arXiv:2603.26742v1 [cs.CL] 23 Mar 2026

多语言视觉语言模型的推理能力是否相同？印度语言跨语言视觉推理审计

Swastik R

印度信息技术学院赖久尔分校

摘要

视觉语言模型在数学、科学和空间推理基准测试中表现出色，但这些评测绝大多数都以英语进行。我开展了首个针对印度语言的跨语言视觉推理审计。我们使用 IndicTrans2 将 MathVista、ScienceQA 和 MMMU 中的 980 道题目翻译成印地语、泰米尔语、泰卢固语、孟加拉语、卡纳达语和马拉地语，并以 Gemini 2.0 Flash 对每种语言的 50 个样本进行交叉验证（译者间一致性为 0.79–0.84）。我们对涵盖 7B 开源模型至 GPT-4o 的 8 个视觉语言模型进行了评测，覆盖全部七种语言，共获得 68,600 条推理记录，其中包括纯文本和思维链消融实验。结果发现，从英语切换到印度语言后，准确率下降了 9.8–25 个百分点；达罗毗荼语系语言的降幅最多比印度-雅利安语支语言高 13.2 个百分点。思维链提示并未带来帮助，反而使孟加拉语（下降 −14.4 个百分点）和卡纳达语（下降 −11.4 个百分点）的表现变差，揭示出以英语为中心的推理链所造成的问题。专为 23 种语言打造的 Aya-Vision-8B，在达罗毗荼文字系统上的准确率仍下降 28.5 个百分点；仅靠多语言预训练无法迁移视觉推理能力。我将发布翻译后的基准以及所有模型输出。[1]

1 引言

印度的 2.6 亿学龄儿童主要在地方语言授课学校就读，这些学校的教学语言为印地语、泰米尔语、泰卢固语、孟加拉语、卡纳达语或马拉地语 [印度政府人力资源发展部， 2020]。教育科技平台开始整合视觉语言模型（VLM），用于自动化辅导和评估，由此引发了一个实际问题：这些模型能否用学生自己的语言推理数学和科学问题？

现有基准测试无法回答这个问题。针对印度语言的视觉语言模型评估考察了文化理解 [Surana 等人，2026年，Maji 等人， 2025], 事实性视觉问答 [Faraz 等人， 2025], 或广泛的多语言能力 [Khan 等人， 2025], 但没有一项专门考察数学、科学或空间推理。有人提出了面向印度的视觉语言模型架构，例如 Chitrarth [Khan 等人， 2025] 但目前还没有基准测试来检验这些模型或其他任何模型是否真正能够解答卡纳达语科学问题或泰米尔语几何题。

为填补这一空白，我使用 Indic-Trans 将 MathVista [Lu 等人、 2024],ScienceQA [Lu 等人、 2022], 以及 MMMU [Yue 等人、 2024 年 a] 中的 980 个视觉推理问题翻译成六种印度语言 2 [Gala 等人， 2023]。所有语言版本使用相同的图像，仅问题文本有所变化。我们评估了八种视觉语言模型（GPT-4o、Gemma 3-27B、Llama-4-Maverick、Qwen3-VL-30B、Qwen2.5-VL-32B、Qwen2.5-VL-7B、InternVL2.5-8B 和 Aya-Vision-8B），并在 Qwen2.5-VL-7B 上进行了仅文本和思维链消融实验，共生成 68,600 条推理记录（图 1）。

[1]代码和数据：https://github.com/QuantumByte-01/multilingual-vlm-reasoning-audit；HuggingFace：https://huggingface.co/datasets/Swastikr/multilingual-vlm-reasoning

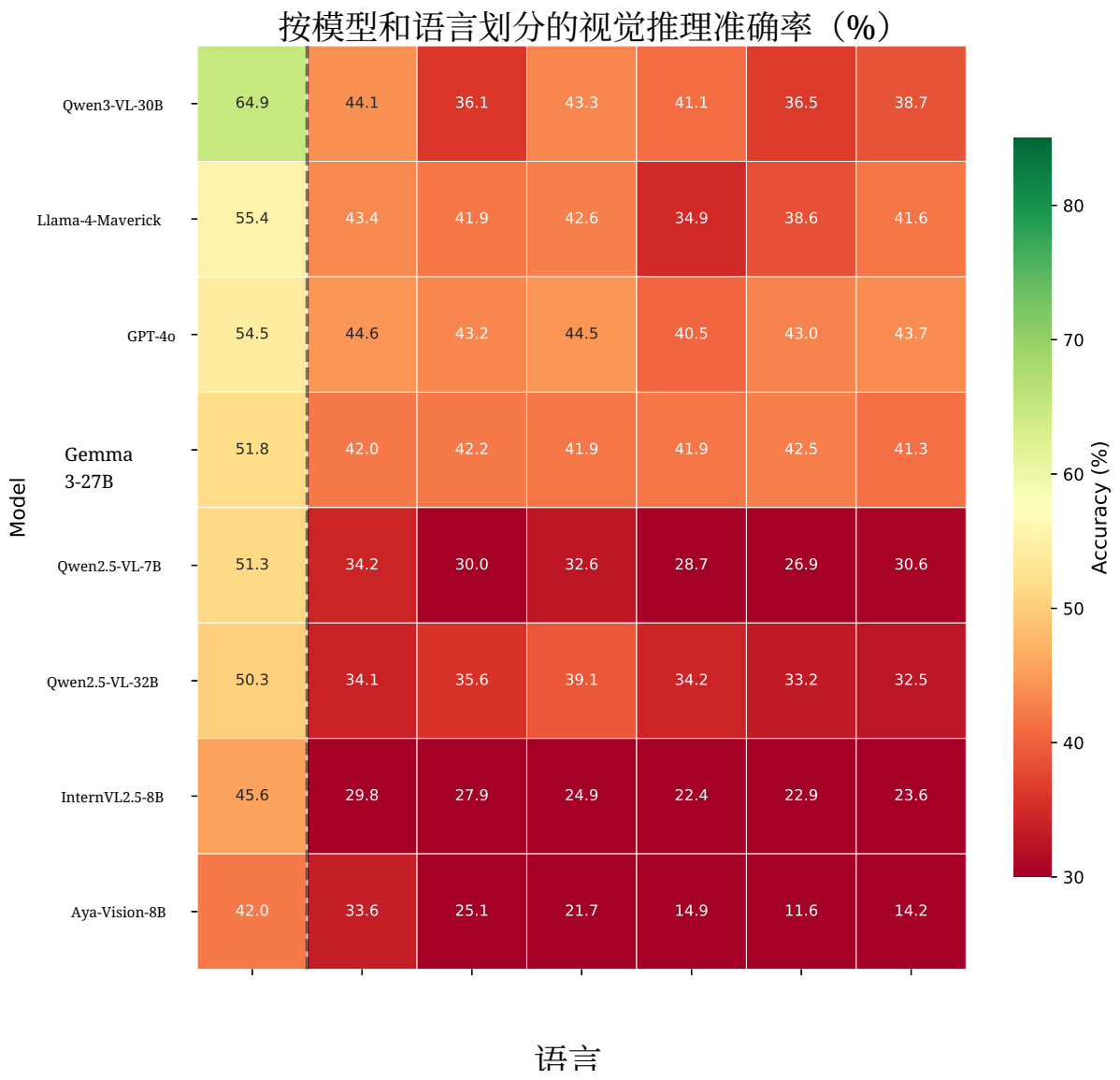


图 1：八种模型在七种语言上的准确率热力图。所有模型从英语（最左列）切换到印度语言时，准确率均有所下降；与印度-雅利安语支语言（hi、bn、mr）相比，达罗毗荼语系语言（ta、te、kn）的颜色始终更深。

主要发现如下：

- Gemma 3-27B 的英语准确率平均降幅最小（9.8 个百分点）；GPT-4o 的印度语言准确率最高（43.2%）；开源模型的准确率下降 15–25 个百分点。
- 对于所有模型而言，达罗毗荼语系语言都比印度-雅利安语支更难，其中 Aya-Vision-8B 的语系差距达到 13.2 个百分点。
- 思维链提示在印度语言中适得其反：相较于标准提示，孟加拉语准确率下降了 14.4 个百分点，卡纳达语下降了 11.4 个百分点，这表明模型的推理链被锁定在英语中。
- 移除图像会使英语准确率下降 15.5 个百分点，但对印度语言的影响仅为 4.9–9.7 个百分点，这意味着当文本理解能力较弱时，模型对视觉信息的利用不足。
- 在 Qwen2.5-VL 语系内从 7B 扩展至 32B，使印度语言上的准确率提高了 4.3 个百分点，而英语上的准确率下降了 1.0 个百分点；规模扩大拓宽了语言覆盖范围，却没有提升推理能力。

2 相关工作

印度语言视觉语言模型基准。IndicVisionBench [Faraz 等人， 2025] 涵盖 10 种印度语言，包含超过 3.7 万个问答对，但重点评估文化理解和光学字符识别，而非推理。Khan 等人提出的 [2025] BharatBench 框架涵盖语音、光学字符识别和嵌入，但未单独评估科学、技术、工程和数学任务。VIRAASAT [Surana 等人， 2026]以及 DRISHTIKON [Maji 等人， 2025] 评估文化知识和地理多样性。Chitrarth [Khan 等人， 2025] 提出了一种面向印度的视觉语言模型架构，但没有进行专门针对推理的评估。本研究的不同之处在于固定图像、仅改变问题语言，因此任何准确率变化都可直接归因于语言。

印度以外的多语言视觉语言模型评估。 MaRVL [Liu 等人， 2021] 测试了五种语言（包括泰米尔语）中基于视觉的推理能力，但依赖具有文化背景的图像，且早于当前的视觉语言模型。xGQA [Pfeiffer 等人， 2022] 将 GQA 扩展至七种语言，但不包括任何印度语言。VLURes [Atuhurra 等人，2025] 对包括乌尔都语在内的四种语言进行了基准测试。ALM-Bench [Vayani 等人， 2024] 以及 CVQA [Romero 等人， 2024] 涵盖了部分印度语言，但并未聚焦于推理。Pangea [Yue 等人，2024年b], 是一个支持 39 种语言的多模态大语言模型，是最接近的开源多语言项目，但它同样缺少以推理为重点的基准测试。

视觉推理基准。 MathVista [Lu 等人， 2024] 评估视觉情境中的数学推理；MATH-Vision [Wang 等人， 2024] 将其扩展至竞赛级问题。ScienceQA [Lu 等人， 2022] 涵盖带图表的幼儿园至高中科学。MMMU [Yue 等人， 2024年a]面向涵盖 30 个学科的大学级科学、技术、工程和数学。以上三者均仅支持英语；我将它们扩展至跨语言场景。

大语言模型中的多语言推理。 Shi 等人 [2023]表明，在 10 种语言的小学数学任务中，英语思维链的表现优于母语思维链，但研究未涉及图像。 Ahuja 等人 [2023]则将多语言生成评估扩展到更多任务和语言。 Saji 等人 [2025] 研究了语言推理模型在纯文本基准（MGSM、GPQA）上的多语言推理，发现无论提示使用何种语言，模型都会默认生成英语推理链。本文将这一研究方向拓展到视觉领域；在该领域，图像或许能部分弥补文本理解能力的不足。

IndianNLP 资源。 IndicTrans2 [Gala 等人， 2023] 提供针对全部 22 种印度宪定语言的开源神经机器翻译，其表现优于谷歌翻译和 NLLB-54B。我将其作为主要翻译系统，并辅以 Doddapaneni 等人 [2023]所述的更广泛的 Indic NLP 基础设施。

3 方法论

3.1 数据集构建

我从三个来源整理了 980 个视觉推理问题：

- **MathVista** [Lu 等人， 2024]：从 testmini 划分中选取 400 个问题，涵盖几何、代数、统计学、图表解读和文字题。排除了与语言无关的模式匹配题（例如智商测试序列）。
- **ScienceQA** [Lu 等人， 2022]：400 个配有图像的自然科学问题，涵盖 K-12 阶段的物理学、化学和生物学。
- **MMMU** [Yue 等人，2024年a]：180 个大学阶段的 STEM 问题（数学、物理学、工程学）。

每个问题都配有一张图像，该图像在不同语言版本中保持不变；只有文字会发生变化。

表1：包含文字系统、语系和近似母语使用者人数的目标语言。

语言	Code	文字系统	语系	母语（百万）
英语	en	拉丁字母	—	—
印地语	hi	天城文	印度-雅利安语支	340
孟加拉语	bn	孟加拉语	印度-雅利安语支	265
马拉地语	mr	天城文	印度-雅利安语支	83
泰米尔语	ta	泰米尔语	达罗毗荼语系	75
泰卢固语	te	泰卢固语	达罗毗荼语系	75
卡纳达语	kn	卡纳达语	达罗毗荼语系	44

表2：在每种语言的50个随机样本上，IndicTrans2 与 Gemini 2.0 Flash 的译者间一致性。

	hi	ta	bn	te	kn	mr
一致性	0.844	0.811	0.806	0.803	0.791	0.791

3.2 语言

我选取了六种印度语言，涵盖两个主要语系和五种文字系统（表1）。印地语和马拉地语使用相同的天城文，但语言各不相同，因此我们能够将词汇和训练数据效应与文字系统层面的效应区分开来。泰米尔语和卡纳达语同属达罗毗荼语系，但使用不同的文字系统，可提供互补的对照。

3.3 翻译

所有980道题目均使用 IndicTrans 翻译成每种目标语言2 [Gala 等人， 2023]。数学符号（π， $\surd$， $=$）、阿拉伯数字、国际单位制单位和答案标签（A/B/C/D）均予以保留，符合印度 STEM 教育的惯例。技术术语由 IndicTrans2 翻译；指令文本则针对每种语言进行完整翻译，以测试多语言指令遵循能力。

翻译质量通过以下方式验证：使用 Gemini 2.0 Flash 对每种语言随机抽取的 50 个样本进行独立翻译，并计算译者间一致性（表 2）。得分范围为 0.79（卡纳达语、马拉地语）至 0.84（印地语），均高于 0.65 的接受阈值。

3.4 模型

评估了八种视觉语言模型（表 3）。三种开源模型（Qwen2.5-VL-7B、InternVL2.5-8B、Aya-Vision-8B）通过 vLLM 在配备 192 GB 显存的 AMD MI300X GPU 上以 bfloat16 运行。云端模型则通过 Google AI（Gemma 3-27B）、DeepInfra（Qwen2.5-VL-32B、Qwen3-VL-30B、采用 FP8 的 Llama-4-Maverick）和 OpenAI（GPT-4o）访问。

两项消融实验使用 Qwen2.5-VL-7B 作为具有代表性的开源模型：（i）**无图像**，即移除图像，仅提供问题文本；（ii）**思维链** （CoT），即在提示中附加“Think step by step”，并使用目标语言。

3.5 评估协议

每个模型都会收到翻译后的问题、原始图像，以及一条特定语言指令，要求在选择题中给出字母答案（A–D），或在开放式问题中给出数字。为确保可复现性，温度设为 0。

对于选择题（980 题中的 762 题），我会从回答中提取第一个独立出现的 A–D 字母，同时也识别天城文及其他文字系统的特定选项标签。对于自由作答的数值题（980 题中的 218 题），会将特定文字系统的数字（天城文、孟加拉语、泰米尔语、泰卢固语、卡纳达语）转换为

Tab表 3：评估的模型。“总体”是所有七种语言的宏平均准确率 .

模型	访问权限	参数量	总体（%）
GPT-4o [OpenAI， 2024]	OpenAI API	—	44.8
Qwen3-VL-30B [Bai 等人，2025a]	DeepInfra	30B	43.5
Gemma 3-27B [Gemma Team、Google DeepMind， 2025]	Google AI	27B	43.4
Llama-4-Maverick [MetaAI， 2025]	DeepInfra	17B×128E	42.6
Qwen2.5-VL-32B [Bai 等人，2025b]	DeepInfra	32B	37.0
Qwen2.5-VL-7B [Bai 等人，2025b]	MI300X（vLLM）	7B	33.5
InternVL2.5-8B [Chen 等人， 2024]	MI300X（vLLM）	8B	28.1
Aya-Vision-8B [Dash 等人， 2025]	MI300X（vLLM）	8B	23.3

阿拉伯数字，并应用 ±5% 的容差。部分模型的提取失败率不可忽视（Aya-Vision-8B 为 33.9%，InternVL2.5-8B 为 25.0%）；这些模型报告的准确率应视为下限。

所有准确率均附有 95% 自助法置信区间（N=2，000 次重采样）。英语与各印度语言之间的成对比较采用 McNemar检验；所有比较均在$p < 0.001$水平上显著。

4 结果

4.1 总体准确率

表 4 展示了完整结果。GPT-4o 在印度语言上的平均准确率最高（43.2%）；而 Gemma 3-27B 尽管英语准确率较低，准确率却仅下降 9.8 个百分点，在所有受测模型中语言鲁棒性最佳。反观另一极端，Qwen3-VL-30B 的英语准确率最高（64.9%），但切换到印度语言后准确率下降了 24.9 个百分点，这表明英语表现出色并不意味着能够自动推广到跨语言场景。

在 7–8B 开源模型中，准确率下降幅度介于 20.4 个百分点（InternVL2.5-8B）至 21.9 个百分点（Aya-Vision-8B）之间。所有模型在各种语言上的拒答率均低于 3%，因此准确率损失反映的是真正的推理失败，而非模型拒答。

4.2 语系效应

表5 和图 2 分别按语系和语言细分了准确率下降情况。达罗毗荼语系语言（泰米尔语、泰卢固语、卡纳达语）的准确率下降幅度始终大于印度-雅利安语支语言（印地语、孟加拉语、马拉地语）。GPT-4o 的差距较小（1.7 个百分点），Gemma 3-27B 的差距几乎不存在（0.1 个百分点）；而 Aya-Vision-8B 的差距则达到 13.2 个百分点，尽管该模型的预训练明确涵盖了多种印度语言。针对印度语言训练的模型仍可能存在 13 个百分点的语系差距，这表明瓶颈在于推理迁移，而非表层的语言覆盖。

表4：各模型和语言的准确率（%）。英语列显示95%自助法置信区间；所有非英语比较均在 $p < 0.001$（McNemar检验）下具有显著性。“下降”是英语准确率与六种印度语言的平均准确率之差。

模型	en	hi	ta	te	bn	kn	mr	Drop
Qwen3-VL-30B	**64.9** [61.9–68.0]	44.1	41.1	36.5	43.3	38.7	36.1	24.9 个百分点
Llama-4-Maverick	55.4 [52.2–58.6]	43.4	34.9	38.6	42.7	41.6	41.9	14.9 个百分点
GPT-4o	54.5 [51.4–57.6]	**44.6**	**40.5**	**43.0**	**44.5**	**43.7**	**43.2**	**11.3** 个百分点
Gemma 3-27B	51.8 [48.9–55.0]	42.0	41.9	42.5	41.9	41.3	42.2	9.8 个百分点
Qwen2.5-VL-7B	51.3 [48.2–54.4]	34.2	28.7	26.9	32.7	30.6	30.0	20.8 个百分点
Qwen2.5-VL-32B	50.3 [47.1–53.5]	34.1	34.2	33.2	39.1	32.5	35.6	15.5 个百分点
InternVL2.5-8B	45.6 [42.5–48.9]	29.8	22.4	22.9	24.9	23.6	27.9	20.4 个百分点
Aya-Vision-8B	42.0 [38.8–45.0]	33.6	14.9	11.6	21.7	14.2	25.1	21.9 个百分点
消融实验（*Qwen2.5-VL-7B*）								
无图像	35.8	26.0	19.0	20.7	24.5	23.7	25.1	12.7 个百分点
思维链	49.5	25.0	29.7	22.8	18.3	19.2	33.3	24.8 个百分点

表 5：各语系内平均的准确率相较于英语的下降幅度（百分点）。“差距”指达罗毗荼语系的下降幅度减去印度-雅利安语支的下降幅度。

模型	印度-雅利安语支	达罗毗荼语系	Gap
Aya-Vision-8B	15.2	28.5	**+13.2**
InternVL2.5-8B	18.1	22.7	+4.6
Llama-4-Maverick	12.8	17.0	+4.3
Qwen2.5-VL-7B	19.1	22.6	+3.5
Qwen2.5-VL-32B	14.1	17.0	+3.0
Qwen3-VL-30B	23.7	26.1	+2.4
GPT-4o	10.4	12.1	+1.7
Gemma 3-27B	9.8	9.9	∼0

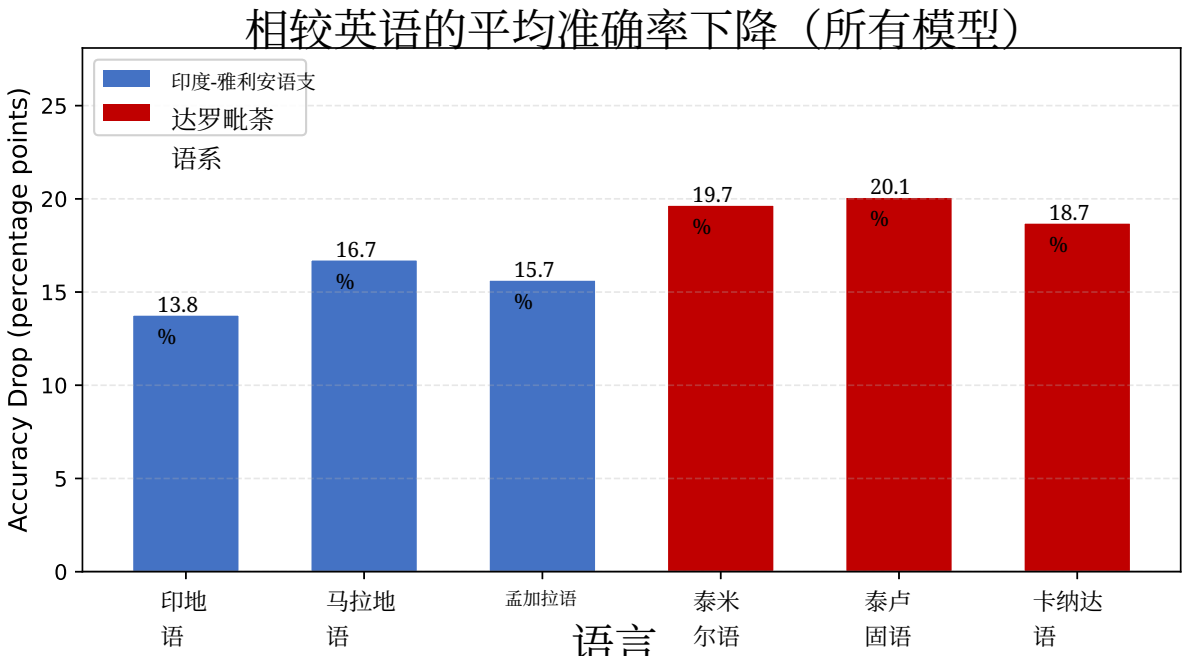


图 2：各语言相较英语的平均准确率下降，按模型取平均。达罗毗荼语系语言（ta、te、kn）的准确率下降始终大于印度-雅利安语支语言（hi、bn、mr）。

表6：按源数据集划分，从英语准确率到印度语言平均准确率的下降幅度（百分点）。

模型	MathVista	ScienceQA	MMMU
GPT-4o	17.2	7.0	7.3
Gemma 3-27B	24.3	−2.7	5.4
Llama-4-Maverick	15.2	15.7	12.3
Qwen3-VL-30B	29.3	21.5	22.8
Qwen2.5-VL-32B	10.2	23.2	10.3
Qwen2.5-VL-7B	20.5	24.8	12.9
InternVL2.5-8B	8.1	36.8	11.4
Aya-Vision-8B	9.3	38.4	13.1

文字系统与语言。 印地语和马拉地语都使用天城文，但在各模型中，马拉地语的准确率平均比印地语低 3.0 个百分点。即使是 Aya-Vision-8B，这一差距依然存在：尽管两种语言使用相同的文字系统，马拉地语的准确率仍比印地语低 8.5 个百分点。训练数据量似乎是主导因素，而非文字编码。

4.3 各数据集模式

表 6 和图 3 按源数据集细分准确率下降情况。两个模式尤为突出。首先，对于表现较好的模型，MathVista 导致的准确率下降最大（Qwen3-VL-30B 下降了 29.3 个百分点），这与数学文字题需要精准语言理解的情况相符。其次，Gemma 3-27B 在 ScienceQA 上的准确率下降为 负值 （−2.7 个百分点）：其印度语言准确率略高于英语。一种合理的解释是，ScienceQA 中的生物学和物理学图表包含了足够的视觉信息，使模型无论面对哪种问题语言都能得出正确答案；而 Gemma 的后训练对齐尤其善于利用这一点。

另一方面，Aya-Vision-8B 和 InternVL2.5-8B 在 ScienceQA 上的准确率下降了 36–38 个百分点，这表明它们严重依赖英语科学词汇。

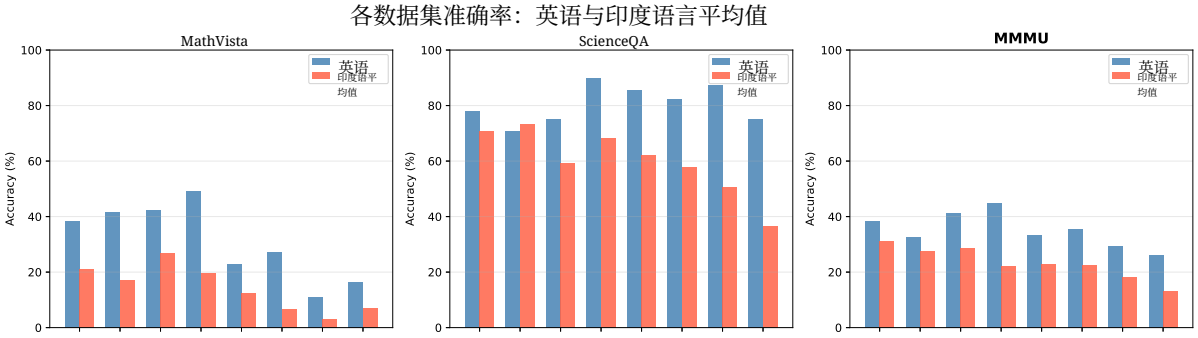


图3：英语与印度语言平均水平的各数据集准确率。对大多数模型而言，MathVista 的降幅最大。Gemma 3-27B 在 ScienceQA 上的降幅为负（印度语言的表现优于英语）。

图4通过同时展示三个数据集上各语言的准确率，进一步补充了上述结果。在 MathVista 和 MMMU 上，印地语和孟加拉语的表现始终优于达罗毗荼语系语言；而在 ScienceQA 上，这一差距有所缩小，因为图表层面的视觉信号在一定程度上弥补了语言理解能力较弱的不足。

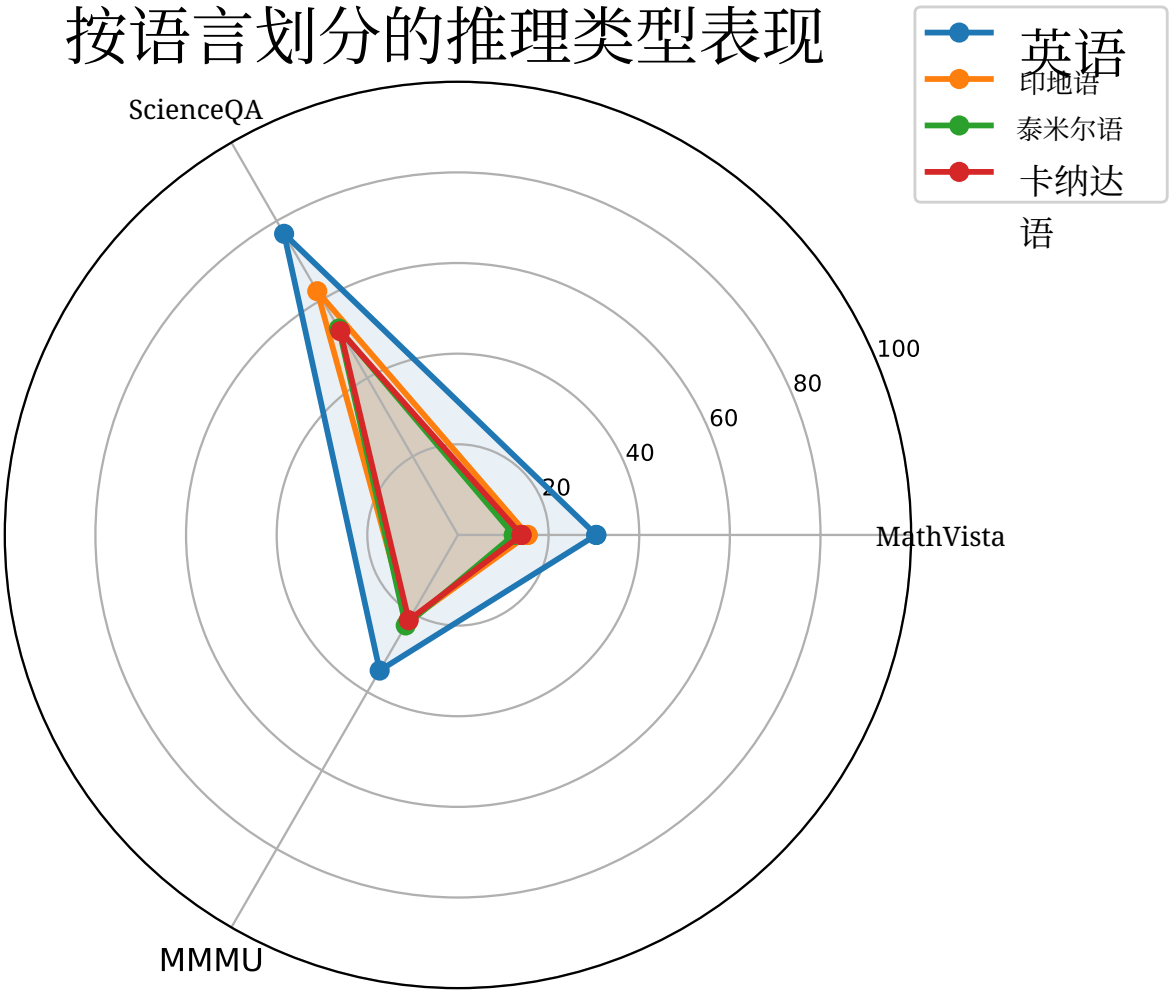


图4：各语言在每个源数据集上的准确率雷达图（对所有模型取平均）。英语作为外围参照；达罗毗荼语系语言（泰米尔语、泰卢固语、卡纳达语）在 MathVista 和 MMMU 上聚集在更靠近中心的位置，证实推理密集型任务受到的影响最大。

4.4 跨语言一致性

对于每个问题，我都会检查模型是否针对所有获得有效答案的语言给出了相同的提取出的答案（仅纳入七种语言中至少有四种成功提取答案的问题）。观察到的一致率介于 67.1%（Aya-Vision-8B）和 80.8%（Llama-4-Maverick）之间，远高于四选一多项选择题中独立随机猜测所预期的接近零的水平（图 5）。所有模型都进行了实质性的跨语言推理，但一致性最低的模型也正是准确率下降最大的模型，这表明语言依赖性不稳定与准确率损失可能有着共同原因。

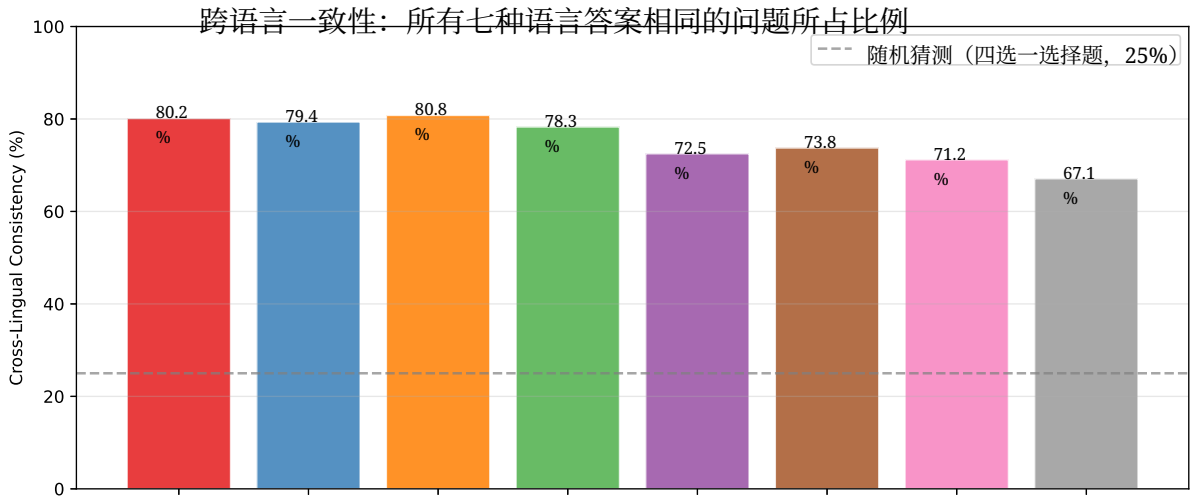


图5：跨语言一致性：在所有具有有效提取结果的语言中，获得相同提取答案的问题所占百分比（≥4 共7项）。

4.5 消融：移除图像与思维链

图6 展示了 Qwen2.5-VL-7B 三个变体在各语言上的准确率。

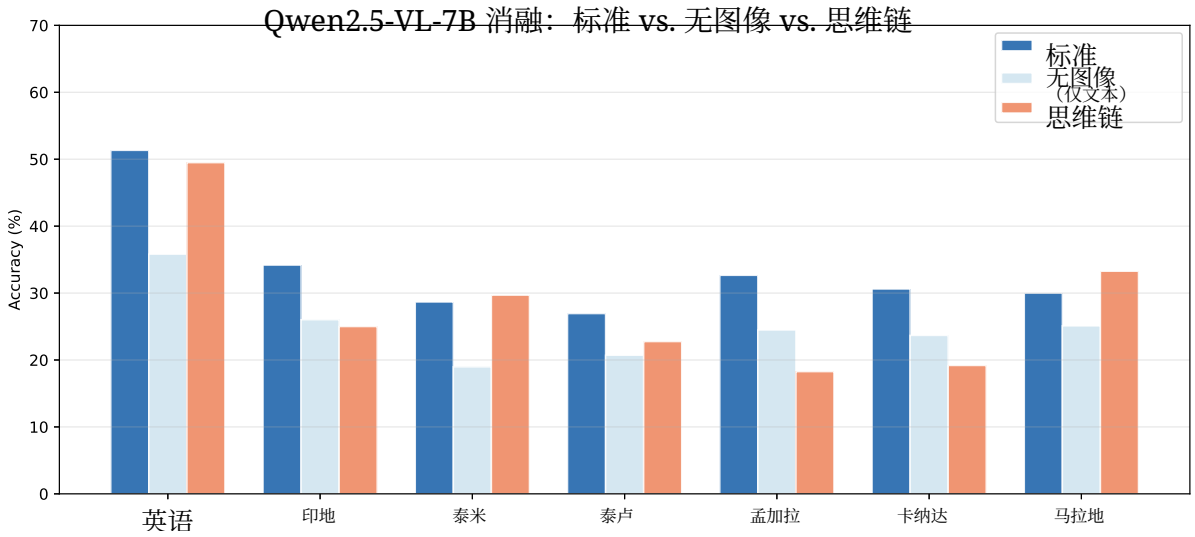


图6：Qwen2.5-VL-7B 的消融结果。移除图像会降低所有语言的表现；思维链会影响大多数印度语言的表现，而英语基本不受影响。

图像移除。 从提示中移除图像会使英语准确率下降 15.5 个百分点（从 51.3% 降至 35.8%）。印度语言准确率下降幅度较小，介于 4.9 个百分点（马拉地语）和 9.7 个百分点（泰米尔语）之间。这种不对称性颇具启示：模型在文本理解较好的情况下，能从图像中提取更多信息；当文本理解能力较弱（如使用印度语言时），图像本就未得到充分利用，因此移除图像造成的损失较小。

思维链。 添加“逐步思考”几乎不会改变英语准确率（−1.8 个百分点），却会显著降低孟加拉语（−14.4 个百分点）、卡纳达语（−11.4 个百分点）和印地语（−9.2 个百分点）的准确率。一个有趣的例外是马拉地语：思维链实际上提高了3.3 个百分点的准确率，这可能反映出用于推理任务的天城文训练数据较为丰富。总体而言，模型无法在不够熟练的语言中维持连贯的逐步推理；强行要求其这样做，会产生杂乱无章的中间步骤，进而损害最终答案。这与仅文本任务中的发现一致： Shi 等人 [2023]指出，英语思维链优于母语思维链；同时，这一发现也扩展到了视觉推理。

4.6 模型规模的影响

对比 Qwen2.5-VL-7B 和 Qwen2.5-VL-32B（属于同一模型系列）可以发现，较大的模型在印度语言上的表现提高了 4.3 个百分点，而在英语上的表现下降了 1.0 个百分点。英语到印度语言的准确率差距从 20.8 个百分点（7B）缩小到 15.5 个百分点（32B）：这是显著的缩减，但远未弥合差距。模型规模似乎扩大了语言覆盖范围，却没有从根本上改善跨语言推理迁移。

5 分析

图 7 展示了各模型在印度-雅利安语支和达罗毗荼语系语言上的性能降幅；所有数据点都位于对角线上方，证实达罗毗荼语系语言始终更难。

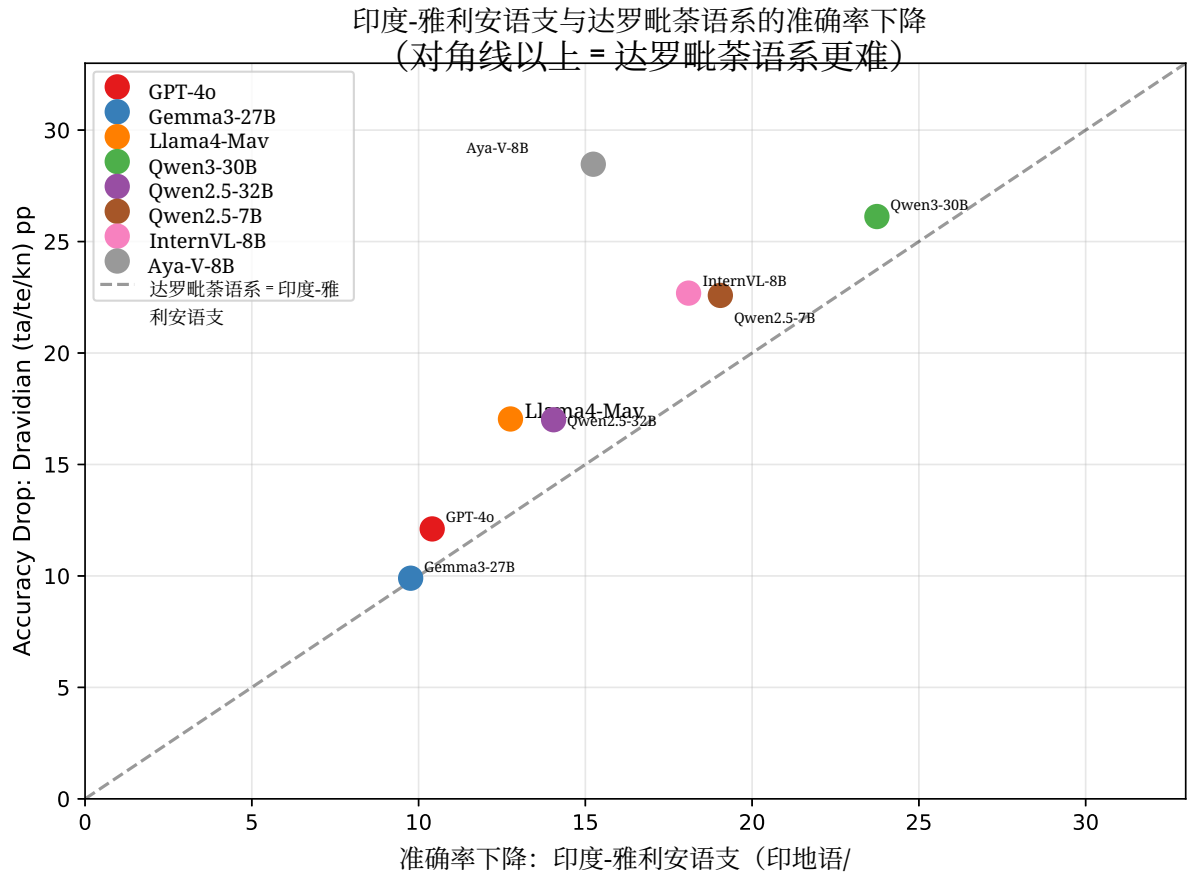


图 7：从英语到印度-雅利安语支和达罗毗荼语系语言的准确率下降（百分点）。位于对角线上方的点表示达罗毗荼语系语言更难。Gemma 3-27B 是唯一接近对角线的模型。

5.1 英语词元泄漏

为探查是否存在隐蔽的英语推理，我测量了非英语回答中含 ≥3 个字母的英语单词所占比例（图 8）。在印度各语言的回答中，Llama-4-Maverick 的英语词元平均占比高达 32%，尤为突出。显然，它使用英语进行推理，只是表面上切换为目标文字，但其跨语言一致性最高（80.8%），这意味着其英语泄漏推理往往正确。

与此相反，Aya-Vision-8B 生成的英语词元仅占 0.8%，但一致性最低（67.1%）。忠实地使用目标文字生成内容，并不意味着推理正确。

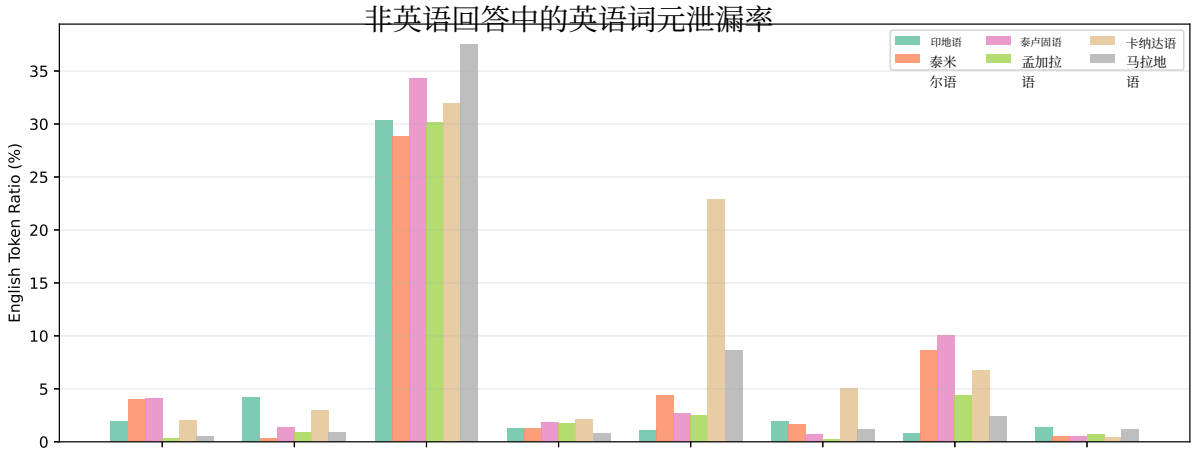


图 8：非英语回答中的英语词元泄漏（ ≥3 字母英语单词的百分比）。Llama-4- Maverick 的语码转换现象严重，占 32%。

5.2 语言混淆

我将超过 15 个字符、且不含预期 Unicode 文字范围内任何字符的回答标记为混淆（图 9）。Qwen3-VL-30B（混淆率 1.2%）和 Aya-Vision-8B（2.5%）始终使用目标文字作答。GPT-4o 的表面混淆率为 45%，但这很大程度上是其简短的单字母选择题答案（例如“B”）造成的假象，因为这类答案本身就不含任何特定文字的字符。

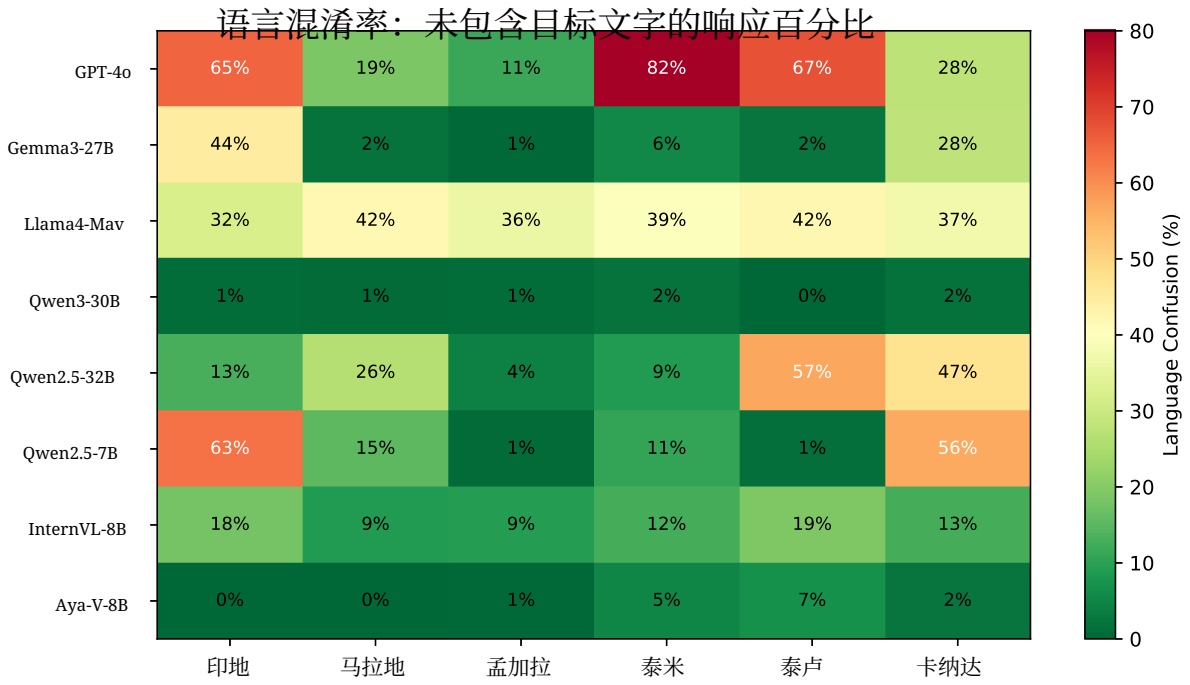


图 9：语言混淆率：冗长（>15字符）响应中不含预期 Unicode 文字范围内任何字符的百分比。

5.3 响应长度模式

回答长度揭示了不同模型的策略差异（图10）。无论使用哪种语言，Llama-4-Maverick 生成的输出都始终冗长（~840–1,070 个字符）；它总是会写出完整的推理过程。Aya-Vision-8B 使用印度语言作答时，回答更长（310–514 个字符），而英语回答为 123 个字符；这可能反映出它信心较低，并通过生成更多内容来弥补。InternVL2.5-8B 则恰好相反：英语回答只有 6 个字符（单字母选择题答案），而印度语言回答则是完整句子，这在一定程度上解释了它较高的混淆率。

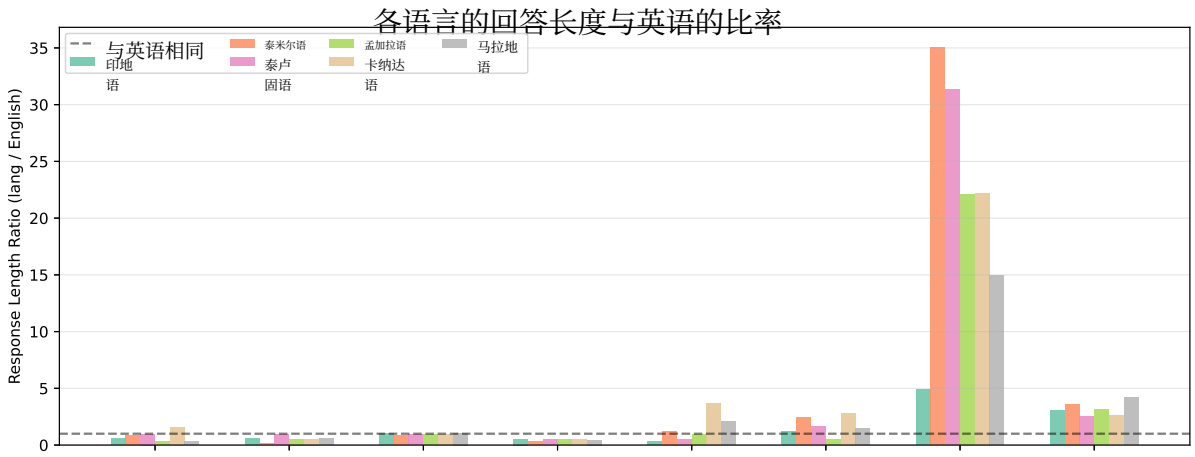


图 10：相对于英语的回答长度比。数值大于 1 表示目标语言中的回答更长。

5.4 选择题与自由作答题

自由作答数值题受到的影响尤为严重。Qwen2.5-VL-32B 在英语选择题上的准确率为 63.9%，但在自由作答数值题上仅为 2.8%：它会生成冗长的推理，却无法给出简洁明确的最终数值，导致答案提取失败。选择题与自由作答题之间相差 61 个百分点，是这组数据中最大的差距，也凸显出报告的准确率在很大程度上受答案提取质量影响。InternVL2.5-8B 也呈现出类似趋势（选择题为 52.5%，自由作答题为 21.6%）。

6 讨论

教育方面的影响。印度《国家教育政策》 2020 [印度政府人力资源发展部， 2020] 规定五年级之前必须以母语授课。如果目前表现最佳的视觉语言模型在用印度语言提问时，STEM推理能力下降11.3个百分点，而开源替代方案下降15至25个百分点，那么在地方语言授课学校中将这些模型用作辅导工具，可能会系统性地使非英语学生处于不利地位。达罗毗荼语授课学校面临的问题最为严峻，具体取决于模型，表现降幅可达12至28个百分点。

多语言预训练还不够。Aya-Vision-8B接受了23种语言的训练，其中有几种是印度语言。它能够准确使用目标文字作答（英语泄漏率为0.8%），也很少拒答。然而，它在达罗毗荼语系语言上的表现下降了28.5个百分点，跨语言一致性也最低。该模型可以生成泰米尔语文本，却无法用泰米尔语进行推理。要弥合这一差距，可能需要使用 STEM 专项数据进行多语言推理微调，而非仅仅扩大预训练语料库。

作为异类的 Gemma 3-27B。Gemma 3-27B 是唯一一个达罗毗荼语系与印度-雅利安语系差距接近于零（0.1 个百分点）的模型，而且在 ScienceQA 上从英语切换到其他语言时，准确率实际上有所提升。虽然完整的训练方案尚未公开，但这表明，在后训练期间进行有针对性的多语言对齐，可以有效拉平不同语系之间的表现差异，这是未来模型的一项具体设计目标。

思维链：谨慎使用。思维链提示通常被认为有助于推理任务。结果表明，在印度语言中，思维链提示会适得其反，马拉地语除外。当模型缺乏足够的语言流畅度，无法用目标语言维持连贯

的逻辑时，推理链就会成为瓶颈。在多语言环境中部署视觉语言模型的从业者应避免使用思维链提示，除非模型经过专门调优，能够生成非英语推理链。

隐藏的英语模式。Llama-4-Maverick 的英语词元泄漏率为 32%，跨语言一致性为 80.8%，两者结合表明，该模型可能采用这样的策略：在内部用英语推理，再在表层进行翻译。这种方式能得出正确的选择题答案，但在辅导、解释和反馈等生成任务中可能会失效，因为这些任务要求用目标语言流畅作答。

7 局限性

这些翻译采用 IndicTrans2 完成，并通过 Gemini 2.0 Flash 进行交叉验证（一致率为 0.79–0.84）。机器翻译错误，尤其是技术术语方面的错误，可能会引入噪声；若能扩大样本并进行母语者验证，将有助于提升基准测试的可靠性。印度 22 种列定语言中仅涵盖 6 种；奥迪亚语、阿萨姆语等资源匮乏的语言，以及 100 多种非列定语言，仍未经过测试。源基准测试面向英语使用者设计，可能存在文化或课程偏见，并对所有译文产生同等影响。Aya-Vision-8B 和 InternVL2.5-8B 的提取失败率较高，分别为 33.9% 和 25.0%，这意味着这些模型的实际推理能力可能高于报告中的数值。基于 API 的模型可能会随时间更新；这些结果反映的是 2026年3月 的一次快照。

8 结论

本次审计涵盖 980 道题、七种语言、八个模型和 68,600 条推理记录，记录了一个系统性问题：视觉语言模型使用印度语言而非英语进行推理时，准确率会损失 9.8–25 个百分点。达罗毗荼语系语言始终比印度-雅利安语支语言更难处理。思维链提示在大多数印度语言中适得其反；仅靠多语言预训练无法迁移推理能力；扩大模型规模虽能缩小差距，却无法消除差距。这些发现表明，在视觉语言模型部署于印度教育环境之前，必须进行多语言推理评估。

翻译后的基准、模型输出和评估代码可在 https://github.com/QuantumByte-01/multilingual-vlm-reasoning-audit 和 HuggingFace 上公开获取。[2]

致谢

开源模型推理在配备 192GB HBM3 内存的 AMD Instinct MI300X GPU 上进行。云端推理使用了 Google AI、DeepInfra 和 OpenAI API。

参考文献

Kabir Ahuja、Harshita Diddee、Rishav Hada、Millicent Ochieng、Krithika Ramesh、Prachi Jain、Akshay Nambi、Tanuja Ganu、Sameer Segal、Maxamed Axmed、Kalika Bali 和 Sunayana Sitaram。《MEGA：生成式 AI 的多语言评估》。载于 *EMNLP* 论文集，2023年。

Jesse Atuhurra 等人。《VLURes：低资源语言中视觉语言模型视觉与语言理解能力的基准测试》。*arXiv* 预印本 *arXiv:2510.12845*，2025年。

Shuai Bai、Yuxuan Cai、Ruizhe Chen 等人。Qwen3-VL 技术报告。*arXiv* 预印本 *arXiv:2511.21631*，2025a。

[2] https://huggingface.co/datasets/Swastikr/multilingual-vlm-reasoning

Shuai Bai 等人。Qwen2.5-VL 技术报告。arXiv 预印本 arXiv:2502.13923，2025b。Zhe Chen、Jiannan Wu、Wenhai Wang、Weijie Su、Guo Chen、Sen Xing、Muyan Zhong、Qinglong Zhang、Xizhou Zhu、Lewei Lu、Bin Li、Ping Luo、Tong Lu、Yu Qiao 和 Jifeng Dai。InternVL：扩展视觉基础模型，并使其适用于通用视觉语言任务。在 CVPR 论文集中，2024年。Saurabh Dash、Yiyang Nan、John Dang、Arash Ahmadian 等人。Aya vision：推动多语言多模态领域的前沿。arXiv 预印本 arXiv:2505.08751，2025年。Sumanth Doddapaneni 等人。致力于不让任何印度语言掉队：为印度语言构建单语语料库、基准和模型。在 ACL 论文集中，2023年。Ali Faraz 等人。IndicVisionBench：对视觉语言模型中的文化与多语言理解进行基准测试。arXiv 预印本 arXiv:2511.04727，2025年。Jay Gala、Pranjal A. Chitale、Raghavan, A K、Varun Gumma、Sumanth Doddapaneni、Aswanth Kumar、Janki Nawale、Anupama Sujatha、Ratish Puduppully、Vivek Raghavan、Pratyush Kumar、Mitesh M. Khapra、Raj Dabre 和 Anoop Kunchukuttan。IndicTrans2：为全部22种印度宪定语言提供高质量、易获取的机器翻译模型。《机器学习研究汇刊》，2023年。Gemma Team、Google DeepMind。Gemma 3 技术报告。arXiv 预印本 arXiv:2503.19786，2025年。Shaharukh Khan 等人。Chitrarth：为十亿人搭建视觉与语言之间的桥梁。arXiv 预印本 arXiv:2502.15392，2025年。Fangyu Liu、Emanuele Bugliarello、Edoardo Maria Ponti、Siva Reddy、Nigel Collier 和 Desmond Elliott。跨语言与文化的视觉接地推理。在 EMNLP 论文集中，2021年。Pan Lu、Swaroop Mishra、Tony Xia、Liang Qiu、Kai-Wei Chang、Song-Chun Zhu、Oyvind Tafjord、Peter Clark 和 Ashwin Kalyan。学会解释：通过思维链进行多模态推理，以回答科学问题。载于《神经信息处理系统进展》，2022年。Pan Lu、Hritik Bansal、Tony Xia、Jiacheng Liu、Chunyuan Li、Hannaneh Hajishirzi、Hao Cheng、Kai-Wei Chang、Michel Galley 和 Jianfeng Gao。MathVista：评估基础模型在视觉情境中的数学推理能力。在 ICLR 论文集中，2024年。Arijit Maji 等人。DRISHTIKON：用于测试语言模型对印度文化理解能力的多模态多语言基准。在 EMNLP 的 MAINS研讨会论文集中，2025年。Meta AI。Llama 4 系列：原生多模态人工智能创新新时代的开端。https://ai.meta.com/blog/llama-4-multimodal-intelligence/，2025年。印度人力资源发展部、印度政府。《2020年国家教育政策》。https://www.education.gov.in/sites/upload_files/mhrd/files/NEP_Final_English_0.pdf，2020年。OpenAI。GPT-4o 系统卡。技术报告，2024年。Jonas Pfeiffer、Gregor Geigle、Aishwarya Kamath、Jan-Martin Steitz、Stefan Roth、Ivan Vuli´c 和 Iryna Gurevych。xGQA：跨语言视觉问答。载于 ACL论文集，2022年。David Romero 等人。CVQA：文化多样性多语言视觉问答基准。载于《神经信息处理系统进展》，2024年。Alan Saji、Raj Dabre、Anoop Kunchukuttan 和 Ratish Puduppully。推理通用语：多语言人工智能的一把双刃剑。arXiv 预印本 arXiv:2510.20647，2025年。

Freda Shi、Mirac Suzgun、Markus Freitag、Xuezhi Wang、Suraj Srivats、Soroush Vosoughi、Hyung Won Chung、Yi Tay、Sebastian Ruder、Denny Zhou、Dipanjan Das 和 Jason Wei。语言模型是多语言思维链推理器。载于 *ICLR* 论文集，2023年。Harshul Raj Surana 等人。VIRAASAT：探索印度文化推理的新路径。*arXiv* 预印本 *arXiv:2602.18429*，2026年。Ashmal Vayani 等人。所有语言都很重要：在文化多样的100种语言上评估大型多模态模型。*arXiv* 预印本 *arXiv:2411.16508*，2024年。Ke Wang 等人。使用 MATH-Vision数据集衡量多模态数学推理。*arXiv* 预印本 *arXiv:2402.14804*，2024年。Xiang Yue、Yuansheng Ni、Kai Zhang、Tianyu Zheng、Ruoqi Liu、Ge Zhang、Samuel Stevens、Dongfu Jiang、Weiming Ren、Yuxuan Sun 等人。MMMU：面向专家级通用人工智能的大规模多学科多模态理解与推理基准。载于 *CVPR* 论文集，2024年a。Xiang Yue 等人。Pangea：面向39种语言的完全开放多语言多模态大语言模型。*arXiv* 预印本 *arXiv:2410.16153*，2024年b。