教育中大型语言模型的多语言性能偏差

Vansh Gupta[★†] **Sankalan Pal Chowdhury**[★†]
Vilém Zouhar[†] **Donya Rooein**[‡] **Mrinmaya Sachan**[†]

{ guptav,spalchowd,vzouhar,msachan } @ethz.ch donya.rooein@unibocconi.it
[†]ETH Zurich [‡]Bocconi University

Abstract

大型语言模型（LLM）在教育环境中被越来越多地采用。这些应用程序超越了英文，但目前的 LLM 仍然主要以英语为中心。在这项工作中，我们确定其在非英语语言的教育环境中使用是否合理。我们评估了流行的 LLM 在四种教育任务中的表现：识别学生误解、提供针对性反馈、互动辅导，以及对六种语言（印地语、阿拉伯语、波斯语、泰卢固语、乌克兰语、捷克语）的翻译评分，此外还有英语。我们发现，在这些任务上的表现与训练数据中语言的代表性多少存在一定关系，资源较少的语言任务表现较差。尽管这些模型在大多数语言中表现尚可，但与英语相比，频繁的性能下降是显著的。因此，我们建议实践者在使用之前，首先验证 LLM 在目标语言中的教育任务表现良好。

1 引言

教育是一项多语言、多文化的事业。基于人工智能的技术最近显示出提高学生学习体验的潜力，世界各地的教育系统正越来越多地采用这些工具 (Gligorea et al., 2023) 。从个性化教学和针对性反馈到合适的内容生成和互动辅导，这些工具提供了解决重要教育挑战的方案 (Leon, 2024; Rooein et al., 2024; Mosher et al., 2024) 。大型语言模型如 GPT、Gemini 和 Llama (OpenAI, 2023; Team, 2024; Roumeliotis et al., 2023) 变得特别有影响力，早期证据表明它们能够支持教师或为学生学习提供支架 (Kasneci et al., 2023; Alqahtani et al., 2023) 。

尽管大多数这些大型语言模型（LLMs）在多语言语料库上进行了训练，但它们仍然以英语为中心。在教育环境中，未充分适应本地语言可能会降低其效用，并通过偏袒主导语言和文化加剧现有的不平等。在 LLMs 应用的每个领域都会出现多语言问题。尽管风险很高，但在教育领域应用却是十分重要的，因为 LLMs 已经被广泛使用。如果没有针对不同语言的教育任务进行严格评估，将 LLMs 应用于课堂可能会带来新的危害形式，包括错误信息、与课程的不一致或文化上不适当的内容。

在这项工作中，我们对前沿大型语言模型在多种语言下完成教育任务的能力进行了一项实证研究。我们确定了四个与教育相关的任务（识别学生的误解、提供针对性反馈、互动辅导和翻译评分），这些任务都有明确的与语言无关的指标。然后，我们在包括英语在内的六种语言（印地语、阿拉伯语、波斯语、泰卢固语、乌克兰语和捷克语）中评估了几种前沿大型语言模型（Claude、Gemini、GPT4o、Llama 和 Mistral）的表现。

我们的结果表明，尽管英语的表现仍占主导地位，但至少对于 GPT4o 和 Gemini-2.0-flash 这两个最佳模型而言，其他语言与之差距不大。我们还发现，与英语提示相比，使用任务语言的提示很少有帮助。

★ Co-first authors sorted alphabetically.

[0]We release the collected dataset and code publicly at github.com/eth-lre/multilingual-educational-llm-bias. The dataset comprises 91,000 automatically evaluated model outputs across six languages, four tasks, and five models.

2 方法

我们根据三个期望选择我们的任务集合：

- 与教育相关：我们专注于 LLMs 在担任导师、教师或教学助理角色时遇到的任务。我们不涉及像问题解答或数学题解答之类的任务，这些任务虽然可能与教育相关，但更是主要在其他情境中研究的一般性任务。
- 具有语言成分：我们避开那些仅使用符号表示的任务，例如，解数学方程。如果方程以数学符号提供，那么在不同语言之间任务将保持不变，使得多语言性能的问题变得无关紧要。
- 语言不变评价：最后，我们需要在不同语言间保持可比性的评价指标，以便高效地比较不同语言的性能。这意味着我们不能依赖像 BLEU 或 COMET (Papineni et al., 2002; Rei et al., 2020) 这样的语言相关指标。

基于这些，我们选择了以下四个任务：

任务 1：错误观念识别。 教学的一个重要方面是解决学生的错误概念，这首先需要识别学生的错误概念 (Liu et al., 2023) 。我们在 扩展边缘检测插值方法 数学问题数据集的基础上构建了这一任务，该数据集包含数千道四选一的选择题。对于许多错误的选项，我们有专家注释的错误概念，这些概念可能导致学生选择该选项。我们利用这些概念来构建我们的任务。LLM 会被提供一个选择题，学生的（错误）答案，以及四个可能的错误概念。候选的错误概念包括由专家识别的真实错误概念和从数据集中其他存在的错误概念中随机选择的三个干扰项。LLM 必须从这四个选项中选出正确的错误概念（见 Example 1 了解示例）。我们通过报告预测学生错误概念的准确性来评估 LLM 的表现。由于模型必须从四个选项中选择一个，随机基准的准确率为 25 %。

任务 2：反馈选择。 纠正学生错误概念的一个关键步骤是生成反馈以缓解这些错误。上述讨论的 扩展开边缘插值 数据集还包括我们在此部分中使用的所有选项的反馈信息。LLM 再次获得一个选择题、学生的答案，并且此时有一组四个可能的反馈，LLM 必须从中选择与学生答案相对应的反馈。请注意，虽然有 4 种可能的反馈，但其中一个对应正确答案。这个容易识别，因为它强化了学生的回答，而对应错误答案的反馈则试图让学生意识到他们的错误。作为示例，请参阅 2 两部分中的选项 C，它们是各自问题中唯一没有以负面语气开头的选项。因此，如果选择的答案也是问题的正确答案，LLM 可能会使用一些浅层语义来选择正确的反馈，而这是我们想要避免的。因此，我们确保所选答案始终不正确。随机基准的准确率为 25 % ，如果在错误答案的反馈中进行选择，准确率为 33 % 。

对于更复杂的误解，单轮反馈通常不足以解决问题，修正误解需要学生与教师之间进行多轮对话，也称为辅导。这涉及教师 LLM（大型语言模型）试图帮助学生识别并纠正其解答中的错误。我们通过让教师 LLM 辅导一个较弱的 LLM，作为学生，来评估其辅导能力。教师和学生都获得了问题，但只有教师 LLM 可以访问正确答案。学生 LLM 被指示除非看到强有力的理由，否则坚持错误的解答。教师和学生轮流发送信息，教师的目标是让学生模型得出正确答案，而不直接透露答案。如果学生 LLM 提出答案，则认为教师取得成功。如果教师在学生得到答案之前透露了答案，则被视为告知。如果在没有告知的情况下取得成功，则称为调整后的成功。该任务最终通过辅导得分进行评估，该得分是成功率和调整后成功率之间的调和平均值。

这项任务与该列表中的其他任务至少在两个重要方面有所不同。首先，它是一个多轮对话任务，因此没有猜测答案的空间。其次，最终评估取决于学生 LLM 的表现，因此学生 LLM 的多语言能力也限制了该任务的适用性。这些因素使得这个任务对 LLM 来说运行得更慢并且更复杂。

一个在语言学习中使用大型语言模型(LLM)增多的常见教育领域。这个领域的一个代表性任务就是给学生提供的翻译打分。虽然我们缺乏跨语言的适当数据集，这些数据集包含翻译及其相应的评分，但我们可以通过以下事实来近似这个任务：一段句子的机器翻译应该比一个单词被随机单词替换的精确翻译获得更高的评分。

我们使用来自 Duolingo 的英语 → 西班牙语 SLAM 数据集 (Settles, 2018) 的英语句子，这些句子被机器翻译成其他语言。我们选择该数据集是因为它旨在用于翻译，因此应包含较少难以翻译的句子。我们过滤掉不以句点结尾或少于五个单词的简单句子。对于每个翻译的句子，我们通过将句子中的一个单词替换为从数据集中的其他句子中随机选择的不同单词，来创建一个相应的扰乱翻译，从而破坏翻译的流利性和准确性。LLM 将对原始版本和扰乱版本分别进行评分，评分从 1（完全不正确）到 5（完美不误），并期望给扰乱版本分配一个严格较低的分数。因此，一个随意给出分数的模型将得分约为 40 %。

我们选择了六种语言进行实验：印地语、泰

	Language family	Script	Wikipedia	CommonCrawl	Speakers
English	Germanic	Latin	6973K	42.8 %	1500M
Hindi	Indo-Iranian	Brahmic	165K	0.20 %	609M
Arabic	Afro-Asiatic	Abjad	1259K	0.68 %	411M
Farsi	Indo-Iranian	Abjad	1034K	0.74 %	127M
Telugu	Dravidian	Brahmic	111K	0.02 %	96M
Ukrainian	Slavic	Cyrillic	1371K	0.62 %	39M
Czech	Slavic	Latin	566M	0.10 %	12M

Table 1: 语言信息、使用者数量（民族语 2025 ），以及被测试语言在自然语言处理中的全球表现（2025 年 3 月维基百科文章和 CommonCrawl 中的比例）。

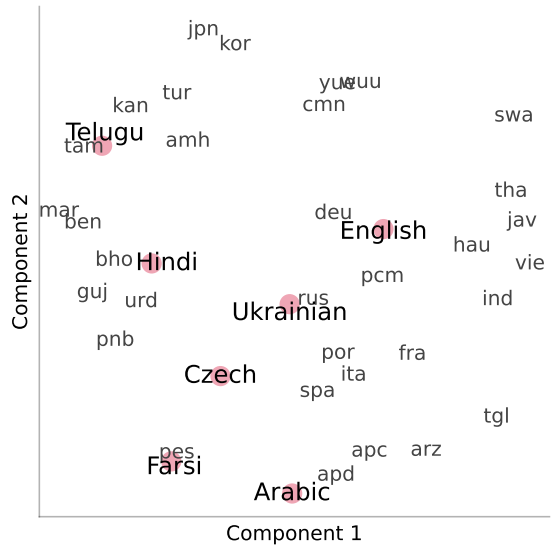

Figure 1: 基于 URIEL/lang2vec 的语法特征对语言进行多维尺度缩放投影。我们实验中使用的语言已被突出显示并显示全名，其他语言以 ISO 639/集 2 的形式展示。

卢固语、波斯语、乌克兰语和捷克语，此外还包括英语用于比较。这些语言选择反映了多样的语言特性、训练数据中的不同表示水平以及不同的语言家族。印地语（印度-雅利安语）和泰卢固语（达罗毗荼语）代表了来自印度次大陆的重要语言，它们使用婆罗米系文字，并且在 CommonCrawl 和维基百科中均表示不足。波斯语和阿拉伯语提供了对从右到左的辅音音素文字上大型语言模型性能的洞察，而乌克兰语和捷克语则让我们通过分别使用西里尔字母和拉丁字母研究中等资源形态丰富语言中的泛化能力。

为了评估我们所选语言的类型学多样性，我们使用了 URIEL 类型学数据库 (Littell et al., 2017) 和 `lang2vec` ，它们基于一系列类型学、谱系和地理特征提供了语言的稠密向量表示。根据该软件包的推荐，我们提取了 `syntax` 特征，并对 40 种语言集进行了 k-NN 预测以补充缺失值，这些语言集是根据我们的核心实验语言和全球使用最广泛语言（根据 Ethnologue (Eberhard et al., 2025) ）构建的联合集。我们使用多维尺度法将每种语言特征向量投影到二维中，从而生成一个二维语言相似性图。这使我们能够可视化（见 Figure 1 ）我们所选语言的相对句法多样性，并确认这些语言跨越了一个广泛的类型学空间。可视化表明，我们的语言选择（突出显示部分）在类型学景观中分布良好。

我们通过机器翻译获取所有上述语言的任务。按照 GPT4 技术报告的 (**?**, Figure 5) ，我们使用 Azure Translate 将所有示例翻译成目标语言。然而，这为非英语语言的任务引入了额外的噪声源。事实上，在手动审核一些翻译后，确实看起来这些翻译虽然不错，但不如它们的英文版本易于理解。这一发现进一步得到了翻译的 $\text{COMET}_{23}^{\text{DA.XL}}$ (Rei et al., 2023) 分数的证实（见 **??** ）。这意味着我们观察到的英语和其他语言表现之间的任何差异不能被确凿地归因于被测试的 LLM。然而，我们仍然可以在同一种语言中比较不同 LLM 的性能，因为所有 LLM 使用的是相同的翻译。此外，如果至少有一个 LLM 在给定语言的任务中表现良好，我们可以相当确信该任务语言对的翻译也是足够好的。

我们评估了六种因其多语言能力而受到称赞的最先进的大型语言模型：GPT-4o、Gemini 2.0 Flash、Claude 3.7 Sonnet、Llama 3.1 405B、Mistral Large 2407 和 Command-A。我们将所有采样参数保持为默认设置。对于提示，我们使用了一种简单的流水线推理提示方法，首先要求模型解释为何选择某个答案，然后在单独的提示中要求选择该答案。根据文献，目前不清楚将提示本身翻译成目标语言或保持其为英文是否有利，因此我们尝试了两种选项。

对于每个任务，我们使用 1000 个样本来报告我们的结果，这些样本是从数据集中随机抽取的，除了辅导任务中的 200 个样本，该任务是多轮对话。

Language	English prompt						Translated prompt					
	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	97.6 %	96.2 %	95.1 %	94.0 %	95.0 %	95.3 %	97.6 %	96.2 %	95.1 %	94.0 %	95.0 %	95.3 %
Hindi	· 94.5 %	· 93.2 %	· 91.9 %	· 89.8 %	· 91.8 %	· 93.2 %	· 95.5 %	· 93.6 %	· 89.6 %	· 90.4 %	· 90.6 %	· 91.4 %
Arabic	· 95.9 %	· 93.0 %	· 92.0 %	⋆ 86.0 %	· 92.6 %	· 93.4 %	· 95.9 %	· 93.0 %	· 92.8 %	· 90.9 %	· 92.0 %	94.0 %
Farsi	· 94.8 %	· 93.3 %	· 93.0 %	· 87.5 %	· 92.7 %	· 93.1 %	· 95.1 %	· 94.4 %	★ 68.0 %	· 88.3 %	★ 66.9 %	· 93.6 %
Telugu	· 95.2 %	· 92.2 %	· 89.9 %	· 86.9 %	· 89.7 %	⋆ 85.5 %	· 94.2 %	· 90.8 %	★ 68.6 %	⋆ 83.6 %	★ 35.5 %	★ 77.9 %
Ukrainian	· 95.7 %	94.9 %	· 92.9 %	93.3 %	94.4 %	94.9 %	· 95.6 %	· 94.3 %	★ 56.6 %	· 90.4 %	94.2 %	93.9 %
Czech	96.9 %	95.1 %	94.5 %	92.3 %	94.5 %	94.1 %	96.6 %	95.8 %	★ 70.2 %	⋆ 81.6 %	★ 41.0 %	94.5 %

Table 2: 误解识别任务的结果（准确性）。我们用单侧 95 % 置信度 t 检验标记出显著低于英语的结果（至少 10 % = ★，至少 5 % = ⋆，否则为 · ）。

Language	English prompt						Translated prompt					
	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	53.4 %	38.2 %	17.0 %	51.1 %	48.5 %	39.7 %	53.4 %	38.2 %	17.0 %	51.1 %	48.5 %	39.7 %
Hindi	· 48.7 %	35.6 %	· 13.0 %	· 43.6 %	· 40.5 %	· 31.6 %	★ 32.1 %	★ 13.4 %	⋆ 6.2 %	· 44.3 %	★ 18.6 %	★ 18.8 %
Arabic	· 49.6 %	⋆ 28.7 %	· 13.9 %	· 45.3 %	⋆ 38.8 %	· 33.3 %	· 48.8 %	★ 10.7 %	16.3 %	48.1 %	★ 27.8 %	⋆ 28.9 %
Farsi	50.2 %	⋆ 27.9 %	· 11.3 %	· 44.9 %	· 41.3 %	⋆ 30.9 %	· 45.9 %	· 31.6 %	16.3 %	· 44.0 %	★ 33.5 %	· 35.5 %
Telugu	· 45.2 %	⋆ 27.6 %	· 10.4 %	· 43.4 %	★ 34.0 %	⋆ 26.3 %	★ 13.9 %	★ 12.7 %	⋆ 6.1 %	⋆ 37.7 %	★ 15.5 %	★ 9.5 %
Ukrainian	50.3 %	· 33.2 %	· 13.0 %	· 44.8 %	· 41.3 %	· 32.2 %	★ 35.9 %	★ 19.6 %	⋆ 8.1 %	52.8 %	★ 31.0 %	⋆ 27.2 %
Czech	49.9 %	37.8 %	· 14.1 %	· 46.5 %	· 41.6 %	⋆ 30.7 %	⋆ 42.7 %	⋆ 26.1 %	19.2 %	· 46.6 %	⋆ 35.5 %	· 35.6 %

Table 3: 反馈选择任务的结果（准确率）。我们用单侧 95 % 置信度 t 检验标记结果比英语显著低的情况（至少低 10 % = ★，至少低 5 % = ⋆，否则 · ）。

	Harmonic mean						Success/1-Telling					
	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	94.7 %	97.0 %	22.1 %	93.0 %	82.0 %	95.5 %	96.0/2.5 %	97.5/1.0 %	96.5/84.0 %	93.5/1.0 %	82.0/0.0 %	96.0/1.0 %
Hindi	90.5 %	92.7 %	24.2 %	⋆ 72.2 %	73.5 %	· 88.4 %	95.0/8.5 %	93.0/0.5 %	89.5/75.5 %	77.5/10.0 %	73.5/0.0 %	91.0/5.0 %
Arabic	91.4 %	89.7 %	24.3 %	· 84.2 %	75.2 %	87.4 %	94.5/5.9 %	90.0/0.5 %	91.0/77.0 %	86.0/3.5 %	75.5/0.5 %	93.0/10.5 %
Farsi	· 85.6 %	⋆ 81.3 %	28.7 %	77.2 %	· 65.8 %	· 77.8 %	89.0/6.5 %	87.5/11.5 %	91.5/74.5 %	78.0/1.5 %	69.5/7.0 %	91.0/23.0 %
Telugu	★ 50.1 %	★ 39.5 %	27.7 %	· 58.9 %	★ 2.9 %	★ 40.7 %	77.5/40.5 %	77.5/51.0 %	85.5/69.0 %	61.0/4.0 %	59.0/57.5 %	63.5/33.5 %
Ukrainian	91.2 %	91.5 %	23.5 %	· 81.2 %	71.5 %	90.9 %	93.0/3.5 %	92.0/1.0 %	91.5/78.0 %	84.0/5.5 %	71.5/0.0 %	93.5/5.0 %
Czech	★ 43.8 %	★ 44.1 %	17.2 %	70.2 %	★ 2.9 %	★ 21.5 %	65.5/32.5 %	73.5/42.0 %	90.0/80.5 %	71.5/2.5 %	52.5/51.0 %	77.0/64.5 %

Table 4: 辅导任务的结果（调和平均数、成功率和告知）。在成功和告知中，如果结果显著低于英语（至少比英语差 10 % = ★，至少比英语差 5 % = ⋆，否则为 · ），则用单侧 95 % 置信 t 检验进行标记。告知指标被反转，即数值越高越好。

Language	English prompt						Translated prompt					
	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
Hindi	91.5 %	74.1 %	92.1 %	77.6 %	82.4 %	77.9 %	93.8 %	88.5 %	56.5 %	86.5 %	87.6 %	81.3 %
Arabic	98.6 %	97.9 %	99.2 %	98.8 %	97.5 %	99.0 %	98.8 %	98.3 %	67.2 %	98.6 %	97.8 %	97.9 %
Farsi	95.3 %	93.5 %	96.0 %	96.4 %	92.3 %	96.6 %	96.8 %	96.0 %	67.0 %	96.4 %	94.1 %	96.2 %
Telugu	77.2 %	33.7 %	81.0 %	51.9 %	48.7 %	25.2 %	82.8 %	46.8 %	40.7 %	82.1 %	67.1 %	15.6 %
Ukrainian	98.0 %	97.3 %	96.9 %	96.5 %	97.3 %	98.3 %	98.1 %	97.9 %	85.3 %	97.7 %	98.4 %	98.2 %
Czech	98.7 %	98.3 %	98.9 %	98.3 %	97.5 %	98.8 %	99.3 %	98.8 %	80.8 %	98.7 %	99.5 %	99.2 %

Table 5: 翻译评分任务的结果（准确率）。

3 结果

在本节中，我们描述了五个流行的大型语言模型在 Section 2 中描述的四项任务上的结果。主要结果如 Tables 2 to 5 所示。

英语对大型语言模型来说最简单。 英语与其他语言之间的差距通常是很大的。平均来看，在所有任务（不包括翻译）和模型中，英语的表现为 70.9 %，而印地语为 63.1 %，捷克语为 55.3 %，乌克兰语为 67.8 %，泰卢固语为 49.7 %，波斯语为 66.2 %，阿拉伯语为 67.4 %。这本身并不能明确说明这种差距是由于大型语言模型的能力不足还是翻译质量较差所致。泰卢固语效果不佳主要是由 Command-A 和 Mistral 造成的。前者的表现不足为奇，因为在我们的列表中，泰卢固语是唯一一种未被 Command-A 官方支持的语言 (Cohere et al., 2025)。另一方面，Mistral 仅列出了 12 种支持的语言，而我们只测试了其中的印地语和阿拉伯语。泰卢固语在 CommonCrawl 和维基百科中的代表性也最低，因此这个结果是可以预期的。对于捷克语辅导表现低下的手动分析显示，交流过程中各种语言的正式程度不断转换，令人分心。此外，捷克课堂上使用的语言具有特殊性，可能在互联网上并不常见。

模型性能和一致性。 Mistral 在非英语语言中的一致性最差（平均偏差 [1] =0.204）。例如，它在捷克语和泰卢固语的辅导任务中完全失败，尽管在同一任务的其他语言中表现尚可。Command-A 稍好一些（平均偏差 $^{??}$ =0.179）。此外，即便我们将泰卢固语排除在分析之外，Command-A 仍然是列表中倒数第二差。另一方面，Gemini 是最一致的（平均偏差 $^{??}$ =0.075），而且也是第二好的表现者（平均得分为 74.3 %）。GPT4o 是表现最好的模型（平均 76.9 %），而 Claude 表现最差（平均 50.0 %），主要由于反馈和辅导任务。

尽管与误解识别任务相似，但在反馈任务中的表现最差。尽管 Claude 仍然是表现最差，同时表现比随机更差的模组，所有模型都在这一任务中表现不佳。在 Table 9 的进一步分析中显示，所有模型倾向于默认与正确答案对应的反馈，其思维过程是“无论学生的错误是什么，这是提供给学生最多正确答案信息的反馈。”大多数模型在翻译评估任务中表现良好，准确性甚至高于人类注释者，这些注释者接受了与之类似扰动的注意力检查 (Kocmi et al., 2024; Zouhar et al., 2025) 。它们在误解任务中也表现不错，大多数得分百分数（至少在英文提示情况下）都在 90 分以上。辅导任务似乎在各个模型和语言之间表现最不一致。一般来说，所有模型在捷克语和泰卢固语中表现不佳，同时 Claude 在所有语言中都表现不佳。避免直白陈述似乎是所有模型在这方面面临的更大挑战，尽管成功率也不太一致。

英文和翻译的提示。 除了辅导任务外，使用英语提示比使用翻译提示表现更好（平均 71.6 % 和 66.2 %）。例外是 GPT、Llama、Gemini 和 Mistral 在翻译任务中，以及 Gemini 在反馈任务中，虽然大多数情况下差别不是很大。注意，有些较差的表现可能归因于提示被翻译和检查正确性，而不是直接用目标语言书写，这可能引入一些翻译腔。无论如何，我们认为最好保持提示为英语。对于设计多语言应用的英语开发人员来说，一个额外注意点是，保持提示为英语可以确保思维链保持英语，使得进行合理性核查更加容易。

4 相关工作

在庞大的多语言数据集上训练的大型语言模型在诸如文本生成、翻译和对话 (Brown et al., 2020) 等任务中表现出色，使其成为智能辅导系统（ITS）(Corbett et al., 1997; Pal Chowdhury et al., 2024) 中的有前途的工具。先前的研究探索了它们在教育环境中的应用，例如动态学生互动 (Schmucker et al., 2023) 、模拟专家和新手行为 (Liu et al., 2023) ，以及解决数学文字问题 (Opedal et al., 2023) 。

除了数学背景，LLMs 也被探索用于其他形式的学习。(Cui and Sachan, 2023) 研究了 LLMs 在语言学习者的自适应和个性化练习生成中的应用，而 (Wang et al., 2023) 则研究了对话式辅导策略如何帮助学生理解。此外，LLMs 还被用来评估语法正确性和翻译准确性 (Kocmi and Federmann, 2023; Omelianchuk et al., 2024; Freitag et al., 2024) , 促进自动化作文评分 (Pack et al., 2024) ，并在二语写作中提供纠正性反馈 (Han et al., 2024) 。

虽然 LLM 在英语方面表现出色，但它们在其他语言方面的能力往往有所不同，这反映了在预训练语料库中高资源语言的过度代表性。例如，Koto et al. (2023) 介绍了 IndoMMLU，它揭示了印尼语和英语环境之间显著的性能差异。类似地，Holtermann et al. (2024) 调查了 LLM 在 137 种语言中的表现，并将性能差异归因于标记化策略。Li et al. (2024); Armengol-Estapé et al. (2022) 进一步发现预训练数据比例与性能之间存在很强的相关性，再次确认了高资源语言和低资源语言之间的差距。对于加泰罗尼亚语，Armengol-Estapé et al. (2022) 发现尽管 GPT-3 在生成任务中表现良好，但其理解能力受到该语言中等代表性的限制。

最近的工作开始更加直接地解决多语言教育问题。Cohere For AI 的 Kaleidoscope (Salazar et al., 2025) 和 Aya (Üstün et al., 2024) 等项目旨在支持文化多样的语言，而 SEA-HELM (Susanto et al., 2025) 和 ECLeKTic (Goldman et al., 2025) 则分别强调东南亚和跨语言背景下的文化基础评估。这些努力突出了需要多语言基准来超越以英语为中心的评估。

先前的教育学研究倾向于评估单一的大语言模型在单语环境中的表现。我们通过在多任务中进行基准测试填补了这一空白。具体而言，我们在多个模型和语言中进行零样本实验，以更好地分析它们在现实世界中的适用性。

5 结论

我们分析了六个著名的最新大语言模型在除了英语之外的六种语言上执行四项教育任务的表现。我们的发现是，虽然英语的表现仍然优于其他语言，但与其他模型的表现落差并不总是很大。特别是，我们发现 GPT4o 和 Gemini 2.0 在所有语言上的表现始终如一，只有少数例外。我们还注意到，在解决多语言任务时，英语提示效果与目标语言书写的提示效果相当，

[1] 我们计算每个任务在六种语言中的标准差，然后计算平均值。

甚至更好。这为轻松将为英语开发的应用程序移植到不同语言中提供了机会。然而，我们也注意到某些模型在某些任务和语言上的表现可能会很差，因此建议在部署之前，先验证特定模型在特定语言上的具体教育相关任务的表现是否良好。

6 局限性

所展示的实验可以自然地扩展到更多的语言。选定的语言代表了作者熟悉度与语言多样性之间的平衡，这对于有意义的定性分析是必要的，并由它们在 URIEL 特征空间中的分布所证明。同样地，我们只涵盖了六种 LLM。在这两种情况下，实验成本（参见 Table 6 ）都变得过于昂贵，这促使了本文中的数据发布以支持进一步研究。

此外，如前所述，翻译质量仍然是一个关注点。一个更为彻底的评估将涉及每个任务的人力翻译，类似于多语言基准测试 MMLU (Xuan et al., 2025) ，但若要对我们所有的任务这样进行处理将会消耗大量资源。

最后，这组任务并不能完整地代表教育领域中的问题，主要是因为大多数更复杂的任务缺乏明确的与语言无关的度量标准。

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , pages 4232–4267. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) , pages 2598–2637. Association for Computational Linguistics.

Mistral AI. 2024. Large enough: Introducing mistral large 2. Accessed: 2024-09-08.

Bashar Alhafni, Sowmya Vajjala, Stefano Bannò, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2024. LLMs in education: Novel perspectives, challenges, and opportunities. Preprint , arXiv:2409.11917.

Abdullah M. Almasoud, Muhammad Rafay Naeem, Muhammad Imran Taj, Ibrahim Ghaznavi, and Junaid Qadir. 2025. Toward inclusive educational AI: Auditing frontier LLMs through a multiplexity lens. ArXiv , abs/2501.03259.

Tariq Alqahtani, H. Badreldin, Mohammed A. Alrashed, Abdulrahman I. Alshaya, S. Alghamdi, Khalid Bin saleh, Shuroug A. Alowais, Omar A. Alshaya, I. Rahman, Majed S Al Yami, and Abdulkareem M. Albekairy. 2023. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Research in social & administrative pharmacy : RSAP .

Anthropic. 2024. Introducing claude 3.5 sonnet. Accessed: 2024-09-08.

Sabrina Argoub. 2022. The NLP divide: English is not the only natural language - polis.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In Proceedings of the Thirteenth Language Resources and Evaluation Conference , pages 3056–3068. European Language Resources Association.

Benjamin S. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher , 13(6):4–16.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. Preprint , arXiv:2005.14165.

Peter A. Cohen, James A. Kulik, and Chen-Lin C. Kulik. 1982. Educational outcomes of tutoring: A meta-analysis of findings. American Educational Research Journal , 19(2):237–248.

Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, and 1 others. 2025. Command a: An enterprise-ready large language model. arXiv preprint arXiv:2504.00698 .

Albert T. Corbett, Kenneth R. Koedinger, and John R. Anderson. 1997. Chapter 37 - intelligent tutoring systems. In Marting G. Helander, Thomas K. Landauer, and Prasad V. Prabhu, editors, Handbook of Human-Computer Interaction (Second Edition) , second edition edition, pages 849–874. North-Holland, Amsterdam.

Peng Cui and Mrinmaya Sachan. 2023. Adaptive and personalized exercise generation for online language

learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 10184–10198. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. Ethnologue: Languages of the World , 28 edition. SIL International, Dallas, Texas. Online version: http://www.ethnologue.com.

Wikimedia Foundation. 2024. List of wikipedias by language group. Accessed: 2024-09-08.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In Proceedings of the Ninth Conference on Machine Translation , pages 47–81. Association for Computational Linguistics.

Ilie Gligorea, Marius Cioca, Romana Oancea, A. Gorski, Hortensia Gorski, and Paul Tudorache. 2023. Adaptive learning using artificial intelligence in e-learning: A literature review. Education Sciences .

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2025. ECLeKTic: A novel challenge set for evaluation of cross-lingual knowledge transfer. Preprint , arXiv:2502.21228.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. Preprint , arXiv:2407.21783.

Albert Gu and Tri Dao. 2023. Mamba: Linear-Time sequence modeling with selective state spaces. Preprint , arXiv:2312.00752.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2024. LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. In Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U) , pages 284–293. Association for Computational Linguistics.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with multiq. Preprint , arXiv:2403.03814.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. Preprint , arXiv:2305.07004.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. Preprint , arXiv:2310.06825.

Enkelejda Kasneci, Kathrin Seßler, S. Küchemann, M. Bannert, Daryna Dementieva, F. Fischer, Urs Gasser, G. Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, J. Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, T. Seidel, and 4 others. 2023. ChatGPT for good? on opportunities and challenges of large language models for education. Learning and Individual Differences .

Blanka Klimova, Marcel Pikhart, and Liqaa Habeb Al-Obaydi. 2024. Exploring the potential of ChatGPT for foreign language education at the university level. Frontiers in Psychology , 15:1269319.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation , pages 193–203. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In Proceedings of the Ninth Conference on Machine Translation , pages 1440–1453. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on IndoMMLU. Preprint , arXiv:2310.04928.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In Findings of the Association for Computational Linguistics: EMNLP 2023 , pages 13171–13189. Association for Computational Linguistics.

Maikel Leon. 2024. Leveraging generative AI for on-demand tutoring as a new paradigm in education. International Journal on Cybernetics & Informatics .

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. Preprint , arXiv:2404.11553.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers , pages 8–14. Association for Computational Linguistics.

Naiming Liu, Shashank Sonkar, Zichao Wang, Simon Woodhead, and Richard G. Baraniuk. 2023. Novice learner and expert tutor: Evaluating math reasoning abilities of large language models with misconceptions. Preprint , arXiv:2310.02439.

Itai Mondshine, Tzuf Paz-Argaman, Asaf Achi Mordechai, and Reut Tsarfaty. 2024. HeSum: a novel dataset for abstractive text summarization in Hebrew. In Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024) , pages 26–36. Association for Computational Linguistics.

Maggie A. Mosher, Lisa Dieker, and Rebecca Hines. 2024. The past, present, and future use of artificial intelligence in teacher education. Journal of Special Education Preparation .

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) , pages 394–403. Association for Computational Linguistics.

Nvidia. 2022. Transformer model.

Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) , pages 17–33. Association for Computational Linguistics.

Andreas Opedal, Niklas Stoehr, Abulhair Saparov, and Mrinmaya Sachan. 2023. World models for math story problems. Preprint , arXiv:2306.04347.

OpenAI. 2019. Language models are unsupervised multitask learners.

OpenAI. 2023. GPT-4 technical report. Preprint , arXiv:2303.08774.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence , 6:100234.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In Proceedings of the Eleventh ACM Conference on Learning @ Scale , L@S '24, page 5–15, New York, NY, USA. Association for Computing Machinery.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics , pages 311–318. Association for Computational Linguistics.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, and 15 others. 2023. RWKV: Reinventing RNNs for the transformer era. Preprint , arXiv:2305.13048.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. Preprint , arXiv:2305.09857.

Ricardo Rei, Nuno M. Guerreiro, Jos textasciitilde A© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In Proceedings of the Eighth Conference on Machine Translation , pages 841–848. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pages 2685–2702. Association for Computational Linguistics.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) , pages 54–67. Association for Computational Linguistics.

Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos. 2023. Llama 2: Early adopters' utilization of meta's new open-source pretrained model. Preprints .

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In Findings of the Association for Computational Linguistics: ACL 2022 , pages 2340–2354. Association for Computational Linguistics.

Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh,

Fabian Farestam, Angelika Romanou, Danylo Boiko, Dipika Khullar, Mike Zhang, Dominik Krzemiski, Jekaterina Novikova, Luísa Shimabucoro, Joseph Marvin Imperial, Rishabh Maheshwary, Sharad Duwal, Alfonso Amayuelas, Swati Rajwal, Jebish Purbey, and 25 others. 2025. Kaleidoscope: In-language exams for massively multilingual vision evaluation. Preprint , arXiv:2504.07072.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2023. Ruffle & riley: Towards the automated induction of conversational tutoring systems. Preprint , arXiv:2310.01420.

Burr Settles. 2018. Data for the 2018 duolingo shared task on second language acquisition modeling (SLAM).

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xian Bin Yong, Weiqi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. SEA-HELM: Southeast asian holistic evaluation of language models. Preprint , arXiv:2502.14301.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. Preprint , arXiv:2312.11805.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 15894–15939. Association for Computational Linguistics.

Lingzhi Wang, Mrinmaya Sachan, Xingshan Zeng, and Kam-Fai Wong. 2023. Strategize before teaching: A conversational tutoring system with pedagogy self-distillation. In Findings of the Association for Computational Linguistics: EACL 2023 , pages 2268–2274. Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. Preprint , arXiv:2503.10497.

Tiffany Zhu, Kexun Zhang, and William Yang Wang. 2024. Embracing AI in education: Understanding the surge in large language model use by secondary students. Preprint , arXiv:2411.18708.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. Preprint , arXiv:2406.12419.

Input	Options
Question: Which number is the greatest? Student answer: 6.079	A: Confuses powers and multiples, B (correct): When multiplying a surd by an integer to write as one surd, squares the number under the surd rather than the number in front, C: When ordering integers, orders from the digits at the end instead of the digits at the start, D: Does not understand that when multiplying both sides of an equation by an amount every term must be multiplied by the same amount
Question: What is the lowest common multiple of 8 and 4 ? Student answer: 4	A: Subtracts instead of adds when answering worded problems, B (correct): Confuses factors and multiples, C: Rounds up instead of down, D: Adds instead of multiplying when expanding bracket

Example 1: Two examples of the misconception identification task (English).

Input	Options
Question: 6 pencils cost č1.50. How much do 3 pencils cost? Student answer: 25p	A (student answer): I think you have made an arithmetic error when halving č1.50. Use short division to divide by two, B: I think you have used the incorrect notation for money. Consider how the monetary values in the question are written, C (correct answer): If 6 pencils cost č1.50, then 3 pencils cost half of č1.50, which is č0.75 or 75p., D: I think you have found the cost for one pencil. The question asks for the cost of 3 pencils.
Question: A film starts at 8.50pm. The film lasts 2 hours and 52 minutes. What time does the film finish? Student answer: 11.02pm	A (student answer): This isn't quite right. Remember that there are 60 minutes in an hour, not 100 :), B: I think you've confused your method a little. Noticing that 2 hours and 52 minutes is just 8 minutes less than 3 hours is super, just make sure you add and subtract in the correct directions though :), C (correct answer): Almost there! Take care to notice how many hours and minutes you're adding here. Is your answer 2 hours and 52 minutes later than 8.50pm?, D: Adding 2 hours to 8.50pm gives 10.50pm. Adding 10 minutes on takes us to 11.00pm, and adding the remaining 42 minutes gives 11.42pm.

Example 2: Two examples of the feedback selection task (English).

Math Problem	Student's (Incorrect) Solution	Correct Solution
Sam sells bread. He has a target of selling 120 crates of bread in a week. One week he was closed on Monday and Friday. Over the weekend he sold 20 crates. On Tuesday he sold 15 crates, on Wednesday 12 crates, and Thursday 18 crates. By how many crates was Sam off from his target for the week?	Sam had 5 days to sell bread because he was closed on Monday and Friday. He sold a total of 20 + 15 + 12 + 18 = 65 crates of bread from Tuesday to Thursday. Adding the 20 crates he sold over the weekend, Sam sold a total of 65 + 20 = 85 crates of bread in a week. Sam was off from his target by 120 - 85 = 35 crates of bread.	During the whole week Sam sold 15 + 12 + 18 + 20 = 65 crates. Sam was off his target by 120 - 65 = 55 crates.
Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?	Sophia used 4 out of the 12 gallons of gas in her tank, so there are 12-4 = 8 gallons of gas left in the tank. If Sophia can drive 100 miles on 4 gallons of gas, then she can drive 100/4 = 25 miles per gallon. Therefore, with 8 gallons of gas left in the tank, Sophia can drive 25 x 8 = 200 miles on a single tank of gas.	To find miles per gallon, divide 100 miles / 4 gallons = 25 miles per gallon. To find how far Olivia can go on a single tank, multiply 25 miles per gallon Œ 12 gallons = 300 miles.

Example 3: Two examples of the tutoring task.

English Source	Original Translation	Perturbed Translation	Language
It is a kind of tomato.	वह एक तरह का टमाटर है।	भाई एक तरह का टमाटर है।	Hindi
	.هذا نوع من الطماطم	.هذا نوع من التفاح	Arabic
	.این نوعی گوجهفرنگی است	.این رنگ گوجهفرنگی است	Farsi
	ఇది ఒక రకమైన టమోటా.	ఇది కనుగొంటాము రకమైన టమోటా.	Telugu
	Він впливають різновидом томатів.	Він є різновидом томатів.	Ukrainian
	Je to druh rajete.	matka to druh rajete.	Czech

Example 4: A single example of the translation grading task for non-English languages.

Model	API	Total	Miconception	Feedback	Tutoring	Translation
Mistral	Mistral API	$ 410	$ 135	$ 135	$ 90	$ 50
Claude	Anthropic	$ 470	$ 150	$ 150	$ 105	$ 75
Command	Cohere	$ 400	$ 125	$ 125	$ 85	$ 65
Llama	Together.ai	$ 450	$ 145	$ 145	$ 100	$ 70
GPT4o	Open AI	$ 60	$ 18	$ 18	$ 14	$ 10
Gemini	Google Genai	$ 20	$ 7	$ 7	$ 4	$ 2

Table 6: 实验的大致费用。不包括税费或货币转换费用。总成本约为 $ 1810，此外还有大约 $ 500 用于初步实验。

	English prompt						Translated prompt					
Language	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	0.0 %	0.7 %	0.1 %	2.1 %	0.0 %	0.9 %	0.0 %	0.7 %	0.1 %	2.1 %	0.0 %	0.9 %
Hindi	0.0 %	1.6 %	0.4 %	2.3 %	0.0 %	0.4 %	0.0 %	0.9 %	0.2 %	2.5 %	0.0 %	0.1 %
Arabic	0.1 %	1.7 %	0.2 %	2.1 %	0.0 %	0.2 %	0.0 %	1.1 %	0.3 %	2.2 %	0.0 %	0.1 %
Farsi	0.0 %	1.8 %	0.2 %	2.0 %	0.0 %	0.3 %	0.0 %	1.6 %	0.2 %	2.9 %	0.0 %	0.1 %
Telugu	0.0 %	0.1 %	0.0 %	2.2 %	0.0 %	0.4 %	0.0 %	0.3 %	0.1 %	1.7 %	0.0 %	0.0 %
Ukrainian	0.1 %	1.6 %	0.1 %	2.2 %	0.0 %	0.3 %	0.0 %	1.8 %	0.4 %	1.6 %	0.0 %	0.1 %
Czech	0.1 %	1.6 %	0.1 %	1.9 %	0.0 %	0.7 %	0.0 %	1.4 %	0.0 %	0.7 %	0.0 %	0.5 %

Table 7: 识别误解任务的响应错误率。

	English prompt						Translated prompt					
Language	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	0.0 %	0.3 %	0.0 %	1.3 %	0.0 %	0.0 %	0.0 %	0.3 %	0.0 %	1.3 %	0.0 %	0.0 %
Hindi	0.0 %	0.0 %	0.0 %	1.1 %	0.0 %	0.1 %	0.0 %	0.0 %	0.0 %	1.0 %	0.0 %	0.2 %
Arabic	0.0 %	0.0 %	0.0 %	1.5 %	0.0 %	0.1 %	0.0 %	0.0 %	0.0 %	2.1 %	0.0 %	0.2 %
Farsi	0.0 %	0.0 %	0.0 %	1.2 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	1.1 %	0.0 %	0.1 %
Telugu	0.0 %	0.0 %	0.0 %	1.7 %	0.0 %	0.1 %	0.0 %	0.2 %	0.0 %	1.8 %	0.0 %	0.1 %
Ukrainian	0.0 %	0.0 %	0.0 %	0.9 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	1.3 %	0.0 %	0.0 %
Czech	0.0 %	0.0 %	0.0 %	1.1 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	3.0 %	0.0 %	0.0 %

Table 8: 反馈选择任务的响应错误率。

	English prompt						Translated prompt					
Language	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	23.7 %	45.8 %	75.0 %	27.8 %	23.0 %	35.1 %	23.7 %	45.8 %	75.0 %	27.8 %	23.0 %	35.1 %
Hindi	30.3 %	42.3 %	79.6 %	34.9 %	29.9 %	47.9 %	55.5 %	78.9 %	87.7 %	33.2 %	68.9 %	71.3 %
Arabic	28.0 %	54.1 %	79.5 %	35.3 %	32.9 %	44.0 %	21.9 %	81.7 %	73.6 %	22.4 %	36.4 %	49.5 %
Farsi	28.5 %	52.5 %	82.5 %	31.6 %	32.6 %	45.5 %	21.6 %	52.1 %	75.1 %	29.3 %	30.3 %	28.9 %
Telugu	29.2 %	55.6 %	81.5 %	35.2 %	33.1 %	52.4 %	78.3 %	73.8 %	89.5 %	37.9 %	70.9 %	78.8 %
Ukrainian	27.3 %	49.4 %	80.3 %	32.7 %	33.5 %	45.2 %	47.3 %	69.7 %	87.1 %	20.7 %	46.7 %	49.7 %
Czech	27.9 %	39.5 %	80.2 %	30.2 %	31.8 %	49.0 %	34.9 %	59.5 %	67.8 %	23.0 %	33.1 %	38.0 %

Table 9: 反馈选择任务中默认正确答案的比率。

	English prompt						Translated prompt					
Language	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
Hindi	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.2 %	0.0 %	0.0 %	0.0 %
Arabic	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	3.7 %	0.0 %	0.0 %	0.0 %
Farsi	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Telugu	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.5 %	0.0 %	0.0 %	0.0 %
Ukrainian	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	1.5 %	0.0 %	0.0 %	0.0 %
Czech	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	2.8 %	0.0 %	0.0 %	0.0 %

Table 10: 翻译评分任务的响应错误率。

A 实验提示

A.1 任务：错误概念识别

我们使用了一系列的三个提示：

Systemprompt：

```
You are an expert math tutor who knows about all grade-school level math misconceptions. Your task is
 to select the accurate type of misconceptions your student has based on the (incorrect) answer he/
she gives to a multiple-choice math question. You will be given 4 misconceptions types. Your selected
 misconception type should correspond to the given question and answer. Explain your reasoning
```

*Usermessage*1：

```
Question: {QUESTION}
Selected Answer: {SELECTED_ANSWER}
Misconceptions:
A. {Misconception 1}
B. {Misconception 2}
C. {Misconception 3}
D. {Misconception 4}
```

与选定答案相对应的误解位置在每个问题中旋转。随后的助手信息被存储为思维链。之后，我们发送了第二条用户消息。

*Usermessage*2：

```
Now based on your above explanation, output the option corresponding to the correct misconception.
Only say 'A', 'B', 'C', or 'D' without any other text. Do not say anything else.
```

这一部分的回答是最终答案。我们会重新生成，直到收到 'A'、'B'、'C' 或 'D' 的答案，最多尝试 20 次。如果没有收到答案，则保存 'E' 的回应。

此方法适用于除了 Gemini 之外的所有模型。在 Gemini 的情况下，我们使用 `generate_content` 方法，该方法推荐用于非聊天任务，并允许单个用户消息。在这种情况下，获得连贯思维后，我们使用相同的系统提示进行新查询，但用户消息为：

Geminimessage：

```
You have previously given the following answer and explanation:
{COT}
Now based on your above explanation, output the option corresponding to the correct misconception.
Only say 'A', 'B', 'C', or 'D' without any other text. Do not say anything else.
```

注意，最后一部分与用户消息 2 相同
当使用翻译的提示时，系统提示、用户消息 2 和双子消息会被翻译成目标语言。

A.2 任务：反馈选择

Systemprompt：

```
You are an expert math tutor who specialises in providing precise and helpful feedback for grade-
school level math questions. Your task is to select the correct explanation for a student's given
answer to a multiple-choice math question.

You will be provided with:
- A math question
- A specific answer chosen by the student (which can be correct or incorrect).
- Four possible explanations (labelled A, B, C, and D).
Your selected explanation should accurately correspond to the given answer. Provide your reasoning
for selecting the explanation.
```

*Usermessage*1：

```
Question: {QUESTION}
Selected Answer: {SELECTED_ANSWER}
Feedbacks:
A. {Feedback 1}
B. {Feedback 2}
C. {Feedback 3}
D. {Feedback 4}
```

与所选答案对应的反馈位置在不同问题中会进行旋转。如果它放在位置 A、B 或 C，那么与正确答案对应的反馈将放在位置 D。否则，它将在位置 C。随后的助手信息被储存为思维链。然后，我们发送了第二条用户信息。

*Usermessage*2：

```
Now based on your above explanation, output the option corresponding to the correct explanation. Only
 say 'A', 'B', 'C', or 'D' without any other text. Do not say anything else.
```

对此部分的响应是最终答案。我们重新生成，直到收到 'A'、'B'、'C' 或 'D' 的答案，最多 20 次。如果没有收到答案，则保存 'E' 的响应。

这种方法适用于除 Gemini 之外的所有模型。如果是 Gemini，推荐使用 `generate_content` 方法，该方法适用于非聊天任务，并允许单个用户消息。在这种情况下，在获得思维链后，我们使用相同的系统提示进行新的查询，但用户消息为：

Geminimessage：

```
You have previously given the following answer and explanation:
{COT}
Now based on your above explanation, output the option corresponding to the correct explanation. Only
 say 'A', 'B', 'C', or 'D' without any other text. Do not say anything else.
```

。注意，最后一部分与用户消息 2 相同。在使用翻译后的提示时，系统提示、用户消息 2 和 Gemini 消息被翻译为目标语言。我们手动确保翻译后的格式保持不变。

A.3 任务：辅导

Student system prompt：

```
Student Persona: {STUDENT_PERSONA}

Math problem: {MATH_PROBLEM}

Student solution: {STUDENT_SOLUTION}

Context: You need to role-play the student, {STUDENT_NAME}, while the user roleplays the tutor. {
STUDENT_NAME} thinks their answer is correct. Only when the teacher provides several good reasoning
questions, {STUDENT_NAME} understands the problem and corrects the solution. {STUDENT_NAME} can use
calculator and thus makes no calculation errors. Send <EOM> tag at end of the student message.
```

Teacher system prompt：

```
A tutor and a student work together to solve the following math word problem.
Math problem: {MATH_PROBLEM}
The correct solution is as follows:
{GROUND_TRUTH}

You need to role-play the tutor while the user roleplays the student, {STUDENT_NAME}. The tutor is a
soft-spoken empathetic man who dislikes giving out direct answers to students, and instead likes to
answer questions with other questions that would help the student understand the concepts, so that
she can solve the problem themselves.

{STUDENT_NAME} has come up with a solution, but it is incorrect. Please start the conversation, one
line at a time, aiming to figure out what is {STUDENT_NAME}'s solution and what is wrong with it.
Then try to get her to fix it.
```

对话历史被格式化为用户-助理消息对，用于教师和学生角色。我们手动设置初始消息，以目标语言启动对话。

后续的助手信息被存储为思维链。之后，我们发送了第二条用户信息。

*Usermessage*2 :

```
Now based on your above explanation, output the final score from 1 to 5. Only say '1', '2', '3', '4',
 or '5' without any other text. Do not say anything else.
```

对该部分的响应是最终答案。我们会重新生成直到收到 '1'、'2'、'3'、'4' 或 '5' 的答案，最多重复 20 次。如果没有收到答案，则记录 '0' 的响应。

此方法适用于除 Gemini 之外的所有模型。在 Gemini 的情况下，我们使用 `generate_content` 方法，该方法推荐用于非聊天任务，并允许单一用户消息。在这种情况下，在获得链式思考后，我们用相同的系统提示进行新的查询，但用户消息如下：

Geminimessage :

```
You have previously given the following answer and explanation:
{COT}
Now based on your above explanation, output the final score from 1 to 5. Only say '1', '2', '3', '4',
 or '5' without any other text. Do not say anything else.
```

请注意，最后一部分与用户消息 2 相同。

这个序列对每个句子重复两次，一次使用原始翻译，一次使用扰动后的翻译。然后将分数进行比较。当使用英文提示时，LANGUAGE 字段设置为其英文外名，即 Hindi、Arabic、Farsi、Telugu、Ukrainian 和 Czech。当使用翻译提示时，系统提示、用户消息 2 和 Gemini 消息被翻译成目标语言。我们手动确保翻译后格式的保持。我们也使用语言的内名，即 हिन्दी, فارسی , عربی, తెలుగు, Українська, and Čeština 。

B 翻译质量

正如我们在限制部分提到的，一个大语言模型在某种语言上表现不佳，并不一定意味着大语言模型本身不好。这也可能意味着信息在翻译过程中丢失了。这尤其成问题，因为机器翻译系统很可能同样遭遇大语言模型一开始就面临的资源限制。因此，我们对我们熟悉的语言（即波斯语、阿拉伯语、捷克语和印地语）的小部分已翻译的问题进行了人工调查。对于每种语言，我们分别分析了反馈和误解任务中的 10 个问题，以及翻译评分任务中的 20 个问题。

在波斯语的情况下，唯一重复出现的错误是与数学符号有关，尤其是减号被放在数字的右侧，而不是应该在左侧。然而，这似乎是一个渲染问题，这归因于减号（‘ − ’，U+2212）经常被相似的连字符（‘-’，U+002D）所替代，从而让渲染程序误以为是在渲染文本。由于 LLM 接收的是原始 Unicode 编码作为输入，这不应该成为问题。除此之外，还有一些小的时态错误，但意思是清楚的。

在阿拉伯语中也观察到了符号位置的问题。此外，似乎还存在一些翻译错误。例如，这里用于描述图形移动的“travel”一词被翻译为“liyusaafir”，这更像是在说“旅行”。在翻译任务中，我们没有发现句子中的错误。在捷克语中，错误的主要来源是上下文相关术语的不当使用。例如，在翻译“co-interior（角）”这个词时，省略了“co”前缀，只翻译了“interior”部分。虽然在日常讲话中这样处理还算可以，但在数学术语中，这可能会造成混淆。尽管翻译使理解变得更加困难，但问题的核心意义得以保留。

在印地语中，我们发现了几个例子，由于译者对一词多义的误解，导致印地语句子对说印地语的作者来说难以理解。例如，“round”这个词本来是以“approach（接近）”的意义使用，却被翻译成了“circle（圆圈）”，而“property”这个词本来是以“quality（质量）”的意义使用，却被翻译成了“possessions（财产）”。此外，短语“Not Quite”被翻译成了类似“Not Enough”的意思，可能因为“quite”在印地语中没有对应的词。然而，考虑到上下文，使用“Almost”一词会更符合语气。不过，尽管有很多翻译对于标注文者来说难以理解，但反向翻译它们后，得到了相当不错的结果，这意味着信息没有丢失。

翻译练习显示出很少的错误，可能是因为设计的句子易于翻译。有一两处误译，但总体效果不错。一个小问题是，使用 Python 和正则表达式 `‘\b\w+\b’` 进行的词边界检测有时会在印地语中识别单个字符而不是整个单词。然而，生成的句子仍然有错误，只是与我们预期的错误类型不同。

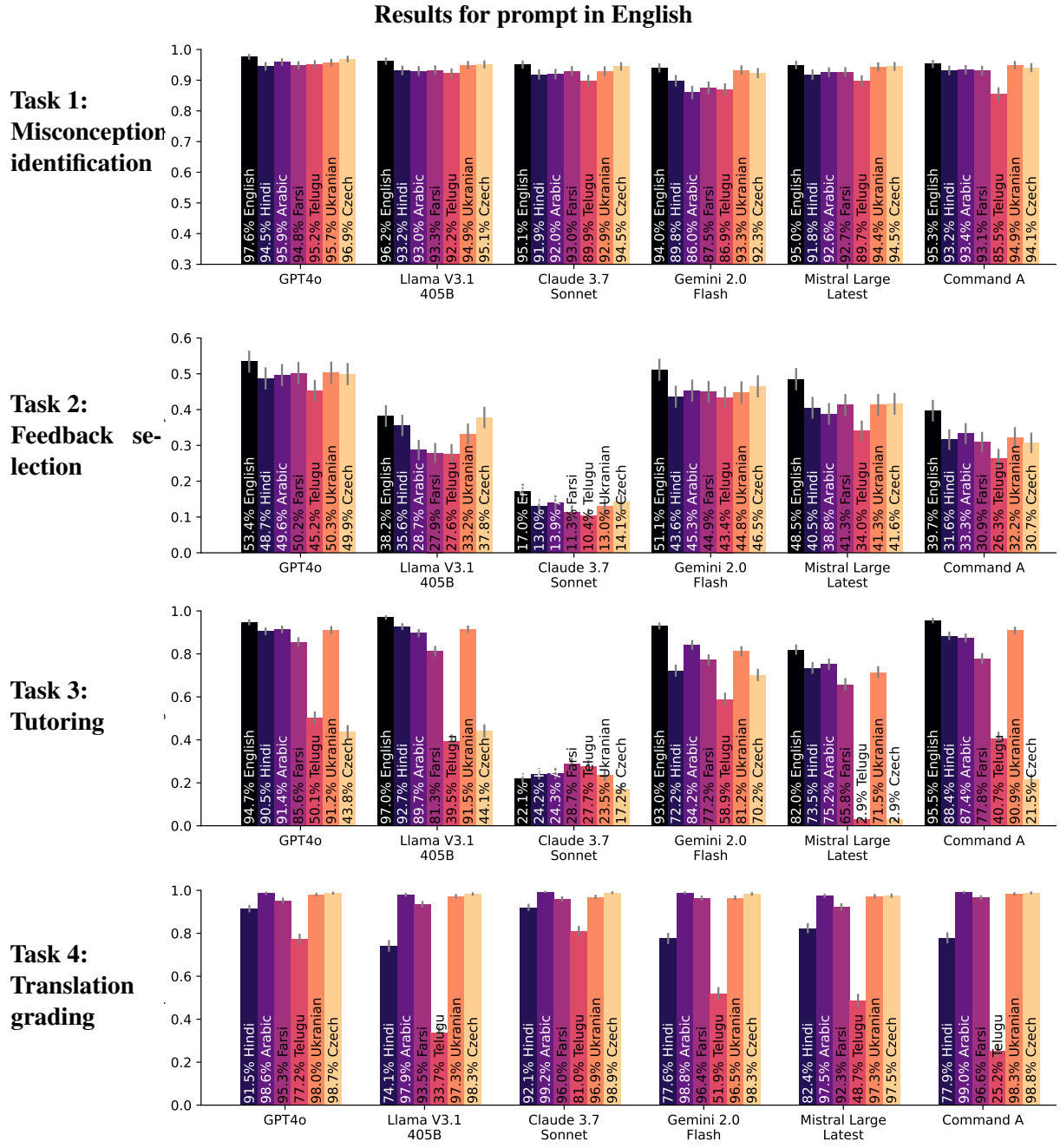

Figure 2: 五个大型语言模型在四个任务上的评估结果。误差条表示 95 % 的置信区间（t 检验）。MathDial 图展示在五次对话转轮后的辅导得分，大多数模型在五次话语对后得分持平。没有英文语言列，因为翻译评估使用英文作为源语言。所有得分范围从 0.0 到 1.0，分数越高越好，但彼此之间不可比。注意截断的 y 轴以获得更好的细节。可视化 Tables 2 to 5 。

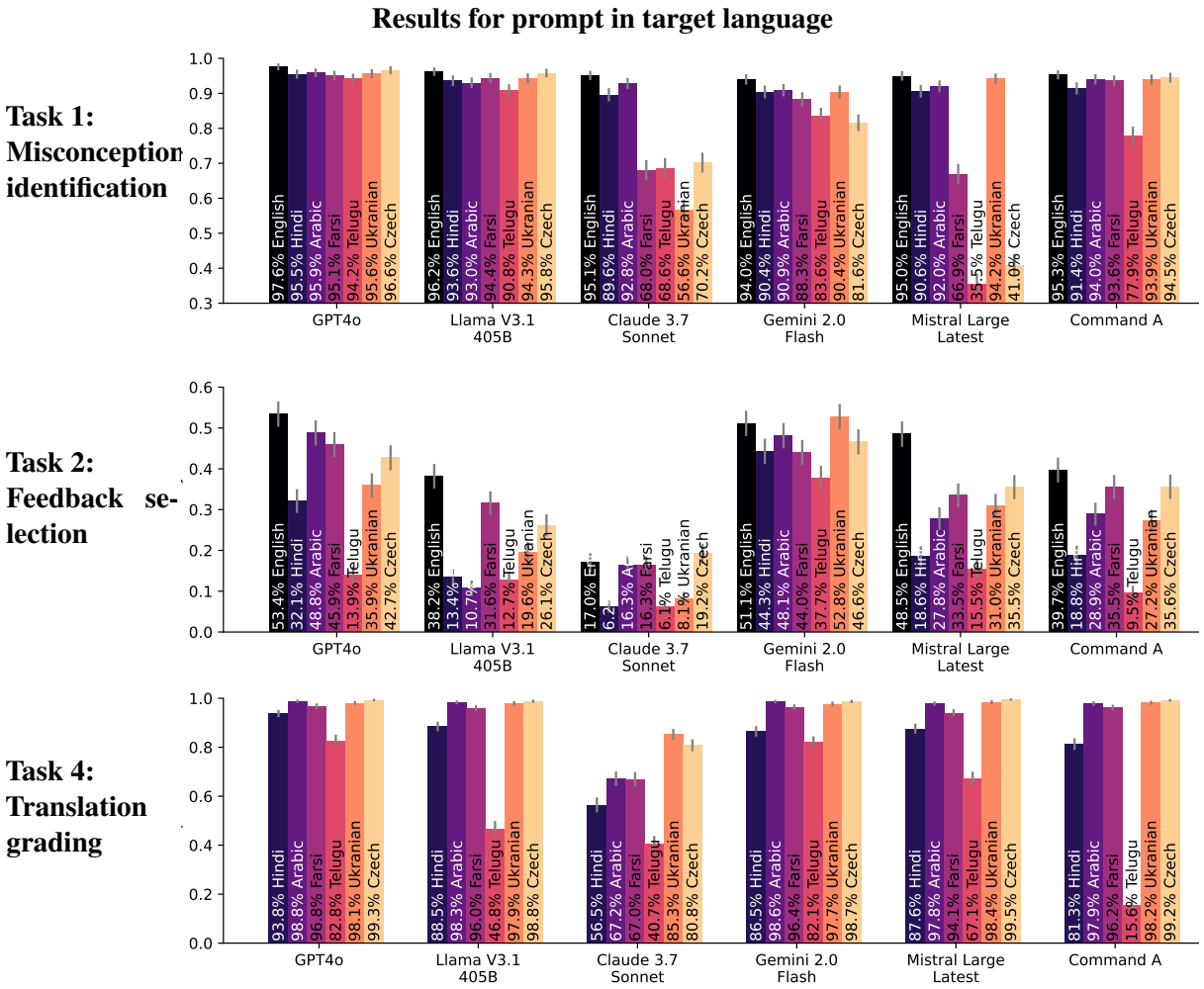

Figure 3: 对五种大型语言模型在四个任务上的评估结果。误差条显示 95 % 置信区间（t 检验）。MathDial 图显示五轮后的辅导得分，大多数模型在 5 对话后趋于平稳。没有英语语言栏是因为翻译评估使用英语作为源语言。所有分数范围从 0.0 到 1.0，分数越高越好，尽管它们之间不可相比。注意，为了更好地展示细节，y 轴被截断。可视化显示 Tables 2 to 5 。