评估大型语言模型在越南普通教育多项选择题中符号绑定能力

Duc-Vu Nguyen*
University of Information Technology
Ho Chi Minh City, Vietnam
Vietnam National University
Ho Chi Minh City, Vietnam
vund@uit.edu.vn

Quoc-Nam Nguyen*
University of Information Technology
Ho Chi Minh City, Vietnam
Vietnam National University
Ho Chi Minh City, Vietnam
20520644@gm.uit.edu.vn

摘要

在这篇论文中，我们评估了大型语言模型（LLMs）在零样本、单样本和少样本设置下执行多项选择符号绑定（MCSB）以完成多项选择问题回答（MCQA）任务的能力。我们的重点是越南语，相比于英语，越南语的挑战性MCQA数据集较少。现有的两个数据集ViMMRC 1.0和ViMMRC 2.0专注于文学。近期越南自然语言处理（NLP）领域的研究集中在2019年至2023年的越南国家高中毕业考试（VNHSGE），以评估ChatGPT的表现。然而，这些研究主要关注于ChatGPT如何逐步解决VNHSGE的问题。我们旨在通过提供结构化的LaTeX公式输入指南来创建一个新颖且高质量的数据集，涵盖数学、物理、化学和生物等领域。这个数据集可以用来评估LLMs以及较小的语言模型（LMs）的MCSB能力，因为它是严格按照LaTeX风格输入的。我们根据上下文确定最可能的答案字符（A、B、C或D），而不是像之前的越南研究那样逐步找到答案。这减少了计算成本，并加速了对LLMs的评估。我们在ViMMRC 1.0和ViMMRC 2.0基准以及我们提出的数据集上对六个知名的LLMs，即BLOOMZ-7.1B-MT、LLaMA-2-7B、LLaMA-2-70B、GPT-3、GPT-3.5和GPT-4.0进行了评估，结果显示了这些模型在越南语MCSB能力上的有希望的结果。该数据集仅用于研究目的，[1]。

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

多项选择题回答，多项选择符号绑定，语言建模，语言模型分析

ACM Reference Format:

Duc-Vu Nguyen and Quoc-Nam Nguyen. 2023. 评估大型语言模型在越南普通教育多项选择题中符号绑定能力. In *The 12th International Symposium on Information and Communication Technology (SOICT 2023), December 7–8, 2023, Ho Chi Minh, Vietnam.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3628797.3628837

*Both authors contributed equally to this research.

SOICT 2023, December 7–8, 2023, Ho Chi Minh, Vietnam

ACM ISBN 979-8-4007-0891-6/23/12…$15.00
https://doi.org/10.1145/3628797.3628837

[1]https://huggingface.co/uitnlp/vigetext_17to23

1 介绍

大型语言模型（LLMs）已成为自然语言处理（NLP）[1, 11, 13, 16] 众多领域中的关键工具。在人工智能驱动的进步时代，LLMs 解决复杂挑战的能力仍然是深入研究和评估的主题。其中一个挑战是多项选择题回答（MCQA）任务的领域，在这个任务中，LLMs 需要理解上下文信息并从一组选项中选出最合适的答案。在这篇论文中，我们深入探讨了 MCQA 中的多项选择符号绑定（MCSB）[15] 这一基本领域，旨在揭示面对这一复杂任务时 LLMs 的熟练程度。

虽然大型语言模型在各种自然语言处理任务中表现出色，但越南语提出了独特挑战和机遇。与英语不同，越南语可用于研究目的的具有挑战性的多项选择题数据集非常有限。现有的数据集，如 ViMMRC 1.0[12] 和 ViMMRC 2.0[9]，主要集中在文学情境上，这在评估大型语言模型跨不同领域的性能时留下了一个显著的空白。

在最近的越南语 NLP 研究中，模型评估主要集中在它们解决 2019 年至 2023 年越南国家高中毕业考试（VNHSGE）问题的能力上。然而，这些研究主要分析逐步解题过程，而不是关注模型在越南语 MCQA 方面的更广泛能力。

认识到需要一个涵盖广泛主题并促进评估 LLMs 的 MCSB 能力的综合数据集，我们创建了一个新颖且高质量的数据集。该数据集包括在数学、物理、化学和生物学科中输入 LaTeX 公式的结构化指南。通过强制执行严格的 LaTeX 格式风格，我们的目标是提供一个标准化和细致的评估环境，不仅可以用于评估 LLMs，还可以用于评估较小的语言模型 (LMs)。

在此研究范围内，我们的主要目标是根据其上下文框架预测给定问题的正确答案字符（A、B、C 或 D）。为了确保全面评估，我们对六个知名的大语言模型进行了性能评估：BLOOMZ-7.1B-MT、LLaMA-2-7B、LLaMA-2-70B、GPT-3、GPT-3.5 和 GPT-4.0。我们的评估涵盖了 ViMMRC 1.0 和 ViMMRC 2.0 基准测试以及我们新颖的数据集。这项详尽分析的结果为了解大语言模型在越南语中的 MCSB 能力提供了宝贵的见解，有望影响该领域的未来研究和发展努力。这些发现为充分利用语言模型的能力铺平了道路，以解决越南语中具有挑战性的多项选择题数据集的稀缺问题，并提高其在专业领域中的熟练程度。

我们的贡献总结如下：

(1) 我们提供了一个新型的高质量数据集，并附有结构化的指南，用于在数学、物理、化学和生物领域输入 LaTeX 公式。
(2) 我们在越南普通教育背景下对大型语言模型的符号绑定能力进行了多项选择题实验。我们的综合评估包括六个突出的大型语言模型，即 BLOOMZ-7.1B-MT、LLaMA-2-7B、LLaMA-2-70B、GPT-3、GPT-3.5 和 GPT-4.0。
(3) 进行了广泛分析和讨论以深入研究 LLMs 对越南多项选择题考试的影响，并探讨 LLMs 在教育中的影响。

2 相关工作

理解并操作语言中的符号对于有效回答多项选择题至关重要。

在他们的工作中，Lai et al. [6] 引入了 RACE，这是一个为了评估阅读理解领域方法而创建的新数据集。该数据集由近 28,000 篇段落和大约 100,000 个问题组成，这些问题是由英语教师开发的，并从中国 12 至 18 岁中学生和高中生参加的英语考试中收集而来。它涵盖了广泛的主题，这些主题经过精心挑选以评估学生的理解和推理能力。

Hendrycks et al. [5] 介绍了 MATH 数据集，这是一个包含 12,500 个具有挑战性的数学问题的新数据集，专门用于竞赛评估。MATH 中的每个问题都包括一个全面的分步解答，为训练模型生成答案推导和解释提供了宝贵资源。

在科学领域，Lu et al. [8] 引入了 SCIENCEQA，一个新的基准数据集，包含大约 21,000 个多模态多项选择题，涵盖各种科学主题。该数据集还包括答案、相关讲座和解释的注释。

Lewis et al. [7] 介绍了 MLQA，一个多语言抽取式问答基准。MLQA 包含七种语言的问答实例：英语、

阿拉伯语、德语、西班牙语、印地语、越南语和简体中文。它包括超过 12,000 个英语实例和每种语言 5,000 个实例，每个都有平均四种语言的平行版本。

Dao et al. [2] 评估了 ChatGPT（2 月 13 日版本），一个大型语言模型，以评估其在回答源自 2019 年至 2023 年越南国家高中毕业考试的英语测试题方面的表现。研究分析的结果显示，ChatGPT 在 50 道题目中平均正确率为 40 题，相当于越南常用的 10 分制中的 7.92 分。值得注意的是，ChatGPT 的回答准确性在不同难度级别的问题上保持一致，这突显了该模型在此特定任务上的熟练程度。

3 数据集

本节介绍了用于评估大型语言模型符号绑定能力的数据集。

3.1 ViMMRC 1.0

Nguyen et al. 组装了一个数据集 **ViMMRC 1.0**，包含 2,783 套选择题及其相应答案。这些问题来自 417 本越南文本，这些文本通常用于小学学生的阅读理解教学。

3.2 ViMMRC 2.0

ViMMRC 2.0 由 Luu et al.引入以扩展早期的 ViMMRC 1.0（在第 3.1 节中描述）数据集，该数据集设计用于越南教科书中的多项选择阅读理解。**ViMMRC 2.0** 包含了 699 篇阅读文章，包括散文和诗歌，以及 5,273 道题目。与之前的版本不同，这个数据集不限制问题必须有固定的四个选项。此外，该新数据集中的问题是设计来更具挑战性的，要求模型全面理解整个阅读文章、问题及每个可用选项的内容才能正确提取答案。

3.3 我们提出的数据集：ViGEText_17to23

我们提出的数据集是通过从公开的互联网来源进行网络爬虫精心组装而成。与 Dao et al. [3] 在 2019 年至 2023 年间开展的先前研究不同，我们的目标是涵盖 2017 年至 2023 年的整个越南普通教育考试范围。这种全面的方法包括了在近年来对几乎所有越南学生都具有重要意义的 2017 年和 2018 年的困难考试。重要的是要强调，确切且毫无疑问正确的答案仅从越南教育部获得。这样做是为了维护我们数据集的最高准确性标准。表 1 显示了 2017 年至 2023 年每个科目的统计信息。

在**数学**，**物理**，**生物学**和**化学**中，标准化需要将几何学转换为几何语言。然而，实施这一更新耗时较长，因此我们将其推迟到未来的工作中。此外，关于**地理**的主题，包含时间元素的陈述在未来近期或远期可能不准确。因此，我们删除了这些陈述以确保数据集保持可重复使用并接受长期评估。

我们的主要目标是从越南基础教育中建立一个全面且高质量的数据集。为了实现这一目标，我们致力于提供精心设计的指南，以准确输入数学、物理、化学和生物中的 LaTeX 公式。这项雄心勃勃的努力背后的主要理由是显著提升符号数学领域的发展。符号数学通常依赖于 LaTeX 标记来确保精确性和灵活性，在各种科学学科中扮演着关键角色。这个数据集创建计划旨在应对研究人员、教育工作者和学生在处理数学表达式、方程和记号时所面临的挑战。

为了更多细节，我们提出的数据集是严格按照 LaTeX 中**数学 JaX**[2] 库的格式标准精心开发的。**为了严格按照规则表示数学公式，确保在所有上下文中进行公平推理，并使进一步的研究能够轻松解析，即使是小型语言模型也是如此**。，我们的 LaTeX 输入指南规定数学公式必须紧密书写（除非在编写化学方程式时，如附录 A 中的化学示例所示），并且只在必要时使用花括号。输入公式后，请确保它与真实试卷中的原始图像完全一致显示。此外，我们的方法采用了“\测”来准确表示化学元素，并且“*\pu*”有效表示测量单位（两者都包含在 **mhchem**[3] 扩展中）。然而，在某些情况下，必须使用大括号，在以下示例中用红色的大括号包围：

[2]https://www.mathjax.org/

[3]https://docs.mathjax.org/en/latest/input/tex/extensions/mhchem.html

年	测试类型		数学	物理	化学	生物学	历史	地理	市民教育	总计
2017	实际		45	37	37	35	40	28	40	262
2018	样本		39	34	37	37	40	24	40	251
	实际		43	33	37	39	40	18	40	250
2019	样本		36	35	38	39	40	22	40	250
	实际		36	36	38	32	40	20	40	242
2020	样本		37	35	39	39	40	21	40	251
	实际	第 1 轮	40	35	35	38	40	18	40	251
		第二轮	41	35	40	38	40	18	40	252
2021	样本		39	36	40	36	40	13	40	244
	实际	第一轮	42	35	40	35	40	13	40	245
		第二轮	41	35	40	37	40	14	40	247
2022	样本		43	35	40	36	40	11	40	245
	实际		42	36	39	37	40	13	40	247
2023	样本		41	36	38	34	40	14	40	243
	实际		41	37	38	33	40	13	40	242
总数			606	530	581	545	600	260	600	3722

表 1: 我们提出的数据集统计。鉴于新冠肺炎，越南教育部作为预防措施在 2020 年和 2021 年采用了区域两轮考试安排。

- 在积分表达式中，考虑以下示例："$\int_0^6{{}f'(x)\,dx{}}$" 表示 $\int_0^6 f'(x)\,dx$。
- 如果函数的输入在最左和最右位置包含括号，或者是一个超过一个字符的非数字字符串，请考虑以下示例："$\ln{{}(5a){}}$" 表示 $\ln(5a)$，"$e=\cos{{}(100{\pi}t+\pi){}}~(\pu{V})$" 表示 $e = \cos(100\pi t + \pi)\ (V)$。

来自我们提出的数据集的样本可以在附录 A 中找到。此外，提供了详细指南以在进一步的研究中使用我们的数据集 [4]。

4 实验

在本节中，我们介绍了基线大型语言模型（参见第 4.1 节）及其评估设置（参见第 4.2 节）。此外，实验结果描述于第 4.3 节。

[4]https://huggingface.co/datasets/uitnlp/ViGEText_17to23

4.1 基线模型

本节介绍了用于评估其符号绑定能力的大语言模型。我们探讨了更高的 MCSB 能力是否会导致多项选择任务的准确性更高。我们在两个文献数据集（ViMMRC 1.0[12] 和 ViMMRC 2.0[9]）和我们提出的数据集上评估五次射击模型性能，所有这些都在第 3 节中进行了介绍。

- **BLOOMZ:** Muennighoff et al. 采用了多语言任务适配 (MTF) 来微调预训练的多语言 BLOOM 和 mT5 模型系列，由此产生了被称为 **BLOOMZ** 和 mT0 的适应版本。他们的研究结果表明，使用英语提示对这些大型多语言语言模型进行针对英语任务的微调能够使它们有效泛化到作为预训练语料库一部分的非英语语言。此外，在使用英语提示对多语言任务进行微调时，这些模型在英语任务和涉及非英语语言的任务上表现出增强的表现力，在零样本场景中取得了众多 SOTA 结果。
- **_LLAMA:_** Touvron et al. 介绍了 LLaMA，一系列基础语言模型，参数数量从 7 亿到令人印象深刻的 70 亿不等。使 LLaMA 引人注目的是其在万亿个标记上的训练，展示了仅使用公开可访问的数据集训练最先进模型的可能性，而无需依赖专有或不可获取的数据源。
- **GPT-3:** Brown et al. 训练了 **GPT-3**，一个具有 1750 亿参数的自回归语言模型，比任何之前的非稀疏语言模型多出 10 倍，并测试了其在少样本设置中的性能。**GPT-3** 在所有任务中均未进行梯度更新或微调，仅通过与模型的文字交互来指定任务和少样本演示。**GPT-3** 在许多 NLP 数据集上表现出色。
- **GPT-4: GPT-4**，由 OpenAI报道的，是一个大型多模态模型，接受文本和图像输入并生成文本输出。**GPT-4** 是一个基于 Transformer 的模型，训练目的是预测文本文档中的下一个标记。尽管它可能无法在所有实际情况下与人类能力相匹配，但在专业和学术基准上表现出色，包括以前十名的成绩通过模拟律师资格考试。

Dưới đây là các câu hỏi trắc nghiệm (kèm đáp án) về toán học
Đề bài:
Cho hàm số $y = \frac{x-2}{x+1}$. Mệnh đề nào dưới đây đúng?
A. Hàm số nghịch biến trên khoảng $(-\infty; -1)$
B. Hàm số đồng biến trên khoảng $(-\infty; -1)$
C. Hàm số đồng biến trên khoảng $(-\infty; +\infty)$
D. Hàm số nghịch biến trên khoảng $(-1; +\infty)$
Đáp án: B

Đề bài:
Trong không gian $Oxyz$, cho mặt cầu (S) tâm $I(1; 3; 9)$ bán kính bằng 3. Gọi M, N là hai điểm lần lượt thuộc hai trục Ox, Oz sao cho đường thẳng MN tiếp xúc với (S), đồng thời mặt cầu ngoại tiếp tứ diện $OIMN$ có bán kính bằng $\frac{13}{2}$. Gọi A là tiếp điểm của MN và (S), giá trị $AM \times AN$ bằng
A. 39
B. $12\sqrt{3}$
C. 18
D. $28\sqrt{3}$
Đáp án:

English version

The following are multiple-choice questions (with answers) about Mathematics
Question:
The given function is $y = \frac{x-2}{x+1}$. Which of the following statements is correct?
A. The function is decreasing on the interval $(-\infty, -1)$.
B. The function is increasing on the interval $(-\infty, -1)$.
C. The function is increasing on the interval $(-\infty, +\infty)$.
D. The function is decreasing on the interval $(-1, +\infty)$.
Answer: B

Question:
In the space $Oxyz$, consider the sphere (S) centered at $I(1; 3; 9)$ with a radius of 3. Let M and N be two points on the Ox and Oz axes, respectively, such that the line MN is tangent to (S). Simultaneously, the circum-sphere of tetrahedron $OIMN$ has a radius of $\frac{13}{2}$. Let A be the point of tangency between MN and (S). The value of $AM \times AN$ is:
A. 39
B. $12\sqrt{3}$
C. 18
D. $28\sqrt{3}$
Answer:

图 1: 一个我们提出的数据集的一次性学习数学示例。在这个一次性学习示例中，有一个指令示例和一个初始不完整的示例。

为了更多细节，BLOOMZ-7.1B-MT、LLaMA-2-7B、LLaMA-2-70B、GPT-3、GPT-3.5 和 GPT-4 被用于评估越南普通教育中多项选择题的符号绑定能力。对于 LLaMA-2-7B 和 LLaMA-2-70B，我们使用 Replicate API[5]。对于 GPT-3、GPT-3.5 和 GPT-4，我们使用 OpenAI API[6]。我们在 2023 年 5 月 12 日版本 [7] 中部署了 GPT-3.5 和 GPT-4。最后，我们使用由 Google Colab 提供的 NVIDIA A100 GPU 对 BLOOMZ-7.1B-MT 模型进行全精度推理。

在 ViMMRC、我们的数据集以及一般的多项选择提示中，我们将一个问题及其答案选项作为一个单一的提示呈现给大语言模型。该提示设计为让模型仅预测一个标记。模型所选的答案对应于概率最高的那个标记。我们将概率最高的选项视为每个样本的预测。

4.2 设置

在本节中，我们的实验设置被指定。

少样本提示遵循 Hendrycks et al.，我们将大型语言模型（在第 4.1 节中介绍）的提示设置为如图 1 所示的形式。我们以 “The following are multiple choice questions (with answers) about [subject].” 开头作为每个提示的起始句。对于零样本评估，我们将问题附加到提示后面。对于少样本评估，在添加问题之前我们会向提示中加入最多 5 个带有答案的示范例子。所有提示都以 “Answer: ” 结尾。然后模型为标记 “A”，“B”，“C” 和 “D”（ViMMRC 2.0 [9] 不限制问题必须有固定四个选项）生成概率，我们将概率最高的选项视为预测结果。为了确保评估的一致性，我们创建了一个测试集，包含从越南教育部于 2017 年发布的**样本测试**中提取的每个学科的 5 个固定的少样本例子。

评估指标： 我们在本研究中初始化了 MCSB 的准确性。公式如下所示，见式 1：

$$\text{Accuracy} = \frac{\text{Number of correct prediction tokens}}{\text{Total number of tokens}} \quad (1)$$

[5]https://replicate.com/

[6]https://openai.com/blog/openai-api

[7]这些版本是在进行通考之前发布的；因此，2023 年没有泄题信息。

最大序列长度： 我们为所有大型语言模型（除 GPT-4 外）设定了最大序列长度为 4096 个标记。由于 OpenAI API 出现了意外错误，我们为 GPT-4 指定了最大序列为 3073 个标记。ViMMRC 2.0 是一个专注于从文学课本中收集文献的语料库，其中包含极长的释义（超过 3073 和 4096 个标记）。因此，为了确保 ViMMRC 2.0 的序列长度不超过 GPT-4 和其他 LLMs（分别为 3073 和 40961 个标记）的最大序列长度，我们实现了 SentenceTransformer[8] [14] 来对段落进行排序并移除不必要的句子。在图 2 中，展示了 GPT-4 在 ViMMRC 2.0 上经过排序和移除前后句子的长度，这一模式同样适用于其他 LLMs。

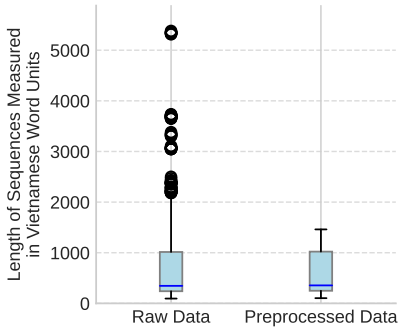

图 2: 序列长度的分布，使用 VnCoreNLP 以越南语词汇单位衡量 [17]，对于原始数据和预处理后的数据，以确保它们不超过 GPT-4 允许的最大序列长度。

最大新标记数： 我们配置了最多 1 个令牌来生成响应，以符合本研究中越南普通教育多项选择题任务的约束。

温度参数： 温度参数已被设置为 0 的值以增强结果的可重复性并方便复现。

分词器： 对于图 2 的分析，我们实现了 tiktoken[9]，这是一个基于字节配对编码（BPE）的高速分词器，特别设计用于补充 OpenAI 的模型。

4.3 结果

[8]https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

[9]https://github.com/openai/tiktoken

方法		ViMMRC 1.0	ViMMRC 2.0
使用 ELMo 提升的分数 [12]		61.81	–
mBERT$_{cased}$ [10]		60.50	–
MMM$_{viBERT\&ViNLI}$ [9]		80.16	58.81
BLOOMZ-7.1B-MT	Zero-Shot	79.96	64.47
	One-Shot	74.51	59.78
	Five-Shot	70.04	56.36
LLaMA-2-7B	Zero-Shot	33.46	31.29
	One-Shot	40.47	35.35
	Five-Shot	39.69	36.70
LLaMA-2-70B	Zero-Shot	75.10	64.83
	One-Shot	78.02	71.33
	Five-Shot	77.82	72.50
GPT-3	Zero-Shot	77.04	68.53
	One-Shot	79.77	70.51
	Five-Shot	79.57	69.61
GPT-3.5	Zero-Shot	82.88	73.49
	One-Shot	83.46	73.49
	Five-Shot	84.44	72.05
GPT-4	Zero-Shot	90.86	84.22
	One-Shot	91.63	85.03
	Five-Shot	90.66	85.84

表 2: 大型语言模型在 ViMMRC 1.0 和 ViMMRC 2.0 数据集上的实验结果。

4.3.1 **实验结果在 *ViMMRC* 上.** 表 2 展示了之前的工作和大型语言模型在 ViMMRC 1.0 和 ViMMRC 2.0 数据集上的结果。结果显示，只有 GPT-3.5 和 GPT-4 超过了旧的 SOTA 模型 [9]，而大多数 LLMs 都超越了除了 BLOOMZ-7.1B-MT 和 LLaMA-2-7B 以外的两个参数最少的 LLM。这表明模型架构、规模与在多项选择数据集上的性能之间存在重要关系。此外，这一结果还表明了 LLMs 在越南文学任务中的 MQCA 有效性。

结果显示，GPT 系列模型的表现优于其他模型；参数越大，GPT 模型表现越好。GPT-4 达到了最高的准确率，并在 ViMMRC 1.0 和 ViMMRC 2.0 数据集上的三种少样本提示场景中超越了其他大语言模型。GPT-3 和 GPT-3.5 也获得了积极的表现，这两种大语言模型均超过了其他模型。Brown et al.还观察到更大的 GPT-3 模型表现更好，尽管进步趋于稳定。

LLaMA-2-7B 和 BLOOMZ-7.1B-MT 是本研究中实现的最小的大型语言模型。然而，这两个模型的结果是矛盾的。根据模型大小和训练，LLaMA-2-7B 表现不佳（在 ViMMRC 1.0 和 ViMMRC 2.0 上表现最差），而 BLOOMZ-7.1B-MT 在零样本情况下超过 LLaMA-2-70B 和 GPT-3，显示了潜在的能力。此外，在一样本情况下，BLOOMZ-7.1B-MT 相比其他大型语言模型（除了 GPT-3.5 和 GPT-4）具有竞争力的结果。然而值得注意的是，BLOOMZ-7.1B-MT 没有利用提示的好处。在我们的评估中，我们观察到 BLOOMZ-7.1B-MT 在零样本情况下达到了其最佳性能，但在过渡到一样本和五样本情况时，表现有所下降。这一观察突显了该模型与其他模型在面对不同水平的上下文信息提示时的行为差异。

4.3.2 **在我们提出的数据集上的实验结果.** 表 3 展示了大型语言模型在我们提出的数据集上的结果。不出所料，GPT-4 在我们提出的数据集的零样本、单样本和五样本场景中分别平均实现了 55.81%、67.87%和 71.24%，并且优于其他 LLMs。GPT 系列的表现也比 LLaMA 和 BLOOMZ 更好。LLaMA-2-70B 在我们的评估中成为第二高表现者，展示了其在符号绑定任务中的出色能力。这一发现表明了 LLMs 在越南普通教育任务上的 MQCA 的巨大效率。

相比之下，虽然 LLaMA-2-7B 仍然是一个有能力的大语言模型，但其性能低于其较大的版本 LLaMA-2-70B。这种差异可以归因于模型大小和训练数据的不同。较小的模型通常在捕捉复杂模式和细微差别方面存在限制，这对于符号绑定任务至关重要。然而，尽管 BLOOMZ-7.1B-MT 与 LLaMA-2-7B 拥有相似的模型规模（70 亿个参数），它却取得了显著的结果。其在少样本设置下的平均准确率分别为 43.14%、35.09%和 38.00%，表现出色。

值得注意的是，在数学领域，GPT-4 在零样本设置中表现不佳，准确率仅为 24.09%，在我们基线中的所有 LLM 中结果最低。紧随其后，GPT-3.5 达到了稍高的准确率 24.42%，标志着在零样本设置中的第二差表现。相比之下，GPT-3 在零样本设置中展示了出色的能力。然而，当我们过渡到少样本设置时，我们观察

大型 语言模型		数学	物理	化学	生物学	历史	地理	市民 教育	平均值
BLOOMZ-7.1B-MT	Zero-Shot	25.25	36.04	34.25	40.00	49.83	48.46	68.17	43.14
	One-Shot	22.77	30.57	30.12	35.05	43.33	24.62	59.17	35.09
	Five-Shot	25.25	28.68	32.19	31.74	40.33	43.46	64.33	38.00
LLaMA-2-7B	Zero-Shot	24.59	23.96	28.57	26.61	28.83	32.31	27.33	27.46
	One-Shot	25.58	23.77	28.74	26.24	28.50	28.08	27.67	26.94
	Five-Shot	27.06	24.15	22.89	26.97	26.33	27.69	33.83	26.99
LLaMA-2-70B	Zero-Shot	32.67	37.55	35.63	41.10	49.00	46.15	53.83	42.28
	One-Shot	35.31	42.83	37.87	37.43	52.00	41.54	67.50	44.93
	Five-Shot	34.16	40.57	36.14	43.30	55.67	41.92	67.67	45.63
GPT-3	Zero-Shot	36.47	36.98	36.49	37.06	43.00	40.00	58.50	41.21
	One-Shot	40.76	38.30	41.14	42.20	43.50	40.38	63.17	44.21
	Five-Shot	40.10	39.43	41.48	43.12	47.83	41.54	67.33	45.83
GPT-3.5	Zero-Shot	24.42	36.04	39.76	45.69	57.50	53.85	67.67	46.42
	One-Shot	38.94	43.96	49.91	50.28	57.50	51.54	69.67	51.69
	Five-Shot	40.26	44.72	50.09	51.38	60.17	51.54	72.67	52.97
GPT-4	Zero-Shot	24.09	47.92	37.69	57.80	75.00	61.15	87.00	55.81
	One-Shot	55.45	66.04	53.36	64.77	78.50	68.46	88.50	67.87
	Five-Shot	56.44	66.60	64.20	69.17	82.17	71.92	88.17	71.24

表 3: 大规模语言模型在我们提出的数据集上的实验结果。

到 GPT-3.5 和 GPT-4 的性能有了显著提升。这一观察强调了少样本方法及其对 GPT 模型和更广泛的 LLM 的影响的重要性，我们在第 5 节中对此进行了详细探讨。此外，在地理、历史和公民教育领域，需要注意的是当前的 LLM 还未达到完美的表现。它们有时会提供这些主题上的不准确答案。需要持续改进以提升其在这些特定领域的准确性和有效性。

5 讨论

如图 3 所示，GPT 模型系列在面对更长的最大序列长度时表现出改进的性能。相反，其他大型语言模型（LLMs）倾向于在受到较短最大序列长度约束时产生更好的结果。此外，值得注意的是一个不同的发现：BLOOMZ 模型在被提示时表现明显下降。总之，这一发现强调了最大序列长度和提示数量对本研究中使用的数据集背景下 LLMs 性能的显著影响。正确的参数对于有效模型至关重要，突显了在 LLM 应用中进行定制调整的需求，尤其是在教育和其他领域。

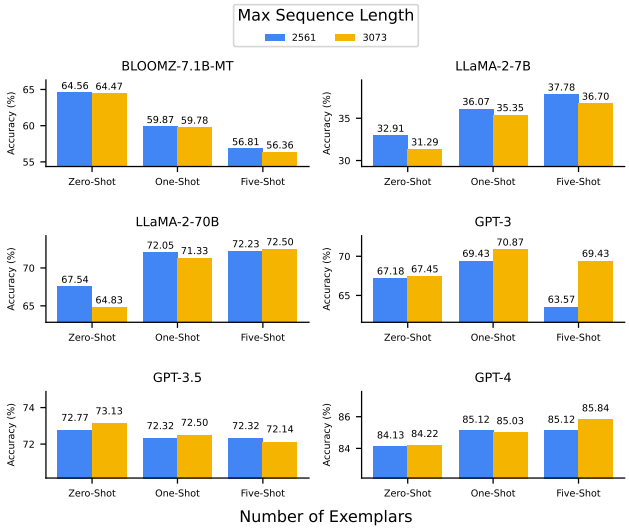

图 3: 具有不同最大序列长度和示例数量的 LLMs 在 ViMMRC 2.0 上的性能评分。

图 4 展示了每个大语言模型在我们提出的数据集上五次射击设置下的平均准确率（从 2017 年到 2023 年）。结果显示，大语言模型在 2017 年和 2018 年的越南普通教育考试中表现挣扎。然而，从 2020 年到 2023 年，大语言模型的表现有了显著的变化。主要原因是考试本身的演变性质。总体而言，GPT-3.5 和 GPT-4 一直表现出最令人印象深刻的表现。

2017 年和 2018 年的考试被公认为最具挑战性的普通教育评估。这些考试包含复杂且要求较高的问题，对 LLMs 构成了严峻的挑战。相比之下，自 2020 年以来，考试有意变得更加简单，题目设计得更容易，不那么繁琐。此外，在这些考试中，GPT-4 和其他 LLMs 在第一轮考试问题中大多解决了 60%，而这些问题对于这些 LLMs 来说比较困难的是剩余的 40%。这是因为第一轮考试中的大多数问题是相对容易的，而剩下的 40%则更加难以解决。这一发现突显了考试难度对 LLMs 表现的关键作用。

6 结论与未来工作

在这项研究中，我们调查了大型语言模型（LLMs）在越南语多项选择题回答（MCQA）中的多选符号绑定（MCSB）能力。我们的贡献包括创建了一个新颖的高质量数据集，对六个著名的大型语言模型进行了严格的评估，并深入分析了它们对越南语 MCQA 的影响，特别是在普通教育方面。

我们的新型数据集严格执行 LaTeX 指南，确保在未来的研究中易于解析。该数据集专为数学、物理、化学和生物等科目的多项选择题设计，填补了越南语领域的一个重要空白。这一标准化资源便于对各种领域的大型语言模型进行全面评估。

我们广泛测试了 BLOOMZ-7.1B-MT、LLaMA-2-7B、LLaMA-2-70B、GPT-3、GPT-3.5 和 GPT-4.0 在越南语多项选择题任务中的表现。我们的评估突显了它们的优势和劣势，加深了我们对它们能力的理解。

限制条件

本文提供了评估而没有详细的分析或解释，指向未来的研究。复制 GPT-X 的结果具有挑战性，由于开源访问受限。然而，所有实验结果都包括 [10]，有助于未来的研发工作。

ACKNOWLEDGMENTS

此项研究得到了 VNUHCM-信息技术大学科学研究支持基金的支持。我们感谢审稿人的宝贵意见，这极大地提高了我们的工作质量。

REFERENCES

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[2] Xuan-Quy Dao, Ngoc-Bich Le, Xuan-Dung Phan, and Bac-Bien Ngo. 2023. An Evaluation of ChatGPT' s Proficiency in English Language Testing of The Vietnamese National High School Graduation Examination. *Available at SSRN 4473369* (2023).

[3] Xuan-Quy Dao, Ngoc-Bich Le, The-Duy Vo, Xuan-Dung Phan, Bac-Bien Ngo, Van-Tien Nguyen, Thi-My-Thanh Nguyen, and Hong-Phuoc Nguyen. 2023. VNHSGE: VietNamese High School Graduation Examination Dataset for Large Language Models. arXiv:2305.12199 [cs.CL]

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask

[10] https://github.com/uitnlp/vigetext_17to23

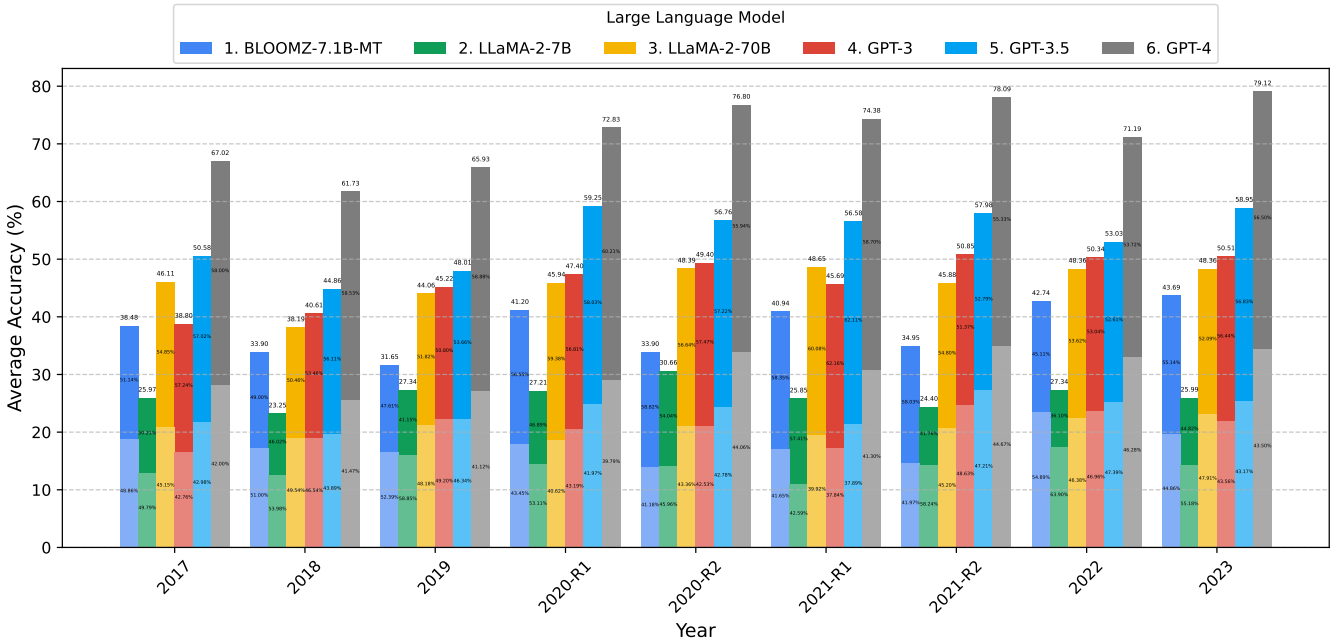

图 4: 从 2017 年到 2023 年，大型语言模型在我们提出的数据集上的平均得分。每列底部的浅色部分表示每个测试的后半部分，因为根据越南教育部的说法，这部分一直比前半部分更具挑战性。

Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[5] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS* (2021).

[6] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 785–794. https://doi.org/10.18653/v1/D17-1082

[7] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7315–7330. https://doi.org/10.18653/v1/2020.acl-main.653

[8] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

[9] Son T. Luu, Khoi Trong Hoang, Tuong Quang Pham, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. A Multiple Choices Reading Comprehension Corpus for Vietnamese Language Education. arXiv:2303.18162 [cs.CL]

[10] Son T. Luu, Kiet Van Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2021. An Experimental Study of Deep Neural Network Models for Vietnamese Multiple-Choice Reading Comprehension. In *2020 IEEE Eighth International Conference on Communications and Electronics (ICCE)*. 282–287. https://doi.org/10.1109/ICCE48956.2021.9352127

[11] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 15991–16111. https://doi.org/10.18653/v1/2023.acl-long.891

[12] Kiet Van Nguyen, Khiem Vinh Tran, Son T. Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Enhancing Lexical-Based Approach With External Knowledge for Vietnamese Multiple-Choice Machine Reading Comprehension. *IEEE Access* 8 (2020), 201404–201417. https://doi.org/10.1109/ACCESS.2020.3035701

[13] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023). https://api.semanticscholar.org/CorpusID:257532815

[14] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[15] Joshua Robinson and David Wingate. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=yKbprarjc5B

[16] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[17] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, New Orleans, Louisiana, 56–60. https://doi.org/10.18653/v1/N18-5012

A 零样本提示用于我们的数据集

```
Dưới đây là các câu hỏi trắc nghiệm (kèm đáp án) về toán học
Đề bài: Một chiếc bút chì có dạng khối lăng trụ lục giác đều có cạnh đáy
$3~\pu{mm}$ và chiều cao bằng $200~\pu{mm}$. Thân bút chì được làm bằng
gỗ và phần lõi được làm bằng than chì. Phần lõi có dạng khối trụ có chiều
cao bằng chiều dài của bút và đáy là hình tròn có bán kính $1~\pu{mm}$.
Giả định $1~\pu{m3}$ gỗ có giá $a~(\pu{triệu đồng})$, $1~\pu{m3}$ than
chì có giá $8a~(\pu{triệu đồng})$. Khi đó giá nguyên vật liệu làm một
chiếc bút chì như trên gần nhất với kết quả nào dưới đây?
A. $9.7{\times}a~(\pu{đồng})$
B. $97.03{\times}a~(\pu{đồng})$
C. $90.7{\times}a~(\pu{đồng})$
D. $9.07{\times}a~(\pu{đồng})$
Đáp án:
```

图 5: 一个数学示例。

```
Dưới đây là các câu hỏi trắc nghiệm (kèm đáp án) về vật lí học
Đề bài: Năng lượng cần thiết để giải phóng một êlectron liên kết thành
êlectron dẫn (năng lượng kích hoạt) của các chất $\ce{PbS}$, $\ce{Ge}$,
$\ce{Si}$, $\ce{CdTe}$ lần lượt là: $0.30~\pu{eV}$; $0.66~\pu{eV}$;
$1.12~\pu{eV}$; $1.51~\pu{eV}$. Lấy $1~\pu{eV}=1.6\times10^{-19}~\pu{J}$.
Khi chiếu bức xạ đơn sắc mà mỗi phôtôn mang năng lượng bằng
$9.94\times10^{-20}~\pu{J}$ vào các chất trên thì số chất mà hiện tượng
quang điện trong xảy ra là
A. $2$
B. $3$
C. $4$
D. $1$
Đáp án:
```

图 6: 一个物理示例。

```
Dưới đây là các câu hỏi trắc nghiệm (kèm đáp án) về hoá học
Đề bài: Cho sơ đồ các phản ứng theo đúng tỉ lệ mol:
(a) $\ce{X + 4AgNO3 + 6NH3 + 2H2O ->[{t\degree}] X1 + 4Ag + 4NH4NO3}$
(b) $\ce{X1 + 2NaOH -> X2 + 2NH3 + 2H2O}$
(c) $\ce{X2 + 2HCl -> X3 + 2NaCl}$
(d) $\ce{X3 + C2H5OH <=>[{\ce{H2SO4} đặc, t\degree}] X4 + H2O}$
Biết $\ce{X}$ là hợp chất hữu cơ no, mạch hở, chỉ chứa một loại nhóm
chức. Khi đốt đốt cháy hoàn toàn $\ce{X2}$, sản phẩm thu được chỉ gồm
$\ce{CO2}$ và $\ce{Na2CO3}$. Phân tử khối của $\ce{X4}$ là
A. $118$
B. $138$
C. $90$
D. $146$
Đáp án:
```

图 7: 一个化学示例。

```
Dưới đây là các câu hỏi trắc nghiệm (kèm đáp án) về sinh học
Đề bài: Ở ruồi giấm, alen $A$ quy định thân xám trội hoàn toàn so với
alen $a$ quy định thân đen; alen $B$ quy định cánh dài trội hoàn toàn
so với alen $b$ quy định cánh cụt. Alen $D$ quy định mắt đỏ trội hoàn
toàn so với alen $d$ quy định mắt trắng. Phép lai $P$:
$\frac{AB}{ab}X^DX^d\times\frac{AB}{ab}X^DY$, thu được $F_1$. Trong tổng
số ruồi $F_1$, số ruồi thân xám, cánh cụt, mắt đỏ chiếm $3.75\pu\%$. Biết
rằng không xảy ra đột biến nhưng xảy ra hoán vị gen trong quá trình phát
sinh giao tử cái. Theo lí thuyết, có bao nhiêu phát biểu sau đây đúng?
I. $F_1$ có $40$ loại kiểu gen.
II. Khoảng cách giữa gen $A$ và gen $B$ là $20~\pu{cm}$.
III. $F_1$ có $10\pu\%$ số ruồi đực thân đen, cánh cụt, mắt đỏ.
IV. $F_1$ có $25\pu\%$ số cá thể cái mang kiểu hình trội về hai tính
trạng.
A. $2$
B. $3$
C. $4$
D. $1$
Đáp án:
```

图 8: 一个生物学示例。