2024)

Table 2: LLMs used in Exp. 1. Top: Group 1: five initial LLMs; Bottom: Group 2: five later LLMs.

up in the next section, but we discuss briefly the red pairs now.

2.2 Red Pairs

Our three mathematically-trained group members tried to analyze the kinds of mistakes LLMs were making in these pairs. We discovered a few patterns of problematic or flawed reasoning:

Ignored context. Sometimes a specific context was mentioned, for instance

- *P: In the nilpotent case, this nerve is known to be a Kan complex.*
  H: This nerve is not known to be a Kan complex.

but it looks like the LLMs discarded the specific context (*the nilpotent case*) and compared the matrix sentences – in the example above this leads to a contradiction – instead of a neutral label. This is similar to the problems with modal and counterfactual reasoning discussed in (Holliday et al., 2024).

Vague quantifiers. We also have problems with vague predicates like *numerous, few, many*, where humans could also disagree amongst themselves: one example from the 'red pairs' set is

- *P: We worked through numerous examples to demonstrate the power of these notions.*
  H: We worked through two examples to demonstrate the power of these notions.

The mathematicians agreed that *numerous examples* should entail *two examples*, but LLMs did not.

Lexical ambiguity. There is lexical ambiguity, for example, with the verb "resemble" which might mean "is almost equal" (for some humans) or "it looks similar to something else, but it is not the same as" – a reason why we might have humans saying both contradiction or entailment in the example:

- *P: The axioms resemble those for monoidal Abelian categories.*
  H: The axioms are the ones of monoidal Abelian categories.

Note that the ambiguity which we call "lexical" here might also be called "pragmatic" because the issue is whether the use of "resemble" here carries the Gricean implicature that if an object A resembles an object B, then A is not, strictly speaking, B at all.

Naming of math entities. There is a problem with naming mathematical entities, e.g. "group B" vs. "group C" if this is only used as a generic name, as an α-variant, then the difference between B and C doesn't matter. But many times we are talking about different groups.

Unknown math concepts. Sometimes one really must know the concepts involved. For example, for the pair

- *P: This paper proposes a recursive definition of V-n-categories and their morphisms.*
  H: This paper proposes a definition of V-categories.

if we know that 'V-n-categories' are 'V-categories', then we can decide on entailment. But how do we know that? The mathematician is at liberty to create concepts and name them in strange ways. For instance a "skew monoidal category" is not a "monoidal category", only an 'almost' monoidal category.

3 Evaluating LLMs on the seed corpus

In our first experiment, we harness LLMs to evaluate the seed corpus.

3.1 Method

The seed corpus was originally judged by five LLMs, the top ones in Table 2. We used the prompt shown in Appendix C. When 4 or 5 LLMs disagreed with the human annotation, we discussed the pair again, throwing it out if it was considered "controversial" by the mathematicians in our group.

We use API services from `together.ai` to query the LLMs, using a script to extract E/C/N judgments from each model's explanation. The algorithm used is simple: it counts the occurrences of a few keywords in the first sentence without semantical analysis. (It works well if the model gives the answer directly.) However, this algorithm can fail. For example, when the model does not follow the instructions strictly we may end up with a pair that is neither E nor C nor N, and as usual in NLI we take N as a catch-all for "not E and not C."

3.2 Results

Performance of 10 LLMs on the seed MathNLI corpus is shown in Table 3, with their confusion matrices shown in Table 4.

Table 3 presents the precision, recall, f1-score and accuracy for 10 LLMs. The overall accuracy is medium to high, ranging from 71% to 91%, suggesting that in general, the LLMs we tested can perform category-theory-related mathematical inference to a certain degree. We note that the first group of LLMs (to the left of the table) are not particularly better than the second group (on the right). This might reflect the fact that the first group were closed-source, while Group 2's models were open-source. The first group has two closed source models: Claude 3 and GPT-4; the others are open source. In particular, Claude 3 seems to still be better than the open-source LLMs, but perhaps more runs are necessary to confirm this.

A main message from Table 4 is that most models struggle with *neutral pairs*, mistakenly categorizing them either as entailment pairs or contradictory pairs. For instance, Llama-3 is particularly bad in that it labels as many as 35% of neutral pairs as contradictions; only 48% of the neutral pairs are correctly classified. Claude 3 is the best in labeling N pairs, with an accuracy of 84.9% for them. On the contrary, most models perform very well on C and E pairs. GPT-4, Llama 3 and Qwen2 correctly labeled more than 90% of the C and E pairs. In fact, C pairs are the easiest for all models, except Llama 2, with most models achieving accuracy greater than 90%. Furthermore, models seldom confuse C and E pairs. For eight out of the ten LLMs, C pairs are never categorized as E pairs.

Only one pair in one model (Gemma2) is classified as C by the machines and E by humans:

- *P: Both of them generalize the concept of algebra on a monad T.*
  H: The concept of algebra on a monad T is more special than both of them.

Note that this pair does not satisfy our criteria of explicit references only. The pair is fairly controversial, as well. All LLMs label it as contradictory, but mathematicians tend to think that generalizing and specializing are antonyms. So whatever "both of them" are, if they are a generalization of the concept of algebra of a monad (as claimed by the premise) then "algebra of a monad" will more specialized than them.

Concerning the Group 1 models: out of 279 samples, there is at least one model that agrees with the human annotator in 271 samples. Hence, there are 8 pairs where none of the 5 initial models agrees with the human label. These eight pairs are recalled in Appendix A. The examples are telling as they point out patterns of reasoning that might be difficult for humans as well. For instance:

- *P: Using these ideas, we also prove that magnetic monopoles form an abelian group.*
  H: Using these ideas, we also prove that monopoles form an abelian group.

Clearly a mathematician would gather that 'magnetic monopoles' form an abelian group, but nothing has been said about non-magnetic monopoles. So neutral is much more reasonable than 'entailment'. (More on this is in the appendix A).

Table 5 discusses unanimity between LLMs. As before we consider two groups of models. Our initial LLMs are unanimous in 163 of the pairs (58.4%). Of these 163, in 155 of the cases, the models' agreed-upon label matches the human annotations. And in 271 of the 279 pairs (including ones where the models were not unanimous), at least one model agreed with the human label. This explains the upper row of the table, and the lower row is similar.

Notice that for the more recent LLMs, unanimity goes up from 58.4% to 68.1%. We do not have a good explanation of this.

4 Using LLMs to generate a MathNLI corpus

4.1 Generation using GPT-4

Our second experiment asked GPT-4 to generate Entailment, Contradiction, and Neutral hypotheses from the Goldilocks sentences in the TAC corpus, resulting in 1157 pairs. The prompt we used is shown below:

		GPT4	L2	L3	**C3**	Mistral	L3.1	Q2	Mixtral	DS	Ge2
C	p	82.9	90.5	70.9	91.8	79.8	88.8	87.3	75.4	**92.7**	85.1
	r	**98.9**	61.3	96.8	95.7	97.8	93.5	95.7	**98.9**	81.7	92.5
	f1	90.2	73.1	81.8	**93.7**	87.9	91.1	91.3	85.6	86.9	88.7
E	p	90.1	73.9	85.4	**93.5**	89.8	83.8	80.8	86.3	82.0	82.8
	r	**97.8**	88.2	94.6	92.5	84.9	89.2	90.3	88.2	78.5	82.8
	f1	**93.8**	80.4	89.8	93.0	87.3	86.5	85.3	87.2	80.2	82.8
N	p	**95.5**	56.2	91.8	87.8	81.8	81.7	84.7	85.5	67.6	75.3
	r	68.8	63.4	48.4	**84.9**	67.7	72.0	65.6	57.0	78.5	68.8
	f1	80.0	59.6	63.4	**86.3**	74.1	76.6	73.9	68.4	72.6	71.9
acc		88.5	71.0	79.9	**91.0**	83.5	84.9	83.9	81.4	79.6	81.4
avg	p	89.5	73.5	82.7	**91.0**	83.8	84.8	84.2	82.4	80.8	81.1
	r	88.5	71.0	79.9	**91.0**	83.5	84.9	83.9	81.4	79.6	81.4
	f1	88.0	71.0	78.3	**91.0**	83.1	84.7	83.5	80.4	79.9	81.1

Table 3: Results of 10 LLMs on the seed MathNLI corpus (precision/recall/F1 per class; accuracy and macro averages). *Closed-source* models are marked with the lavender header; green cells denote row-best scores.

Generate “Entailment”, “Contradiction”, “Neutral” hypothesis of a given sentence.
Here are some examples: `[example_script]`
Sentence: `[context]`

The temperature for the generation was 1. GPT-4 was a good generator of pairs, as we shall see below. But it was not consistent with itself. If it created a pair nominally to be E it could later judge it N or even C. As we see in Table 6, 41.4% of the pairs which GPT-4 created to be neutral it later claims as entailments.

4.2 Checking of a subset, using both humans and LLMs

We chose 89 pairs to conduct manual evaluation and distributed these among the mathematicians of the group. This gave us a set of 89 GPT-4-created/human evaluated pairs. These 89 pairs were also evaluated using GPT-4, Llama 2, Llama 3 and Claude 3, in the first instance. Our mathematicians agree with each other in 80 of the 89 pairs. They agree with 74 (83%) of the GPT-generated labels.

5 Evaluating LLMs on GPT-generated MathNLI corpus

Next, we had the 4 models in Group 1 and 5 models in Group 2 label the 89 pairs. The results are shown in Table 6. The models in group 1 show unanimous agreement in 57 of the pairs (64%), while the models in group 2 do so in 65 (73%). In group 1, for 50 of these 65 pairs (87%), their unanimous label agrees with human labels; while the agreement for group 2 is 57 pairs (88%). Here is our conclusion from this experiment: If we take the unanimous labels from the group 2 models to simply *be* the gold label, then this label is the same as the human label 88% of the time.

The evaluation results on the GPT-generated corpus using the GPT-generated label as the true label are shown in Table 8, with the confusion matrices presented in Table 7. The overall accuracy of LLMs varies between 59.6% and 86.5%, which is relatively lower than the accuracy on the seed corpus.

Our analysis reveals that while the E and C pairs generated by GPT show a certain level of consistency relative to our seed pairs, N pairs are frequently misclassified as E. (This finding echoes what we saw in our previous experiment, but there the pairs were human-generated.) Surprisingly, Llama 2 classifies 75.9% of N pairs as E. Among all evaluated models, Mixtral showed the least susceptibility to this issue, maintaining the highest accuracy of 76.0%. Although its performance on the seed corpus was not outstanding, Mixtral achieved the highest overall accuracy of 86.5% on the GPT-generated corpus.

	(a) GPT4			(b) Llama2			(c) Llama3			(d) Claude3			(e) Mistral		
Gold↓	C	E	N	C	E	N	C	E	N	C	E	N	C	E	N
C	98.9	.0	1.1	61.3	1.1	37.6	96.8	.0	3.2	95.7	.0	4.3	97.8	.0	2.2
E	.0	97.8	2.2	.0	88.2	11.8	4.3	94.6	1.1	.0	92.5	7.5	2.2	84.9	12.9
N	20.4	10.8	68.8	6.5	30.1	63.4	35.5	16.1	48.4	8.6	6.5	84.9	22.6	9.7	67.7

	(f) Llama3.1			(g) Qwen2			(h) Mixtral			(i) DeepSeek			(j) Gemma2		
Gold↓	C	E	N	C	E	N	C	E	N	C	E	N	C	E	N
C	93.5	.0	6.5	95.7	.0	4.3	98.9	.0	1.1	81.7	.0	18.3	92.5	1.1	6.5
E	1.1	89.2	9.7	2.2	90.3	7.5	3.2	88.2	8.6	2.2	78.5	19.4	1.1	82.8	16.1
N	10.8	17.2	72.0	12.0	21.7	66.3	29.0	14.0	57.0	4.3	17.2	78.5	15.1	16.1	68.8

Table 4: Confusion Matrices Comparison for 10 LLMs on the seed MathNLI corpus. Darker green denotes higher scores, while orange shades denote low scores; for both colors, paler shades represent smaller values in the corresponding range.

	unanimous	some agree w/ human	agrees w/ a human
models in group 1	163 (=58.4%)	271 (=97.1%)	155 (=55.6%)
models in group 2	190 (=68.1%)	266 (=95.3%)	178 (=63.8%)

Table 5: Agreement by LLMs on the seed corpus.

Furthermore, it was observed that LLMs tend to identify C pairs within the GPT-generated corpus more accurately than they do within the seed corpus; 8 out of 9 models achieved an F1 score of over 90%. Notably, Gemma2 successfully detected all the C samples in the GPT-generated corpus.

6 Sanity checks

For previous NLI datasets such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), studies have shown that some of the systems which detect "entailment" were working on artifacts of the datasets, instead of doing any logical reasoning (e.g., Poliak et al., 2018). In this section we provide some sanity checks on our corpus.

6.1 Does every assertion entail itself?

We are trying to make reasoning as clear as possible for humans and systems, and in some sense as simple as possible. We thus propose the easiest logical deduction that we expect LLMs (and humans) to make proficiently. Taking a cue from propositional logic, this would be that every proposition A entails itself, $A \to A$. Informally, if one accepts a sentence S as an assumption, then one should draw S as a trivial conclusion. We want to know whether a computational system "knows" this without being told it explicitly. For an LLM, we want to know if our prompt together with the model's training is enough to carry out what we take to be a very easy inference. To see how much noise LLMs could introduce, we ask them whether every one of a set of mathematical assertions entailed itself. It was not very surprising to discover that for one of 180 such pairs, one LLM missed this assertion $A \to A$.

- *P: We do not include any guiding examples.*
  H: We do not include any guiding examples.

This pair was judged N instead of E. But the explanation for the reasoning of the LLM is fascinating:

> Explanation: In this case, the premise and the hypothesis are identical statements. Since they both convey the same information, there is no logical inference or contradiction between them. The hypothesis is neither entailed by the premise nor does it contradict the premise. Therefore, the hypothesis is neutral to the premise.

The first sentence is correct, but immediately following that the model blunders.

The experiment was repeated with many more pairs, and only 3 in 2312 pairs had the same problem. This is a very small number of errors, but this is arguably the easiest inference one can ever make. It is odd that a system which can make deep, complicated mathematical inferences, can get a simple inference like 'A implies A' wrong.

	unanimous	agree w/ at least 1 human	agree w/ all human
human annotator	/	/	80 (= 89.9%)
GPT generator	/	74 (= 83.1%)	65 (= 73.0%)
models in group 1	57 (= 64.0%)	50 (= 56.2%)	43 (= 48.3 %)
models in group 2	65 (= 73.0%)	57 (= 64.0%)	50 (= 56.2%)

Table 6: Experiment 3 Result: total 89 pairs generated by GPT-4

(a) GPT-4

Gold↓	C	E	N
C	96.7	.0	3.3
E	.0	96.7	3.3
N	.0	41.4	58.6

(b) Llama 2

C	E	N
53.3	.0	46.7
.0	100.0	.0
.0	75.9	24.1

(c) Llama 3

C	E	N
96.7	.0	3.3
.0	100.0	.0
.0	51.7	48.3

(d) Claude 3

C	E	N
93.3	3.3	3.3
.0	93.3	6.7
3.4	34.5	62.1

(e) Llama 3.1

C	E	N
93.3	.0	6.7
.0	100.0	.0
3.4	55.2	41.4

(f) Qwen2

C	E	N
93.3	.0	6.7
.0	100.0	.0
3.4	34.5	62.1

(g) Mixtral

C	E	N
96.7	.0	3.3
.0	96.7	3.3
3.4	31.0	65.5

(h) Deepseek

C	E	N
83.3	.0	16.7
.0	90.0	10.0
.0	31.0	69.0

(i) Gemma2

C	E	N
100.0	.0	.0
.0	90.0	10.0
.0	34.5	65.5

Table 7: Confusion Matrices on GPT-generated Corpus

Previous work such as Xu et al. (2024) tries to catalog the kinds of mistakes that LLMs are known to make. They suggest that "to uncover the logical flaws of LLMs, problematic cases will be attributed to five error types from two dimensions, i.e., *evidence selection process* and *reasoning process*." The example above seems clearly a reasoning process kind of error, as the LLM is very clear that both the hypothesis and the premise are 'identical statements'. But from that it concludes that the hypothesis is **not** entailed by the premise.

6.2 Contradictions must be symmetric

Most humans would agree that if a sentence A is contradictory with a sentence B, then sentence B is contradictory with A. That is, being contradictory is a symmetric property. Work in (Kalouli et al., 2017) showed that the humans annotating the corpus SICK did not realize when they had non-symmetric contradictions. We hence checked whether LLMs evaluated contradictions symmetrically in the GPT-generated corpus. This small experiment showed that out of 495 pairs (5 times 93 contradiction pairs), 49 contradictions were not symmetric. This is not as bad as humans did in the paper above, but it still shows a lack of consistency.

6.3 Entailment requires premises and hypothesis

The premise-only work in NLI points to the fact that the labels E, C, and N could be accurately determined without any premise, simply using the hypothesis. To make sure that our corpus does not have the same problem, we run an experiment using a dummy true premise, say, "Right adjoints preserve limits".

We substitute this sentence for the premise in all 279 pairs, and evaluate the new pairs using the Group 2 Models. These models do not suffer from the same problems that earlier ones did; all four essentially classified all of the hypotheses as N, which is correct.

7 Final remarks

We find it useful to discuss our work by seeing how it aligns with the perceptive conclusions drawn by (Madaan et al., 2024).[3] We agree that evaluating models on NLI tasks is still relevant. For Math NLI, we do not find models to be saturated. This contrasts with ordinary language NLI (ONLI). We also confirm their finding that "while the similarity of model distributions with human label distributions increases with scale, it is still much higher

[3] We would compare with other sources, but (Madaan et al., 2024) seems to be the most relevant contemporary paper on this topic.

		GPT4	L2	L3	C3	L3.1	Q2	Mixtral	DS	Ge2
C	precision	100.0	100.0	100.0	96.6	96.6	96.6	96.7	100.0	100.0
	recall	96.7	53.3	96.7	93.3	93.3	93.3	96.7	83.3	100.0
	f1-score	98.3	69.6	98.3	94.9	94.9	94.9	96.7	90.9	**100.0**
E	precision	70.7	57.7	66.7	71.8	65.2	75.0	76.3	75.0	73.0
	recall	96.7	100.0	100.0	93.3	100.0	100.0	96.7	90.0	90.0
	f1-score	81.7	73.2	80.0	81.2	78.9	**85.7**	85.3	81.8	80.6
N	precision	89.5	33.3	93.3	85.7	85.7	90.0	90.5	71.4	86.4
	recall	58.6	24.1	48.3	62.1	41.4	62.1	65.5	69.0	65.5
	f1-score	70.8	28.0	63.6	72.0	55.8	73.5	**76.0**	70.2	74.5
acc		84.3	59.6	82.0	83.1	78.7	85.4	**86.5**	80.9	85.4
avg	precision	86.7	64.0	86.6	84.7	82.5	87.2	87.8	82.3	86.4
	recall	84.3	59.6	82.0	83.1	78.7	85.4	86.5	80.9	85.4
	f1-score	83.8	57.2	80.8	82.8	76.8	84.8	**86.1**	81.1	85.2

Table 8: Results of LLMs on GPT-generated Corpus.

Model	E	N	C
L3.1	.039	.961	.0
Q2	.004	.992	.004
Mixtral	.0	1.00	.0
Ge2	.004	.996	.004

Table 9: Result of Hypothesis only Baseline

than the similarity between two populations of humans, making it a potentially interesting statistic to consider." We have found that models show less of a distribution of labels than humans. We mean that the models are closer to unanimity than humans. Finally, they note a certain "subjectivity": "examples with 'incorrect' predictions are rarely in fact incorrect; most concern questions on which humans may disagree as well." And just as they point out, "The ground truth labels for NLP benchmarks are often decided according to the majority label by human annotators. This simplifies the data annotation process while also making the evaluation easier. However, several previous studies have noted that human disagreements in annotations for NLP datasets reflect the lack of a single ground truth label, rather than noise in the annotation process." Even in mathematical texts, there is room for disagreements between experts.

7.1 Conclusion and future directions

This paper investigates the performance of Large Language Models (LLMs) on Natural Language Inference (NLI) tasks within the domain of research-level mathematics. We explore the complexities of mathematical language compared to everyday language and evaluate LLMs' ability to handle mathematical inferences, noting some surprising strengths and weaknesses.

Contrary to what we initially assumed Math NLI is not much easier than ONLI for LLMs. Challenges include unfamiliar vocabulary (e.g., 'ring', 'field', 'comonad'), multimodal elements like symbols and equations, lack of open-source mathematical resources, and the difficulty of finding expert human annotators.

LLMs show paradoxical performance on math tasks: despite exhibiting impressive capabilities in complex mathematical-like problem-solving, LLMs surprisingly struggle with basic logical reasoning and NLI tasks in mathematics. We have documented issues with self-consistency, which is crucial in mathematics. A sanity check testing whether LLMs correctly identify that a statement entails itself ($A \to A$) revealed a very small number of errors, but the explanations for these errors showed a fundamental reasoning flaw.

Post-GPT LLMs avoid some issues that plagued earlier systems. For example, we expected lexical ambiguity involving math words to cause LLMs to stumble, as in mixing up "stack" (a mathematical concept) with ordinary "stack" (pile). They did not do so.

We provide two corpora intended to support further research in the Math NLI area. One had hypotheses which we wrote ourselves, and the other had LLMs write the hypotheses. We believe that these corpora will help newcomers to this attractive area. And our results give some idea of what is reasonable to expect from this area in the next years.

Further directions include combining our work with theorem provers or other symbolic methods, tests of similarity as opposed to inference, and interactions of our work with running systems in the Math NLI area. We also leave to future work an analysis of the CoT explanations provided by LLMs. For us, this would be especially interesting regarding the red pairs (see Section 2.2).

Limitations

We did not fine-tune to mathematical text the LLMs we use. We also only ran things once. All of our mathematical work was centered on the relatively special area of category theory, since that was the source of our premise pairs. We do not expect significant differences when we pivot to other branches of mathematics.

A more problematic limitation is that from the outset we concentrated on a relatively limited kind of sentence. That is, we aimed for sentences which did not manifest interesting but semantically problematic phenomena like ellipses, temporal reference, poetic language, and the like. In a sense, we aimed for sentences that were close to what one could formalize in standard logic. This concentration was behind our initial choice of 432 sentences from the TAC corpus. We also wanted sentences which were not too short, not too long, and with little or no LaTeX. This also is a limitation.

Acknowledgments

The work of Hai Hu is funded by Shanghai Pujiang Program (22PJC063).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*.

Jacob Collard, Valeria de Paiva, Brendan Fong, and Eswaran Subrahmanian. 2022. Extracting mathematical concepts from text. *Preprint*, arXiv:2208.13830.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wesley H. Holliday, Matthew Mandelkern, and Cedegao E. Zhang. 2024. Conditional and modal reasoning in large language models. *Preprint*, arXiv:2401.17169.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Aikaterini-Lida Kalouli, Valeria de Paiva, and Livy Real. 2017. Correcting contradictions. In *Proceedings of the Computing Natural Language Inference Workshop*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Preprint*, arXiv:2305.00050.

Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank, and Dieuwke Hupkes. 2024. Lost in inference: Rediscovering the role of natural language inference for large language models. *arXiv preprint arXiv:2411.14103*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Anastasiia Sedova, Robert Litschko, Diego Frassinelli, Benjamin Roth, and Barbara Plank. 2024. To know or not to know? analyzing self-consistency of large language models under ambiguity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17203–17217, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625:476 – 482.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2024. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. *Preprint*, arXiv:2306.09841.

A On the LLMs used in this work

See Table 2. We used Qwen2-72B-Instruct, which was released in June 2024. According to the Qwen2 Technical Report, this model outperformed Llama3-70B-Instruct on most benchmarks, including mathematical benchmarks such as GSM8K and MATH.

B Disagreements between models and humans in the seed corpus

1. *P: Using these ideas, we also prove that magnetic monopoles form an abelian group.* H: Using these ideas, we also prove that monopoles form an abelian group.

   Humans say the label is N, as it's only for magnetic monopoles that we have the abelian group. Machines say entailment E, but no mathematician would state the weaker result, if they could prove it without the extra hypothesis.

2. *P: The problem of relating a factorization system to a pointed endofunctor is considered.* H: A pointed endofunctor cannot be related to a factorization system.

   Humans disagree: some say contradiction C, others say N

3. *P: This paper introduces the notions of vector field and flow on a general differentiable stack.* H: This paper generalizes the notions of vector field and flow on a stack.

4. *P: We define eventually cyclic Boolean flows and the eventually cyclic spectrum of a Boolean flow.* H: The definition of the eventually cyclic spectrum of a Boolean flow uses the definition of eventually cyclic Boolean flows.

5. *P: The axioms resemble those for monoidal Abelian categories with the addition of an involutive functor.* H: The axioms are the ones of monoidal Abelian categories.

6. *P: The category of Set-valued presheaves on a small category B is a topos.* H: The category of Set-valued presheaves on a small category C is a topos.

7. *P: The category of Set-valued presheaves on a small category B is a topos.* H: There exists a small category C such that the category of Set-valued presheaves on C is not a topos.

8. *P: Various concerns suggest looking for internal co-categories in categories with strong logical structure.* H: We suggest looking for internal co-categories.

C Seed corpus prompt

Here is the prompt which we used on the seed corpus:

`[Begin prompt head]`
Suppose you are a logician. Your job is to determine the inference relation between the premise and the hypothesis. There could be three answers: (1) the hypothesis is entailed by the premise; (2) the hypothesis is neutral to the premise; (3) the hypothesis contradicts the premise. Please first tell me your answer and explain why this is your answer.
`[End prompt head]`
Premise: `[Premise]`
Hypothesis: `[Hypothesis]`