GSM-Symbolic 复现成果

摘要

复现论文是一篇探究性论文，主要是探究大语言模型中数学推理的局限性，他们对另一个文章的观点表示赞同，即模型并非进行形式推理而是模式匹配。他们提出了一个数据集，这个数据集是 GSM8K 数据集的变体，具体变体方式包括更改名字、数据、从句数量和增加无效信息，他们通过这些变体数据集来测试模型性能的变化。其中 GSM8K 是一个小学数学推理数据集，包含带有详细解答的简单数学问题。GSM8K 数据集被广泛运用在各个数学推理相关论文中。作为评判模型在小学数学问题上回答准确率的一个基本数据集，不过这导致了数据集可能存在数据污染的风险。复现论文利用测试结果的方差参数来评估模型是否有数学推理能力。最后他们证明了大语言模型在数学推理的脆弱性。我的复现结果符合复现论文的结果，未来希望能够更加完整地复现此篇论文。

关键词：大语言模型；数学推理

1 引言

现如今，大语言模型 (LLMs) 在诸多方面表现出色。而它们在数学推理领域上的能力和潜力引起了研究人员和从业者的广泛关注。数学推理需要强大的逻辑推理能力，但是 LLMs 表现出的更多是模式匹配而非形式推理 [4]。LLMs 是否能够进行真正的逻辑推理现在仍然是一个重要的研究焦点。

数学推理是一项关键的认知技能，如果解决了该问题，LLMs 将会进一步发展。GSM8K 数据集 [1] 是一个 8.5K 的小学数学难度的数学推理数据集，每一个实例都包含一个问题、一个解题的过程和一个答案。GSM8K 数据集被广泛运用在各个数学推理相关论文中，作为评判模型在小学数学问题上回答准确率的一个基本数据集。不过，复现论文 [5] 指出，GSM8K 的流行和广泛应用可能会带来数据污染的风险。此外，GSM8K 的静态不变性无法充分体现出 LLMs 的逻辑推理能力。

为了能够有效解决这种局限性，复现论文提出了一个新的数据集 GSM-Symbolic，它是基于 GSM8K 的变体。如图1所示，他们利用符号模板生成问题的变体，通过改变原有问题的名字和数据并且通过一定的方式来保证问题不会出现题目和答案的不一致性。这样通过模板随机生成的变体问题可以有效避免模型的模式匹配偏好。他们探索了目前最先进的 25 个 LLMs 来深入了解目前 LLMs 在数学推理方面的能力。并且，他们着重关注 LLMs 在同一问题不同实例之间所存在的方差分布，这能有效反映 LLMs 的性能，并且证明潜在的数据污染。他们同时增添了其他相关数据集来进一步说明问题，比如增加或减少从句的数量，这样形成的数

据集会来让问题变得更加复杂、引入看似相关但对最终答案无关的信息，这样形成的数据集会揭露模型在推理能力的脆弱。

总的来说，他们的工作全面展示了 LLMs 在数学推理方面的局限性，强调了更可靠评估方式的重要性和进一步研究 LLMs 推理能力的必要性。

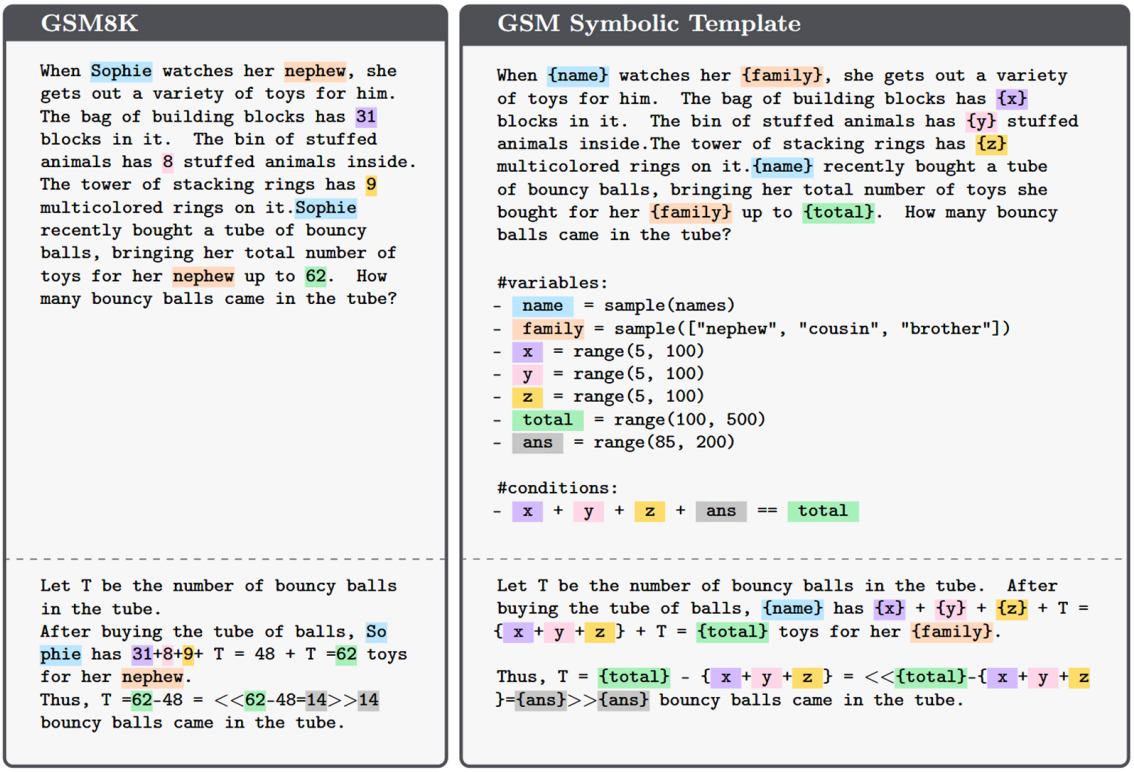


图 1. GSM-Symbolic 模板创建过程

2 相关工作

逻辑推理能力是人工智能应该拥有的能力之一。目前，LLMs 在诸多领域都展现出了巨大的潜能，但是它的推理能力还仍有待确认。目前很多研究都在研究 LLMs 逻辑推理的本质，并且探索 LLMs 是否真正具备逻辑推理能力。

现在提高 LLMs 的推理准确性的办法有很多，包括微调、思维链 CoT [7]、增加提示 prompt 以及为 LLMs 建立暂存区来提供额外记忆等 [6]。虽然这些方式并没有提高模型本身的推理能力，但是还是能有效提高模型的推理准确度。

关于数学推理的数据集很多，包括小学数学问题数据集 GSM8K [1]、竞赛数学问题 MATH [3]、奥林匹克级别的数学问题 MiniF2F [9] 以及数学专家原创的极具挑战性的数学问题基准数据集 FrontierMath [2]。

2.1 数据集

数学推理相关的数据集覆盖面很广，包含了小学数学问题、初高中数学问题、大学数学问题、奥林匹克等竞赛数学问题以及数学家原创的极具挑战的覆盖各个领域的数据集。这些问题通常包含一个问题，一个解题过程和一个答案。尽管有大部分数学问题是无数值解的 [8]，

但是限制数学问题解可以更加有效评估 LLMs 的准确性。在提高 LLMs 在有数值解问题上的准确度后，有望利用 LLMs 解决更多无数值解的任务，比如千禧年大奖问题。

GSM8K [1]：GSM8K 是一个数据集，包含 8500 个由人类问题编写者创作的高质量、语言多样的小学数学应用题。该数据集分为 7500 个训练问题和 1000 个测试问题。这些问题需要 2 到 8 步来解决，解决方案主要涉及执行一系列基本算术运算以获得最终答案。一个聪明的中学生应该能够解决每个问题。它可用于多步数学推理。

MATH [3]：MATH 是一个全新的数据集，包含 12500 个具有挑战性的竞赛数学问题。MATH 中的每个问题都有完整的逐步解决方案，可用于教导模型推导和解释答案。

MiniF2F [9]：MiniF2F 是一个用于奥林匹克级别数学问题陈述的正式数据集，旨在为证明神经定理提供一个统一的跨系统基准。当前的 MiniF2F 基准测试针对 Metamath、Lean 和 Isabelle，包括来自美国数学邀请赛（AIME）、美国数学竞赛（AMC）和国际数学奥林匹克竞赛（IMO）的 488 个问题陈述，以及高中和本科数学课程的材料。

FrontierMath [2]：一个由专家数学家精心编制和审核的数百个原创、极具挑战性的数学问题的基准。这些问题涵盖现代数学的大多数主要分支——从数论和实分析中的计算密集型问题到代数几何和范畴论中的抽象问题。解决一个典型问题需要相关数学分支的研究人员花费数小时的努力，而对于高端问题则需要数天。

3 本文方法

复现论文提出了一系列基于 GSM8K 的变体数据集。基于这些变体数据集，他们选择了市面上 20 个以上的公开模型以及一些封闭模型进行测试，这使得他们的研究可以更加全面地评估目前 LLMs 模型的表现。并且他们在每次评估的时候都采用 8shot 的思维链提示方式和贪心编码，这使得他们的结果具有可复现性，并且避免了模型创新性生成所带来的偏差。

这一章节将展示他们生成问题变体的具体方式以及实验结果，实验结果符合他们的预期，模型在这些数据集上的表现相较于在 GSM8K 数据集上的表现降低了。

3.1 名字和数据变体数据集

GSM-Symbolic 作为 GSM8K 的变体，变化的部分主要是名称和数据两个方面。图1是他们所建立的一个更改名称和数据的模板的例子，他们修改了原数据的名字、数据，并且添加一定的随机函数来动态调整数据，从而使得每次生成的数据不一样。他们根据 GSM8K 数据集的 100 个具体问题建立了这样的 100 个模板，然后利用每个模板随机生成 50 个样本实例，最终生成了 5000 个实例。他们对每个模板生成的样本实例一一进行组合，从而形成了 50 个小数据集，每个小数据集里面有 100 个实例，每个实例都是来自 GSM8K 数据的变体。

如图2所示，他们研究了单改名字，单改数据，两者都改三个方面模型准确度的差异。其中他们着重关注模型在测试过程中表现出的方差。他们发现，问题总体格式没变，只是其中的名称或数据变化会导致模型的预测准确率有大幅度变化。并且模型对修改文本名字时表现得更为稳健，但是对数据的变化表现得更为敏感。

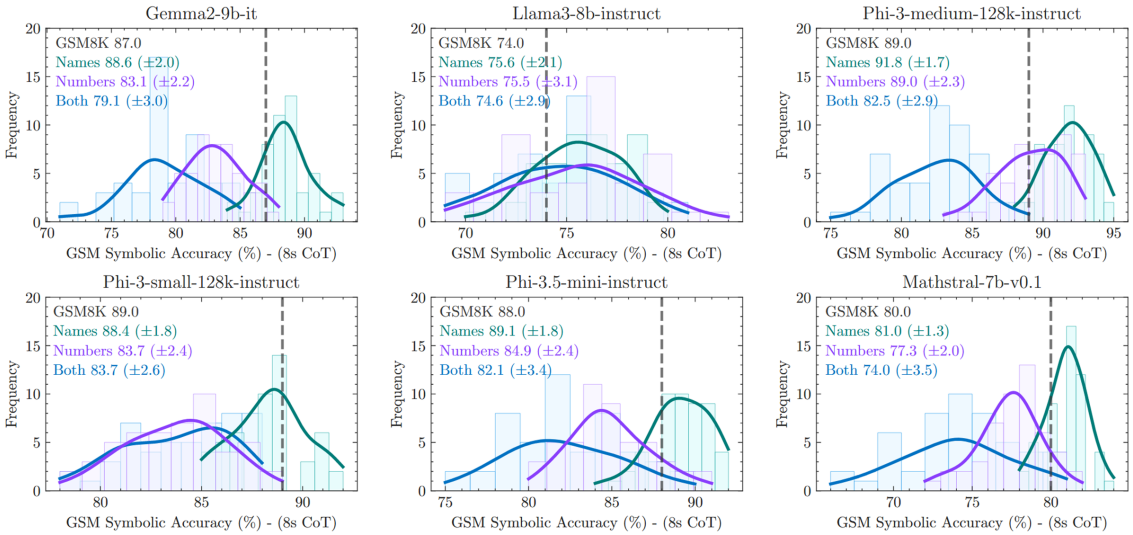


图 2. 名字和数据变体数据集准确率图

3.2 从句变体数据集

如图3所示，在从句更改方面，设置从句数据集有三种类型，分别是删除从句 M1，添加一个从句 P1和添加两个从句 P2，删除从句会明显减弱问题的难度，添加从句会增加问题的难度。

Different Levels of GSM-Symbolic Difficulty

GSM-Symbolic-M1: To make a call from a phone booth, you must pay $0.6 for each minute of your call. ~~After 10 minutes, that price drops to $0.5 per minute.~~How much would a 60-minute call cost?

GSM-Symbolic: To make a call from a phone booth, you must pay $0.6 for each minute of your call. After 10 minutes, that price drops to $0.5 per minute. How much would a 60-minute call cost?

GSM-Symbolic-P1: To make a call from a hotel room phone, you must pay $0.6 for each minute of your call.After 10 minutes, that price drops to $0.5 per minute. After 25 minutes from the start of the call, the price drops even more to $0.3 per minute.How much would a 60-minute call cost?

GSM-Symbolic-P2: To make a call from a hotel room phone, you must pay $0.6 for each minute of your call. After 10 minutes, the price drops to $0.5 per minute. After 25 minutes from the start of the call, the price drops even more to $0.3 per minute. If your total bill is more than $10, you get a 25% discount. How much would a 60-minute call cost?

图 3. 从句问题变体示例

如图4所示，几乎所有模型中性能分布演变的趋势是一致的，即随着难度的增加，性能下降，方差增大。这也更加证明了模型并非进行形式推理而是模式匹配。

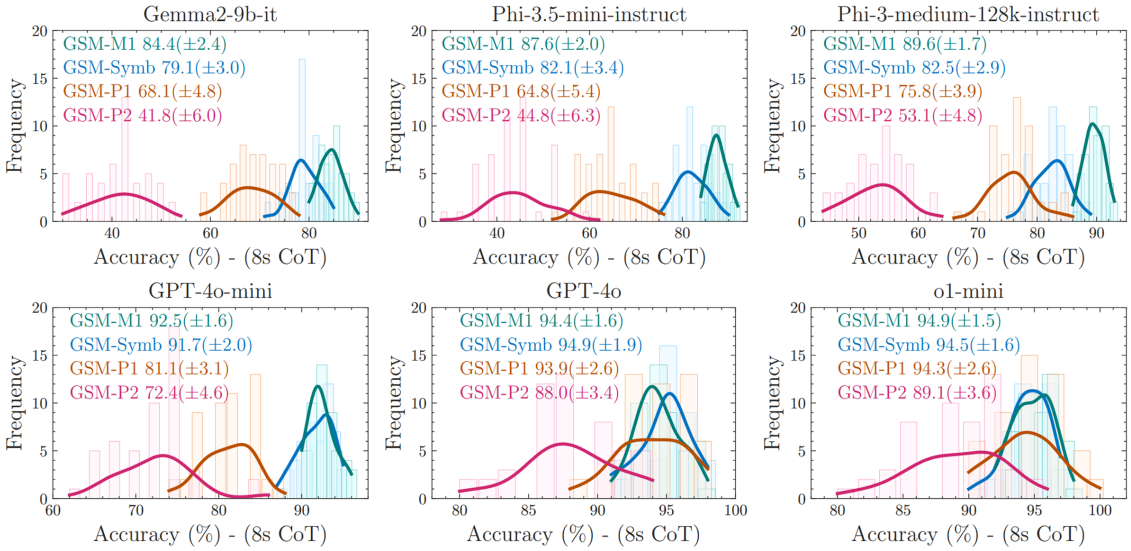


图 4. 从句变体数据集准确率图

3.3 无效信息变体数据集

我们平时使用大语言模型的时候，有时候会发现模型总是把对解决问题没有帮助的信息也纳入考虑。在数学推理上也是一样，他们建立了一个新的数据集 GSM-NoOp，图6展示了该数据集中一个示例。这个数据集在问题中添加了一些看似相关但是对答案无关紧要的从句，当问大模型的时候，他们通常会犯错。比如图中例子，模型会盲目地减去较小水果的数量，他们猜测这可能是因为它们的训练数据集中包含了类似的例子，这些例子需要转换为减法运算。

GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

---

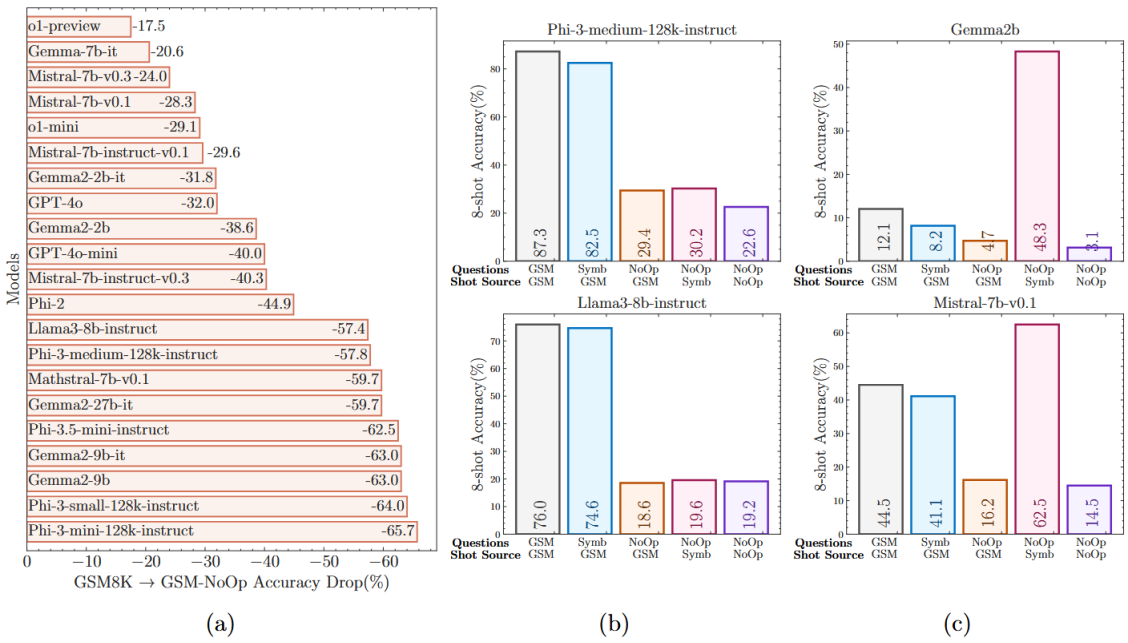
o1-mini: Oliver picks kiwis over three days as follows:
- Friday: 44 kiwis
- Saturday: 58 kiwis
- Sunday: Double the number he picked on Friday, which is $2 \times 44 = 88$ kiwis

However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185 kiwis.

Llama3-8B: Let's break it down step by step:
Oliver picks 44 kiwis on Friday. He picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, which is 2 x 44 = 88 kiwis.
Five of the kiwis picked on Sunday are a bit smaller than average, so we subtract 5 from the total number of kiwis picked on Sunday: 88 - 5 = 83 kiwis. Now, let's add up the total number of kiwis Oliver has:
44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis
So, Oliver has 185 kiwis in total.

图 5. GSM-NoOp 数据集示例

图6(a) 展示了他们测试的模型在该数据集上的性能变化，几乎所有模型在该数据集上的表现都下降了，并且很多新模型的表现比老模型还要差。



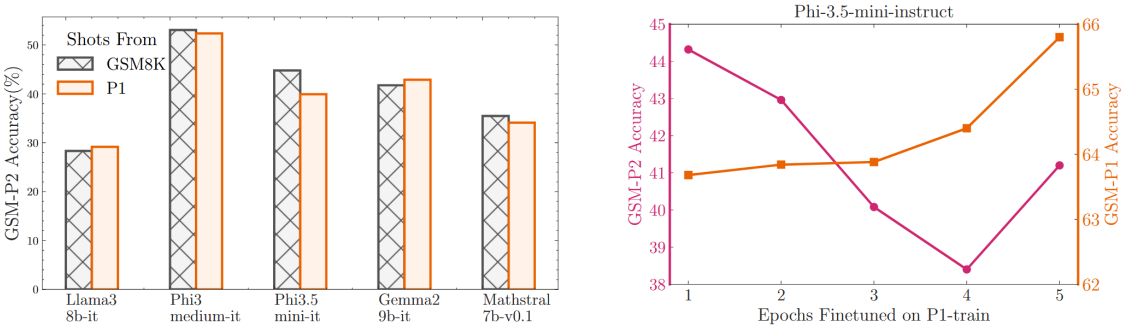
(a) (b) (c)

图 6. (a)GSM8K 到 GSM-NoOp 的准确度变化; (b) 更改 shot 来源模型准确度变化情况 1; (c) 更改 shot 来源模型准确度变化情况 2

3.4 更改 shot 来源

为了进一步展示模型性能的变化，他们通过更改 8shot 中 shot 的来源来看模型的性能下降程度。模型的性能下降大致有两种情况，如图6(b)(c) 所示。图6(b) 中，我们可以看到不管 shot 来源于哪，模型在 NoOp 数据集上的表现都很差，即便模型以 NoOp 作为 shot 来源，模型依旧无法很好地处理 NoOp 问题，有的甚至出现了下降。图6(c) 中，我们可以看到某些模型在使用 GSM-Symbolic 作为 shot 的来源的时候，他们 GSM-NoOp 的表现突飞猛进，这个结果是值得关注的。

他们验证了 shot 来自P1 数据集(增加一个从句的数据集)，模型在P2 数据集上的表现并没有显著变化，如图7(a) 所示。他们还发现如果使用P1 数据集来微调模型，模型在P1 数据集上的表现会提升，但在P2 数据集上的表现却在下降，如图7(b) 所示。他们对此得出的结论是扩大大语言模型的训练数据量并不能有助于提高语言模型的数学推理能力。



(a) (b)

图 7. (a)Gshot 来源于 P1 的准确率变化; (b) 使用 P1 微调给模型带来的改变

4 复现细节

4.1 与已有开源代码对比

本论文没用开源代码，数据集于复现结束前几日才发布，所以此前实验基于的数据集为自己手写代码产生的。参考了 GSM8K 的源代码，并参考了 Github 上作者 PolarisRising-War 复现的 GSM8K 代码项目—— “Math_Word_Problem_Collection”。调用参考了其中的 test_w_calculator.py 文件。除了源代码和参考代码，其余代码均是原创。

采用的模型并非本地部署，为了能够复现出有效的成果，我采用的是 OpenAI 提供的封闭模型 GPT3.5，型号是 gpt-3.5-turbo-0125。

4.2 实验环境搭建

在本地主机上搭建了 pycharm 虚拟环境，在服务器上也同样搭建了虚拟环境，在本地和服务器上同时运作。

5 实验结果分析

图8是我的复现结果。因为他们采用的模型较多，并且没有提供原始数据集，所以我计划复现一至两个模型，生成部分数据集。图中的 “FULL”、“100” 和 “20” 分别代表使用 GSM8K 数据集的全部测试问题、前 100 个问题和前 20 个问题。Cobbe 等人 [1] 提出了 GSM8K 数据集，并且提供了开源数据集和 Demo 代码 (GPT2 的微调和测试代码)，我从该代码入手，按照代码进行微调和测试，得到了经过微调的 GPT2 模型，测试准确率大约是 2%。我将 GSM8K 数据集上传至 OpenAI 进行微调和测试，经过微调的 GPT3.5 模型准确率约为 72%，这个准确度具有一定的参考作用。

由于复现论文采用的是无微调模型的 8shots 方式，所以我采用 8shots 的无微调方式测试了 GSM8K 前 100 个问题的准确度，和前 20 个问题的准确度。之所以要测试前 20 个问题的准确度，主要是因为模板生成困难，我只生成了 20 个模板，即我生成的 50 个小数据集中并没有包含 100 个问题，而是包含 20 个问题。我采用手写规则和代码生成了 20 个模板，最终得到了 20*50=1000 的实例数，每个数据集 20 个问题。其中前 10 个模板是手写得到的，后 10 个模板是采用 GPT4 生成，这样可以减少一定的重复性工作。这 20 个模板都是同时更改了数据和名称的，结果如图8(b) 所示。我得到的结果是更改名称和数据确实会产生较大的方差变化，并且更改数据和名称的平均准确率会比之前用 20 个 GSM8K 样本测试的准确率要低。这符合他们的结果。

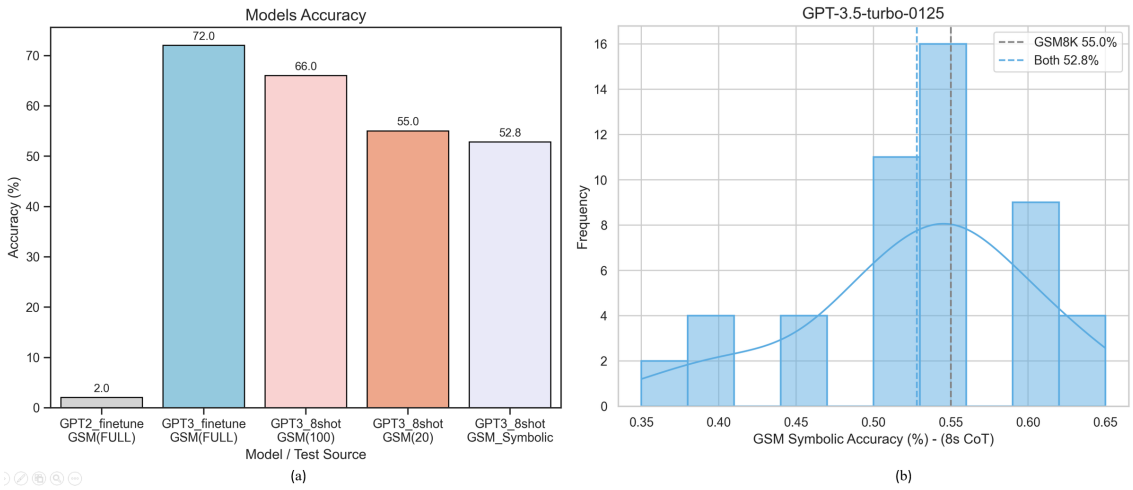


图 8. (a) 一定数据集数量下的模型准确度; (b) 模型在 GSM-Symbolic 数据集上的方差结果

6 总结与展望

复现论文是一篇探索性的论文，主要探索了现在 LLMs 的数学推理能力。得出的结论是：LLMs 实际上做的更多是模式匹配而非逻辑推理，这导致他们在数学推理上表现出脆弱性。本次复现工作复现了 GSM-Symbolic 主要研究部分，但仍有很多不足之处，未来希望能够进一步复现完整，以下是复现的不足总结：

- 复现实例不足。复现论文生成了 5000 个实例，而我只生成了 1000 个实例，基础模板生成的困难导致我的工作并不能完整复现出原论文的全部模板，这是十分可惜的，也是我的不足点；
- 复现不够完整。复现的数据集并没有包括删除和添加从句变体数据集和无效信息变体数据集，导致我无法得复现完整他们这两方面的结果；
- 模型较为单一。只用了一个 GPT3 模型进行测试，并且调用 API 的开销较大，不能随意测试，希望以后能找到一个更好的能够本地部署的模型。

参考文献

[1] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[2] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2024.

[3] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[4] Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo J Taylor, and Dan Roth. A peek into token bias: Large language models are not yet genuine reasoners. *arXiv preprint arXiv:2406.11050*, 2024.

[5] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

[6] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models, 2021. *URL https://arxiv. org/abs/2112.00114*, 2021.

[7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[8] Kaiyu Yang, Gabriel Poesia, Jingxuan He, Wenda Li, Kristin Lauter, Swarat Chaudhuri, and Dawn Song. Formal mathematical reasoning: A new frontier in ai. *arXiv preprint arXiv:2412.16075*, 2024.

[9] Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*, 2021.