MedAraBench：大规模阿拉伯语医学问答数据集与基准测试

Mouath Abu-Daoud[1]，**Leen Kharouf**[1]，**Omar El Hajj**[1]，**Dana El Samad**[1]，**Mariam Al-Omari**[1]，**Jihad Mallat**[3]，**Khaled Saleh**[3]，**Nizar Habash**[2]，**Farah E. Shamout**[1]

[1]纽约大学阿布扎比分校工程学部，阿联酋 [2]纽约大学阿布扎比分校科学学部，
阿联酋 [3]克利夫兰诊所阿布扎比，阿联酋
{mma9138，lk2713，ose6432，dae7168，ma8058，nh48，fs999}@nyu.edu
{mallatj，salehk}@ccad.ae

arXiv:2602.01714v1 [cs.CL] 2 Feb 2026

摘要

阿拉伯语仍是自然语言处理研究中代表性不足的语言之一，在医学应用领域尤为如此，这主要是由于开源数据和基准测试资源有限。资源匮乏阻碍了对大语言模型（LLM）多语言能力的评估与提升。本文介绍 MedAraBench，这是一个大规模数据集，包含涵盖多个医学专科的阿拉伯语选择题问答对。我们通过人工数字化由阿拉伯语地区医疗专业人士编写的大型学术资料库，构建了该数据集。随后，我们进行了广泛的预处理，并将数据集划分为训练集和测试集，以支持该领域未来的研究。为评估数据质量，我们采用了两种框架，即专家人工评估和大语言模型评审。我们的数据集内容多样、质量上乘，涵盖 19 个专业领域和五个难度等级。为进行基准测试，我们评估了八种最先进的开源及专有模型的性能，例如 GPT-5、Gemini 2.0 Flash 和 Claude 4-Sonnet。研究结果表明，仍需进一步进行领域特定增强。我们发布该数据集及评估脚本，旨在扩大医学数据基准测试的多样性，拓展大语言模型评估套件的范围，并增强模型的多语言能力，以便将其部署于临床场景。

1 引言

近年来，大语言模型（LLM）的出现推动了自然语言处理领域的变革性进展。凭借强大的理解和推理能力，以及训练语料库中海量知识的支撑，它们在各种知名基准测试中展现出卓越性能（Brown 等，2020年；Bommasani 等，2022年；Chowdhery 等，2022年）。这包括通用基准测试和领域特定基准（Wang 等，2021；Stahlberg，2020年）。

然而，性能提升在代表性不足的语言和领域之间仍存在差异，尤其是在医学等高风险应用中（Jiang et al., 2025年；Yang et al., 2025年）。例如，阿拉伯语是世界上使用人数最多的语言之一，全球使用者超过4亿。然而，由于专家标注资源有限，阿拉伯语在医学领域的代表性仍然不足（Habash, 2005）。阿拉伯语自然语言处理通常面临固有的语言特有挑战（Habash, 2005；Farghali & Shaalan, 2009；AlMoaiad, 2024），因此这类资源对于

表1：MedAraBench 与现有医学问答基准的比较，包括估计的数据集规模。

基准测试	语言	Type	Size	专家标注	难度映射	专业领域覆盖	阿拉伯语	公开
MedQA	英语、中文	MCQs	60,000	✓	×	✓	×	✓
MedMCQA	英语	MCQs	193,000	✓	×	✓	×	✓
MMLU（USMLE）	英语	MCQs	1,800	×	×	✓	×	✓
MMLU翻译	14种（含阿拉伯语）	MCQs	15,000	✓	×	✓	✓	✓
AraMed	阿拉伯语	QA	270,000	✓	×	✓	✓	×
MedArabiQ	阿拉伯语	阿拉伯语问答和选择题	700	×	×	✓	✓	✓
MedAraBench（我们的工作）	阿拉伯语	MCQs	24,000	✓	✓	✓	✓	✓

asse评估大语言模型的性能尤为重要，尤其是在它们正被部署于各种医学 l
con文本。

近期，医学领域陆续推出了多个基准测试，如表1所示。然而，其中大多数几乎只关注英语（Jin 等人，2021；2019）。近期研究开始着手满足这一需求，但范围和规模仍然有限（Abu Daoud等，2025年）。因此，亟需大规模基准测试来评估并提升大语言模型在阿拉伯语医学推理方面的能力。

为弥补这些不足，本文提出 MedAraBench，这是一个全面的基准测试，用于评估和提升大语言模型在阿拉伯语医学任务上的表现。该数据集由经过筛选的选择题组成，涵盖不同专业领域和难度等级，并与医学教育的各个阶段相对应。我们提出了一套标准化的开发与评估规程，以实现可复现且具有临床意义的大语言模型评估。我们的主要贡献如下：

- 我们推出了 MedAraBench，这是一个大规模阿拉伯语医学基准，包含 19 个医学专科、五个难度等级的 24,883 道选择题。该基准提供标准化的训练集和测试集，以支持对大语言模型进行系统评估并推动其发展。
- 我们通过人工专家评估开展了全面的质量评估，重点关注问题清晰度、临床相关性和医学准确性，并进行了自动化的大语言模型评审分析。
- 我们在零样本设置下，使用 MedAraBench 测试集对 16 个最先进的专有及开源大语言模型进行基准测试，涵盖通用、以阿拉伯语为中心和医学三类模型，以确立未来研究的基线性能。

2 相关工作

近年来，随着简化临床工作流程的应用需求不断增长，人们开发了多个基准数据集，用于评估大语言模型在医学领域的能力。尽管取得了这些进展，阿拉伯语在临床自然语言处理中仍代表性不足，主要原因是缺乏高质量数据来支持阿拉伯语临床应用的构建（Abdelaziz 等人，2025年）。因此，现有基准大多聚焦于英语。例如，大规模多任务语言理解（MMLU）基准包含来自美国医学执照考试（USMLE）的问答对（Hendrycks 等人，2021）。Jin 等人（2020年）推出了 MedQA，这是一个多语言基准数据集，包含从英语和中文医学执照考试中收集的选择题。Pal 等人（2022年）通过 MedMCQA 将这些基准扩展为多语言评估框架，但其阿拉伯语覆盖仍然有限。

近期研究引入了用于阿拉伯语医学评估的新资源。现有数据集的译本（例如将 MMLU 翻译成包括阿拉伯语在内的 42 种语言）提供了有价值的数据，但缺乏适当融入临床实践所需的细微差异（Singh，2025年）。AraMed 提供了一个阿拉伯语医学语料库，以及一个来源于在线医疗平台的带标注的阿拉伯语问答数据集（Alasmari 等人，2024）。MedArabiQ 提供了首批阿拉伯语医学选择题数据集之一，但缺乏专业领域覆盖、难度映射和专家评估（Abu Daoud等，2025年）。

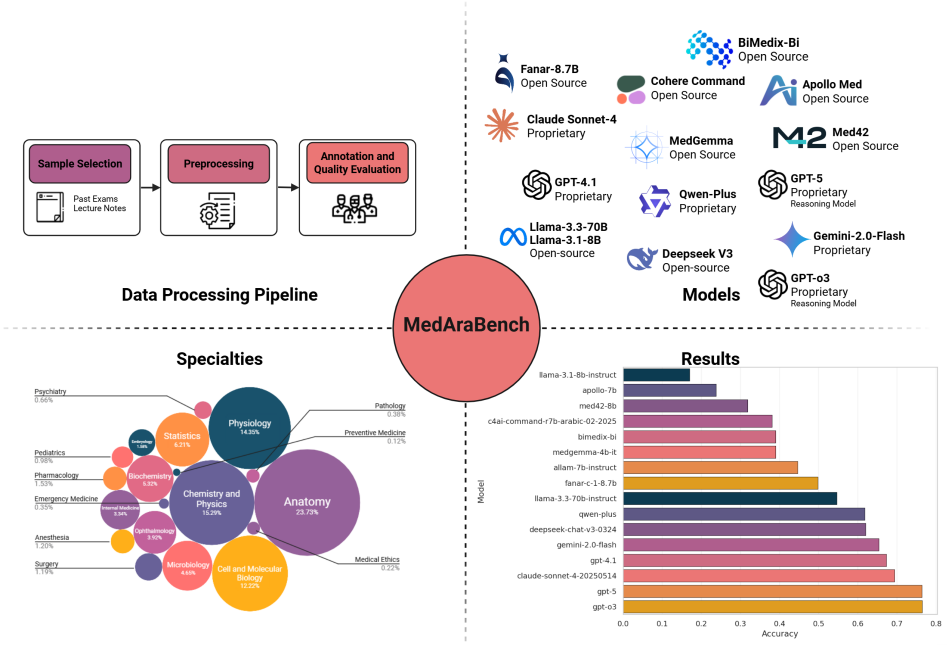


图 1：MedAraBench 概览。

首批阿拉伯语医学选择题数据集之一，但缺乏专业领域覆盖、难度映射和专家评估（Abu Daoud等，2025年）。

目前已有多种评估框架用于评估临床人工智能模型的性能。Kanithi 等人（2024）提出了 MEDIC，这是一个评估大语言模型从医学推理和伦理到上下文学习与临床安全能力的框架。Wang 等人（2024）建议在真实世界输入噪声、对话中断和推理依据方面测试模型。尽管人们对多语言评估的兴趣日益浓厚，但针对服务不足的语言，全面、高质量且具有临床相关性的基准仍存在关键缺口。我们旨在通过推出一项全面的阿拉伯语基准来弥补这一缺口。表1对 MedAraBench 与其他现有医学基准进行了结构化比较，涵盖语言覆盖、专科多样性、专家标注和公开可用性等关键维度。

3 方法

在此，我们介绍数据收集、处理、评估、基准测试、少样本学习和微调的相关步骤，以便促进规范的重复使用和公平比较，并遵循基准构建与评估的最佳实践。图 1 展示了总体概览。

3.1 数据收集和预处理

我们汇总了大量扫描版纸质考试资料，这些资料托管在地区医学院学生主导的社交平台上。该数据集不包含任何个人信息或真实患者数据，因此无需进行匿名化处理；我们的数据收集也符合隐私和伦理标准。考虑到文档的性质，我们聘请了专业打字员将数据数字化。随后，我们汇总这些纸质考试资料，构建了一个多项选择题数据集。

经自然语言处理研究人员人工检查，我们观察到若干文档存在如下问题：正确答案缺失或格式错误、答案选项不完整或重复、格式不规范或字段错位，以及答案键含糊或额外-

的非选择题内容。为确保数据集质量与模型兼容性，我们采用了严格的筛选标准，以剔除存在此类问题的任何题目。筛选过程由五名自然语言处理研究人员人工完成。尽管数据获取与预处理需要大量投入，但这表明该数据集无法以结构化格式公开获取，从而降低了数据污染的可能性。

每个问题都附有若干标注：(i) 答案选项数量（即 ABCD（4 个选项）、ABCDE（5 个选项）和 ABCDEF（6 个选项）），(ii) 对应五年学习阶段（Y1 - Y5）的难度级别，以及 (iii) 医学专科。问题涉及 19 个医学专科：解剖学、麻醉学、生物化学、细胞与分子生物学、化学与物理学、胚胎学、急诊医学、内科学、医学伦理学、微生物学、眼科学、病理学、儿科学、药理学、生理学、预防医学、精神病学、统计学和外科学。从各存储库收集的扫描文档最初按专业领域分类，因此我们可以直接沿用其对每个问题的医学专科分类。

由于包含 6 个答案选项的问题数量较少（9 道），我们首先将其全部剔除，从而将数据分为两类（4 个答案选项和 5 个答案选项）。随后，我们对数据集进行分层随机划分，分为训练集（80%）和测试集（20%）。这样做是为了确保各医学专科在训练集和测试集中均匀分布（例如，如果数据集中有 100 道心脏病学问题，那么随机选取 80 道纳入训练集，其余 20 道纳入测试集）。附录 A 总结了数据集的 token 长度分布、医学专科分布和难度级别分布。

需要注意的是，在评估大语言模型处理标准化阿拉伯语医学任务的能力时，医学术语的一致性至关重要。对于 MedAraBench，我们未对整个数据集进行外部标准化，因为我们认为，从基准测试的角度来看，这可能并非重大问题：现实世界中的医学问答在术语方面未必需要标准化。

我们承认，阿拉伯语医学术语标准化仍是一项持续面临的挑战，阿拉伯语地区的统一词汇框架仍在制定中（Amar，2022年；Attia，2024）。第 3.2 节所述的临床专家评估起到了质量控制的作用；尽管缺乏标准化，评审者在所有指标上的意见一致度仍然很高。

3.2 质量评估

为进一步评估数据集的质量，我们进行了两项分析：人类专家评估和大语言模型评审。

3.2.1 人类专家评估

We 我们设计了专家评估方案，按照以下标准评估数据
ng crite

1. **医学准确性：** 问题、选项及正确答案反映当前循证医学知识的程度（等级：高或低）（Olatunji 等，2024；Iskander 等，2024；Rejeleene 等，2024）。2. **临床相关性：** 问题内容对真实世界医学实践或教育的实际重要性与适用性（等级：高或低）（Iskander 等，2024；Olatunji 等，2024）。3. **问题难度：** 正确回答该问题所需的复杂程度（等级：高或低）（Iskander 等，2024）。4. **问题质量：** 依据既定医学教育标准对选择题构建质量的评估（等级：高或低）（Al-Rukban，2006）：

- 清晰度： 问题清晰、完整且明确。

- 选项同质性： 所有干扰项都合理且类型相似。
- 单一最佳答案 存在一个明确正确的选项。
- 无提示性： 选项不会为其他答案提供线索。

我们从测试集中选取样本进行审查，并依据 Cochran 公式（Cochran，1977年）确定样本量，以构建具有代表性的样本（Hosseini，2024年）。我们以 95% 的置信度估计单一比例，并采用 ± 5 个百分点的误差范围；当真实质量率未知时，使用 $p = 0.5$ 作为保守假设，因为这会使方差最大化，从而得出安全的样本量上限。我们首先假设总体无限，计算估计样本量，公式如下

$$n_0 = \frac{z^2 p(1-p)}{e^2}$$

其中 $z = 1.96$、 $p = 0.5$ 和 $e = 0.05$。然而，由于数据集规模有限，我们随后使用以下公式应用有限总体校正

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

我们招募了两名经委员会认证的临床医生，分别专攻麻醉学和内科学，具备阿拉伯语临床交流能力，且每人拥有超过 20 年的经验。评审采用双盲方式，不向评审者展示模型输出和数据来源；两位评审者依据预先注册的说明，通过$Qualtrics$ （犹他州普罗沃，https://www.qualtrics.com）独立开展评审。Qualtrics 是一个基于 Web 的调查平台，可标准化呈现评估标准，并分别收集两位评审者的回答。

3.2.2 大语言模型评审

鉴于人类专家评估仅覆盖了测试集的一个子集，我们进一步引入了大语言模型评审评估协议，作为数据质量的额外评分。受先前文献（Abu Daoud等，2025年）的启发，我们提示了四个表现最佳的最先进的大语言模型（gpt-03、gemini-2.0-flash和claude-4-sonnet）充当医学教育专家，并按照我们专家质量评估中使用的相同指标，对整个测试集的选择题进行评估：医学准确性、临床相关性、问题难度和问题质量，采用二元（0或1）量表。此外，我们针对医学专家评估的378题数据集，计算了每个模型与专家评审员之间的皮尔逊相关系数。该方法的优势在于，能够在更广泛的数据范围内提供更细致的评估，同时也有助于洞察大语言模型相对于专家标注员在评估阿拉伯语医学数据质量方面的效能。

3.3 基准测试协议

为引入新的基准测试结果，我们在 MedAraBench 基准测试的测试集上评估了 16 个专有模型和开源模型。我们将模型温度设为 0，以确保输出稳定，如先前针对选择题分类的工作所示（Abu Daoud等，2025年）。我们为每个问题选取一个字母作为回答。

- **开源模型：**Llama-3.3-70b-instruct（Grattafiori等，2024a年）、Llama- 3.1-8b-instruct（Grattafiori等，2024b年）、Deepseek-chat-v3-0324（DeepSeek-AI，2024年）、Allam-7b-instruct（Bari，2024年）、Cohere c4ai-comman-r7b-arabic-02-2025（Alnumay，2025年）、Medgemma-4b-it（Sellergren，2025年）、Apollo-7b（Wang，2024年）、Fanar-C-1-8.7b（Abbas，2025年）、BiMedix-Bi-27B（Pieri等，2024年）以及 Med42-8b（Christophe，2024年）。

- **专有模型：**Claude-sonnet-4-20250514（Anthropic，2025年）、Gemini-2.0- flash（Google Cloud，2025年）、GPT-4.1（OpenAI，2025a）、GPT-5（OpenAI，2025b）、GPT-o3（OpenAI，2025c）以及qwen-plus（Yang等人，2024）。

我们的提示词如下文附录B所示。答案通过后处理提取，同时提示模型直接输出答案选项字母（A-D），我们则使用模式匹配解析其文本回复以提取答案。由于模型会自动从输入中检测阿拉伯语文本，因此在API和模型调用中未使用设定的语言参数。我们的基准测试实验结果如下表4所示。

3.4 少样本学习方案

在零样本基准测试方案的基础上，我们针对 LLaMa-3.1-8B-instruct 开展了少样本实验，以评估其上下文学习能力。我们提供了 3 个高质量示例问题；专家评估者认为这些问题在所有评估指标（问题质量、临床相关性、难度和医学准确性）方面均表现出色。这些示例问题经过精心挑选，来自训练集划分，并从测试集中剔除，以确保评估公平。少样本示例涵盖解剖学、生物化学和生理学等多个医学主题，采用阿拉伯语呈现问题、选项和正确答案，以确保与测试题目在语言上的一致性。所选问题和少样本提示见下文附录 B。

3.5 低秩自适应

我们进一步研究了在 Llama-3.1-8B-instruct 模型上使用 QLoRA（量化低秩自适应）进行参数高效微调。模型以 4位精度加载，并在 MedAraBench 训练集划分上进行训练，数据格式为用于多项选择题问答的阿拉伯语提示—回答对。我们使用标准超参数，将 LoRA 适配器应用于关键注意力模块（q proj、k proj、v proj、o proj），并通过批量梯度累积训练最多 800 步。这种方法在保留模型通用能力的同时，使其能够高效适应阿拉伯语医学术语和推理模式。我们使用与零样本和少样本实验相同的测试集评估微调后的模型，以直接衡量 MedAraBench 数据对模型性能的影响。

4 结果

In t本节概述我们的数据集，以及专家质量评估的结果 y
eval、大语言模型评审实验和基准测试实验。

4.1 数据集概述

初始数据集包含 34,333 道选择题。经过人工筛选，初始数据集缩减了约 29%，最终共计 24,883 个样本。训练集包含 19,894 个样本，测试集包含 4,989 个样本。下文附录 A 中的图 A1 展示了数据集概况，并附有其他统计摘要。

4.2 专家质量评估

在 95% 置信水平和 $\pm 5\%$ 误差范围下，Cochran 公式最初得出 $n_0 = 384$ 个问题。经过有限样本调整后，最终样本量为 378。因此，我们的两位标注员完成了对 378 个问题的审查，这些问题作为整个数据集的代表性样本。数据质量评估和标注员间一致性的结果汇总于表 2。结果显示，所有指标的一致性均处于轻微至一般水平，其中医学准确性的一致性最高，科恩卡帕系数为 0.555，一致率为 82%。

此外，我们在附录 C 中提供了按专业领域划分的详细评估结果。为了更好地评估这些结果，我们考察了每个专业领域各自的平均值和一致性得分。具体而言，图 C1、C2、C3 和 C4 展示了各指标下每个专业领域的平均准确率，而表 C1、C2、C3 和 C4 则展示了平均准确率

表2：代表性数据子集的专家质量评估结果。

指标	平均值 [标准差]	百分比一致性	科恩卡帕系数
医学准确性	0.722 [0.448]	82.0%	0.555
临床相关性	0.653 [0.476]	65.6%	0.275
问题难度	0.669 [0.471]	65.6%	0.233
问题质量	0.767 [0.423]	68.3%	0.152

表3：大语言模型评审应用于测试集的评估结果。

模型	评估指标（平均值）			
	医学准确性	临床相关性	问题难度	问题质量
GPT-o3	0.673 [0.469]	0.827 [0.378]	0.588 [0.492]	0.841 [0.366]
Gemini 2.0 Flash	0.717 [0.450]	0.565 [0.496]	0.815 [0.388]	0.774 [0.366]
Claude-4-Sonnet	0.711 [0.453]	0.749 [0.434]	0.576 [0.494]	0.764 [0.425]
GPT-5	0.533 [0.499]	0.610 [0.488]	0.597 [0.490]	0.476 [0.499]

以及各指标下每个专业领域的一致性结果。总而言之，我们的专家质量评估表明，数据质量较高，在测试集的随机样本中具有尚可的一致性水平。

表4：所有评估大语言模型的基准准确率、模型规模和训练数据集规模。

模型类型	模型类别	模型	模型规模	训练数据集规模	总体准确率
专有	通用	claude-sonnet-4-20250514	未知	未知	0.694
		gemini-2.0-flash	未知	未知	0.654
		gpt-4.1	未知	未知	0.673
		gpt-5	未知	未知	0.764
		gpt-o3	未知	未知	**0.765**
开源	通用	deepseek-chat-v3-0324	80亿个参数	1.8万亿	0.620
		qwen-plus	80亿个参数	18万亿	0.618
		llama-3.3-70b-instruct	700亿个参数	15万亿	0.547
		Llama-3.1-8B-instruct	80亿个参数	15万亿	0.170
	以阿拉伯语为中心	fanar-c-1-8.7b	87亿个参数	1万亿	0.498
		allam-7b-instruct	70亿个参数	5.2万亿	0.447
		c4ai-command-r7b-arabic-02-2025	70亿个参数	未知	0.381
	医学	medgemma-4b-it	40亿个参数	4万亿	0.390
		apollo-7b	70亿个参数	未知	0.238
		med42-8b	80亿个参数	151 + 1B	0.318
		bimedix-bi	270亿个参数	6.32亿	0.390

4.3 大语言模型裁判评估

我们的大语言模型评审实验结果汇总于表3和附录D中的表 D1，涵盖全部四项评估指标。结果显示，不同大语言模型之间存在中等程度的一致性，且与专家评估分数相近。为更好地理解这些结果与专家评估之间的比较情况，我们在附录D中详细列出了大语言模型评审结果。

4.4 最先进的大语言模型基准测试

我们在表4中报告了所有基准测试结果。结果表明，在 MedAraBench 上，通用模型和推理优化模型的表现始终优于专业模型或较小型开源模型。GPT-o3 和 GPT-5 的总体准确率均名列前茅（分别为 0.765 和 0.764），而 med42-8b 和 apollo-7b 等模型的得分最低（分别为 0.0.318 和 0.238）。附录E中的表E1和表E2按专业领域和级别，详细列出了15个评估模型的准确率分数。

4.5 少样本学习与 QLoRA

下方表5展示了我们在 Llama-3.1-8B-instruct 上采用少样本学习和 QLoRa 协议的结果。

表5：Llama-8.1B-instruct 的少样本及 QLoRA 微调性能与基线零样本准确率的比较

模型	基线准确率	少样本准确率	LoRA 准确率
Llama-3.1-8B-instruct	0.170	0.191	0.320

少样本和 QLoRA 微调实验相比基线零样本性能取得了显著提升。少样本学习带来了 12.4 % 的小幅提升（从 0.170 提高到 0.191），而 QLoRA 微调则使准确率提升了 88.2%，达到 0.320，模型性能几乎翻倍。

5 讨论

总体而言，本研究推出了 MedAraBench，这是一个包含训练集和测试集的新型 2.5 万题问题数据集。我们报告了最先进的大语言模型的性能基线结果，并指出了模型在零样本设置下能力方面的关键差异。通过揭示阿拉伯语医学理解方面的不足，MedAraBench 为开发更具包容性、多语言且专注于特定领域的语言模型提供了有益见解。
专家质量评估和大语言模型评审实验，为我们了解数据质量以及使用大语言模型评估医学数据集质量是否可行，提供了宝贵见解。具体而言，专家质量评估在全部 4 项评估指标上的平均分介于 0.653 至 0.767 之间，一致率和科恩卡帕系数分别介于 0.656 至 0.820 和 0.152 至 0.555 之间，表明各项指标的一致程度从轻微到一般不等。各评估指标的平均分表明数据质量中等至较高，标注员之间的一致程度一般；但这也说明，我们需要整理更多质量更高、临床相关性更强的基准数据集，才能恰当地评估大语言模型是否已准备好用于临床部署。总体而言，除问题难度外，大语言模型评审结果与专家对数据集质量的评估在所有指标上的一致性较弱至中等。具体来说，在医学准确性（分别为 0.577 和 0.505）、临床相关性（分别为 0.252 和 0.0.377）以及问题质量（分别为 0.407 和 0.336）方面，GPT-o3 与专家 A 和专家 B 的皮尔逊相关系数最高。对于问题难度，所有模型的结果都显示一致性较弱或不存在一致性；其中，Gemini-2.0-Flash 与专家 A 的相关性最高（0.019），Claude-4-Sonnet 与专家 B 的相关性更高（0.039）。我们的结果表明，大语言模型尚不能可靠地评估阿拉伯语医学数据，也凸显了进一步训练的必要性，以使其更好地符合人类专家的标准。

我们的基准评测结果与医学领域大语言模型基准测试的既有研究一致：在多项选择问答等结构化任务中，专有模型通常优于开放获取模型。Chen等人（2025）和Alonso等人（2024）此前已对此进行了验证，他们发现，在多种语言的医学问答任务中，专有模型的准确率表现优于开源模型。Abu Daoud等人（2025）进一步表明，在阿拉伯语医学多项选择题任务中，Gemini 1.5 Pro、Claude 3.5 Sonnet 和 GPT-4 等专有模型表现更优。我们的结果进一步印证了这些发现：所有专有模型的准确率分数都显著高于或接近开源模型。具体而言，专有模型 GPT-o3 和 GPT-5 在 MedAraBench 上的总体准确率最高（分别为 0.765 和 0.764），Claude-Sonnet-4 的表现也很强劲（0.694）。相比之下，即使是规模最大的开源通用模型，例如 llama-3.3-70b-instruct 和 qwen-plus，准确率也仅为 0.547 和 0.618；而以阿拉伯语为中心的领域模型和医学模型，例如 allam-7b-instruct（0.447）、fanar-c-1-8.7b（0.498）和 medgemma-4b-it（0.390），准确率仍低于 0.5。尽管训练数据集规模和

模型规模与性能相关（例如，llama-3.3-70b-instruct 和 deepseek-chat-v3-0324 等较大的模型优于较小的以阿拉伯语为中心的模型或纯医学模型），但这仍不足以让开源模型达到专有大语言模型的水平。我们还观察到，经过推理微调的通用模型始终优于以阿拉伯语为重点的模型或医学专用模型，这凸显了架构和数据规模以及在专有实验室中磨炼出的迁移学习能力的重要性。
专有模型表现更优，可能是因为其训练语料规模更大、在结构化数据集上进行了更强的预训练、接受了更全面的指令调优，并且相较于开源模型，其专有模型采用了专门的强化学习流程。此外，我们的结果显示，推理模型（GPT-5 和 GPT-o3）的表现显著优于其他模型，这表明将推理和可解释性融入整个医学自然语言处理领域，尤其是阿拉伯语医学自然语言处理领域，前景广阔且意义重大。

上文表5中的少样本学习和 QLoRA 实验显示，对于 Llama-3.1-8B-instruct，两种方法相较于零样本基线性能均有显著提升，但效果存在明显差异。少样本学习带来了 12.4% 的小幅提升（从 0.170 增至 0.191），这表明上下文示例有助于模型更好地理解医学问题格式和推理模式。此外，QLoRA 微调使准确率提高了 88.2%，达到 0.320，几乎使模型性能翻倍。QLoRA 微调的显著优势表明，对于将通用模型适配到阿拉伯语专业医学问答任务而言，利用领域特定医学数据进行参数高效适配，远比单独使用上下文学习有效。这些结果凸显了相较于提示工程，采用针对性训练方案对于在专业医学语言理解任务中实现可靠性能的重要性。

我们进一步通过比较当代模型与传统模型在 MedArabiQ 基准测试（Abu Daoud等，2025年）上的表现，评估模型跨代演进情况。下文附录G中的表 G1所示结果表明，所有当代模型在选择题任务上的表现均优于传统模型。此外，这项分析也使我们能够比较 MedArabiQ 与 MedAraBench 两个基准测试的优劣：gemini-2.0-flash、gpt-4.1、gpt-5、gpt-o3 和 qwen-plus 在 MedArabiQ 基准测试上的表现更好，而 claude-sonnet-4-20250514 和 llama-3.3-70b-instruct 在 MedAraBench 上表现更好。这表明 MedAraBench 总体上更具挑战性，凸显了它在推动医学阿拉伯语自然语言处理发展方面的价值。

然而，尽管模型性能代际间有了显著提升，表现最好的模型准确率仍为 0.765，未达到专家水平；这表明，在模型能够用于临床场景之前，仍有明显的提升空间。此外，这也引出了一个重要问题：大语言模型究竟学到了什么？高准确率并不一定意味着模型具备深入理解能力或临床推理能力。相反，模型可能利用统计关联和词汇模式，排除不合理的答案。例如，在预训练期间频繁接触某些疾病-治疗配对，可能使模型能够做出有依据的猜测，却不必推理症状进展或进行鉴别诊断。这一区别至关重要，尤其是在医学教育或决策支持等高风险应用中。未来的研究不应只评估答案是否正确，还应考察模型选择答案的依据，例如通过基于解释的任务或临床医生对模型论证的评分来实现。

6 局限性与未来工作

尽管本研究对大语言模型在阿拉伯语医学选择题上的表现进行了全面评估，并显著提升了基准测试能力，但仍存在若干局限。首先，由于多项选择题数据集的性质，该数据集只能用于评估大语言模型的分类任务，无法评估其生成任务。此外，尽管数据源并非结构化数字格式，需要投入大量工作进行数字化和清理，但我们仍无法完全排除数据污染的可能性。此外，我们的数据假设受试者熟练掌握现代标准阿拉伯语，而

现代标准阿拉伯语虽常用于正式场合，却未必完全符合实际语言使用情况。这可能影响 MedAraBench 对习惯使用方言或接受混合语言教学的学习者或从业者的普适性。

我们的专家质量评估实验还揭示了另一项局限。尽管数据集经过了两位专家临床医生的审阅，我们仍观察到他们的评估偶尔存在不一致。我们承认，专家之间的意见分歧和临床判断固有的主观性十分常见，但也认识到，未来的验证工作需要达成更广泛的共识。此外，与较高难度级别（Y3 为 12.04%、Y4 为 3.38%、Y5 为 5.14%）相比，我们的数据在较低难度级别上分布明显偏多（Y1 占 56%，Y2 占 22.89%）。这种分布是由于从区域医学资源库收集的源材料中，低年级内容所占比例过高。尽管这种评估方式限制了基准测试评估高级临床推理的能力，但它反映了当前阿拉伯语教育资源的现状，并通过为当前各难度级别和专业领域的阿拉伯语医学理解建立基线表现而具有价值，同时也揭示出，即使是基础医学推理，对当前的大语言模型而言仍具挑战性。此外，我们的数据仅为文本形式，因此难以适用于放射学和皮肤病学等专业领域所需的基于图像的推理；未来工作有必要将数据扩展至其他模态。

本研究主要聚焦于模型性能的零样本评估，在不进行进一步适配的情况下对其进行评估。尽管这是一种重要的评估框架，未来研究仍可探索少样本提示、思维链提示策略以及微调对提升模型性能的作用。此外，未来工作还可考虑纳入方言数据，以增强模型在不同阿拉伯语临床场景中的适应能力。未来研究的另一个重要方向是探讨如何利用阿拉伯语讲义，推动医学阿拉伯语自然语言处理超越选择题的范畴。

7 结论

总而言之，我们推出了一个大规模阿拉伯语医学基准测试，旨在评估大语言模型在经过整理的选择题上的零样本性能。MedAraBench 涵盖 19 个医学专科，横跨五个难度等级，为评估大语言模型的阿拉伯语医学推理提供了全面而细致的视角。我们的基准测试推动了阿拉伯语基准测试的发展，并揭示了当前大语言模型在低资源语言任务中的性能局限，以及采用稳健的多语言训练策略的必要性。未来工作应探索微调策略，并整理规模更大、质量更高且针对阿拉伯语医学语境的数据集。我们发布 MedAraBench，希望能为下游临床应用提供支持，并期待它成为推动阿拉伯语自然语言处理与医学人工智能交叉领域持续研究的催化剂。

伦理声明

作者声明，仅使用生成式人工智能工具协助进行 LaTeX 代码清理与格式调整；本项目的全部内容、分析与结论均由参与研究人员撰写并核实。

可复现性声明

The MedAraBench 数据集可在 `https://github.com/nyuad-cai/MedAraBench` 获取

致谢

本研究得到了 Meem基金会、纽约大学阿布扎比分校人工智能与机器人中心（由 Tamkeen 通过纽约大学阿布扎比分校研究院奖 CG010 提供资助）、网络安全中心（CCS）（由 Tamkeen 通过纽约大学阿布扎比分校研究院奖 G1104 提供资助）以及跨学科数据科学与人工智能中心（CIDSAI）（由 Tamkeen 通过纽约大学阿布扎比分校研究院奖 CG016 提供资助）的支持。研究使用了纽约大学阿布扎比分校（朱拜勒）的高性能计算资源。

参考文献

Umar Abbas。Fanar：一个以阿拉伯语为中心的多模态生成式人工智能平台。*arXiv* 预印本 *arXiv:2501.13944*，2025年。

Mariam Essam Abdelaziz、Mohanad A. Deif、Shabbab Ali Algamdi 和 Rania Elgohary。使用 AAFAQ 框架构建的阿拉伯语问题分类基准数据集。*Scientific Data*，2025年。

Mouath Abu Daoud、Chaimae Abou Zahir、Leen Kharouf、Walid AlEisawi、Nizar Habash 和 Farah Shamout。MedArabiQ：阿拉伯语医学任务上的大语言模型基准测试。*PMLR*，2025年。

Mohammed Al-Rukban。选择题测试的编制指南。*Journal of Family Community Medicine*，2006年。URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3410060/。

Ashwag Alasmari、Sarah Alhumoud 和 Waad Alshammari。AraMed：使用预训练变换器语言模型进行阿拉伯语医学问答。载于第六届开源阿拉伯语语料库与处理工具研讨会论文集 *(OSACT)*，第 50–56 页，2024。

Yazeed AlMoaiad。自然阿拉伯语处理中的挑战。*Edelweiss AppliedScience and Technology*，2022年。

Yazeed Alnumay。Command R7B Arabic：一款小型、面向企业、多语言的，文化意识的阿拉伯语大语言模型。*arXiv* 预印本 *arXiv:2503.14603v1*，2025年。

Iñigo Alonso、Maite Oronoz 和 Rodrigo Agerri。MedExpQA：用于医疗问答的大语言模型多语言基准测试。*Artificial Intelligence in Medicine*，155:102938，2024年9月。ISSN 0933-3657。doi: 10.1016/j.artmed.2024.102938。URLhttp://dx.doi.org/10.1016/j.artmed.2024.102938。

KhatraAmar。医学术语翻译中的阿拉伯语形态生产力。博士论文 - 马来西亚理科大学，2022年。

Anthropic。Claudeopus 4 与 claude sonnet 4——系统卡。https://www.anthropic.com/claude-4-system-card，2025年。

Yasser Mohamed Attia。在心脏和脑部疾病医学术语的阿拉伯语翻译中实现动态对等的挑战：困难与建议。塞得港大学文学院学报，30(30):14–31，2024。M Saiful Bari。Allam：面向阿拉伯语和英语的大语言模型。*arXiv*预印本 *arXiv:2407.15390*，2024。Rishi Bommasani、Drew A. Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S. Bernstein、Jeannette Bohg、Antoine Bosselut、Emma Brunskill、Erik Brynjolfsson、Shyamal Buch、Dallas Card、Rodrigo Castellon、Niladri Chatterji、Annie Chen、Kathleen Creel、Jared Quincy Davis、Dora Demszky、Chris Donahue、Moussa Doumbouya、Esin Durmus、Stefano Ermon、John Etchemendy、Kawin Ethayarajh、Li Fei-Fei、Chelsea Finn、Trevor Gale、Lauren Gillespie、Karan Goel、Noah Goodman、Shelby Grossman、Neel Guha、Tatsunori Hashimoto、Peter Henderson、John Hewitt、Daniel E. Ho、Jenny Hong、Kyle Hsu、Jing Huang、Thomas Icard、Saahil Jain、Dan Jurafsky、Pratyusha Kalluri、Siddharth Karamcheti、Geoff Keeling、Fereshte Khani、Omar Khattab、Pang Wei Koh、Mark Krass、Ranjay Krishna、Rohith Kuditipudi、Ananya Kumar、Faisal Ladhak、Mina Lee、Tony Lee、Jure Leskovec、Isabelle Levent、Xiang Lisa Li、Xuechen Li、Tengyu Ma、Ali Malik、Christopher D. Manning、Suvir Mirchandani、Eric Mitchell、Zanele Munyikwa、Suraj Nair、Avanika Narayan、Deepak Narayanan、Ben Newman、Allen Nie、Juan Carlos Niebles、Hamed Nilforoshan、Julian Nyarko、Giray Ogut、Laurel Orr、Isabel Papadimitriou、Joon Sung Park、Chris Piech、Eva Portelance、Christopher Potts、Aditi Raghunathan、Rob Reich、Hongyu Ren、Frieda Rong、Yusuf Roohani、Camilo Ruiz、Jack Ryan、Christopher R´e、Dorsa Sadigh、Shiori Sagawa、Keshav Santhanam、Andy Shih、

Krishnan Srinivasan、Alex Tamkin、Rohan Taori、Armin W. Thomas、Florian Tram`er、Rose E. Wang、William Wang、Bohan Wu、Jiajun Wu、Yuhuai Wu、Sang Michael Xie、Michihiro Yasunaga、Jiaxuan You、Matei Zaharia、Michael Zhang、Tianyi Zhang、Xikun Zhang、Yuhui Zhang、Lucia Zheng、Kaitlyn Zhou和Percy Liang。论基础模型的机遇与风险，2022年。URL https://arxiv.org/abs/2108.07258。

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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever 和 Dario Amodei。《语言模型是少样本学习者》，2020年。URL https://arxiv.org/abs/2005.14165。

Hanjie Chen、Zhouxiang Fang、Yash Singla 和 Mark Dredze。《大语言模型在回答和解释具有挑战性的医学问题方面的基准测试》，2025年。URL https://arxiv.org/abs/2402.18060。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、Parker Schuh、Kensen Shi、Sasha Tsvyashchenko、Joshua Maynez、Abhishek Rao、Parker Barnes、Yi Tay、Noam Shazeer、Vinodkumar Prabhakaran、Emily Reif、Nan Du、Ben Hutchinson、Reiner Pope、James Bradbury、Jacob Austin、Michael Isard、Guy Gur-Ari、Pengcheng Yin、Toju Duke、Anselm Levskaya、Sanjay Ghemawat、Sunipa Dev、Hen- ryk Michalewski、Xavier Garcia、Vedant Misra、Kevin Robinson、Liam Fedus、Denny Zhou、Daphne Ippolito、David Luan、Hyeontaek Lim、Barret Zoph、Alexander Spiridonov、Ryan Sepassi、David Dohan、Shivani Agrawal、Mark Omernick、Andrew M. Dai、Thanumalayan Sankaranarayana Pillai、Marie Pellat、Aitor Lewkowycz、Erica Moreira、Rewon Child、Oleksandr Polozov、Katherine Lee、Zongwei Zhou、Xuezhi Wang、Brennan Saeta、Mark Diaz、Orhan Firat、Michele Catasta、Jason Wei、Kathy Meier-Hellstern、Douglas Eck、Jeff Dean、Slav Petrov 和 Noah Fiedel。《PaLM：借助路径扩展语言建模》，2022年。URL https://arxiv.org/abs/2204.02311。

Clement Christophe. Med42-v2：临床大语言模型套件。*arXiv*预印本 *arXiv:2408.06142*，2024。

William Gemmell Cochran。抽样技术。*John Wiley & Sons*，1977。

DeepSeek-AI。Deepseek-v3技术报告。*arXiv*预印本 *arXiv:2412.19437*，2024 .

Ali Farghali和Khaled Shaalan。阿拉伯语自然语言处理：挑战与解决方案。*ACM*亚洲语言信息处理汇刊（*TALIP*），2009。

Google Cloud。Gemini 2.0 Flash——Vertex AI生成式人工智能。https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash，2025年5月。

A. Grattafiori等。Llama 3模型群。*arXiv*预印本 *arXiv:2407.21783*，2024a .

Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman和Akhil Mathur等。Llama 3模型群。*arXiv*预印本 *arXiv:2407.21783*，2024b。

Nizar Habash。阿拉伯语自然语言处理导论。*ACL*教程，2005。

Dan Hendrycks、Collin Burns、Steven Basart、Andy Zou、Mantas Mazeika、Dawn Song和Jaco b Steinhardt。大规模多任务语言理解的测量，2021。URLhttps://arxiv.org/abs/2009.03300。

Pedram Hosseini。长篇医疗问答基准测试。*arXiv* 预印本，2024。

Shadi Iskander、Nachshon Cohen 和 Zohar Karnin。质量至关重要：评估用于工具使用大语言模型的合成数据。*arXiv* 预印本 *arXiv:2409.16341*，2024。网址 https://arxiv.org/abs/2409.16341。

Luyi Jiang、Jiayuan Chen、Lu Lu、Xinwei Peng、Lihao Liu、Junjun He 和 Jie Xu。中文医学大语言模型基准测试：基于 MedBench 的性能差距分析与分层优化策略，2025年。URL https://arxiv.org/abs/2503.07306。

Di Jin、Eileen Pan、Nassim Oufattole、Wei-Hung Weng、Hanyi Fang 和 Peter Szolovits。该患者患有什么疾病？来自医学考试的大规模开放域问答数据集，2020年。URL https://arxiv.org/abs/2009.13081。

Di Jin、Yansong Pan、Tao Ouyang 和 Pascale Fung。该患者患有什么疾病？来自医学考试的大规模开放域问答数据集。*arXiv* 预印本 *arXiv:2009.13081*，2021年。URL https://arxiv.org/abs/2009.13081。提交至 AAAI 2021。

Qiao Jin、Bhuwan Dhingra、Zhengping Liu、William Cohen 和 Xinghua Lu。Pubmedqa：一个用于生物医学研究问答的数据集。载于 *2019 年自然语言处理实证方法会议暨第九届国际自然语言处理联合会议（EMNLP-IJCNLP）论文集*，第 2567–2577 页，2019 年。

Praveen K. Kanithi、Cl´ement Christophe、Marco A. F. Pimentel、Tathagata Raha、Nada Saadi、Hamza Javed、Svetlana Maslenkova、Nasir Hayat、Ronnie Rajan 和 Shadab Khan。Medic：迈向用于临床应用的大语言模型评估综合框架。*arXiv* 预印本 *arXiv:2409.07314*，2024 年。

Tobi Olatunji、Charles Nimo、Abraham Owodunni、Tassallah Abdullahi、Emmanuel Ayo- dele、Mardhiyah Sanni 和 Chinemelu Aka。Afrimed-qa：一个泛非、多专科的医学问答基准数据集。*arXiv* 预印本 *arXiv:2411.15640*，2024 年。网址 https://arxiv.org/html/2411.15640。

OpenAI。《GPT-4 技术报告》。https://openai.com/index/gpt-4-1/，2025年4月。

OpenAI。《GPT-5 系统卡》。https://cdn.openai.com/gpt-5-system-card.pdf，2025年 b.

OpenAI。《OpenAI o3 和 o4-mini 系统卡》。https://cdn.openai.com/pdf/o3-and-o4-mini-system-card.pdf，2025c。

Ankit Pal、Logesh Kumar Umapathi 和 Malaikannan Sankarasubbu。MedMCQA：用于医学领域问答的大规模多学科选择题数据集。发表于健康、推理与学习会议，第 248–260 页。PMLR，2022年。

Sara Pieri、Sahal Shaji Mullappilly、Fahad Shahbaz Khan、Rao Muhammad Anwer、Salman Khan、Timothy Baldwin 和 Hisham Cholakkal。BiMedix：双语医学混合专家大语言模型。*arXiv* 预印本 *arXiv:2402.13253*，2024。

Rick Rejeleene、Xiaowei Xu 和 Johm Talburt。迈向可信语言模型：探究大语言模型的信息质量。*arXiv* 预印本 *arXiv:2401.13086*，2024年。URL https://arxiv.org/abs/2401.13086。

Andrew Sellergren。MedGemma 技术报告。*arXiv* 预印本 *arXiv:2507.05201*，20 25.

Shivalika Singh。Global MMLU：理解并应对多语言评估中的文化和语言偏见。第*63*届计算语言学协会年会论文集，2025年。

Felix Stahlberg。神经机器翻译：综述与调查，2020年。URL https://arxiv.org/abs/1912.02047。

Bin Wang、Chengwei Wei、Zhengyuan Liu、Geyu Lin 和 Nancy Chen。《噪声指令下大语言模型的韧性》。*EMNLP*，2024。

Cunxiang Wang、Pai Liu 和 Yue Zhang。生成式预训练语言模型能否充当闭卷问答的知识库？，2021。URL https://arxiv.org/abs/2106.01561。

Xidong Wang。Apollo：一款轻量级多语言医疗大语言模型，旨在让60亿人都能受益于医学人工智能。*arXiv* 预印本 *arXiv:2403.03640*，2024。

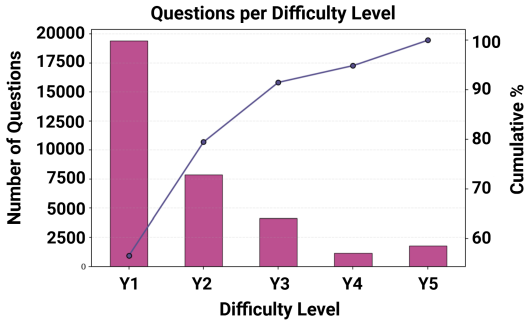
An Yang 等。Qwen2.5 技术报告。*arXiv* 预印本 *arXiv:2412.15115*，2024。网址 https://arxiv.org/abs/2412.15115。

Han Yang、Mingchen Li、Huixue Zhou、Yongkang Xiao、Qian Fang 和 Rui Zhang。用于医疗问答的集成学习大语言模型协同：设计与评估研究。《医学互联网研究杂志》，2025年。

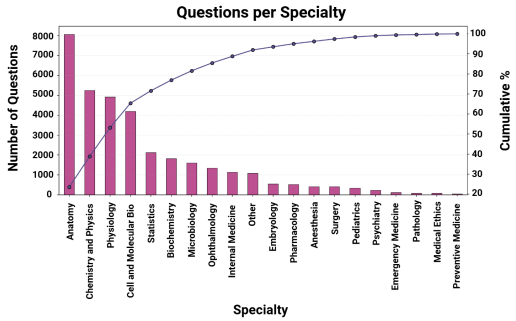
A 数据分析

A.1 数据摘要

下方图 A1 展示了 MedAraBench 数据集的概况，按难度级别和专业领域划分了问题分布。



(a)



(b)

图 A1：数据集按难度级别和专业领域的概览。

从专业领域来看，数据集中最大的子集属于解剖学（6100 道题，占数据集的 24.61%）和生理学（3302 道题，占 13.32%），而最小的子集属于病理学（56 道题，占 0.23%）。表 A1 和上方的图 A1 (b) 展示了按专业领域划分的问题分布详情。

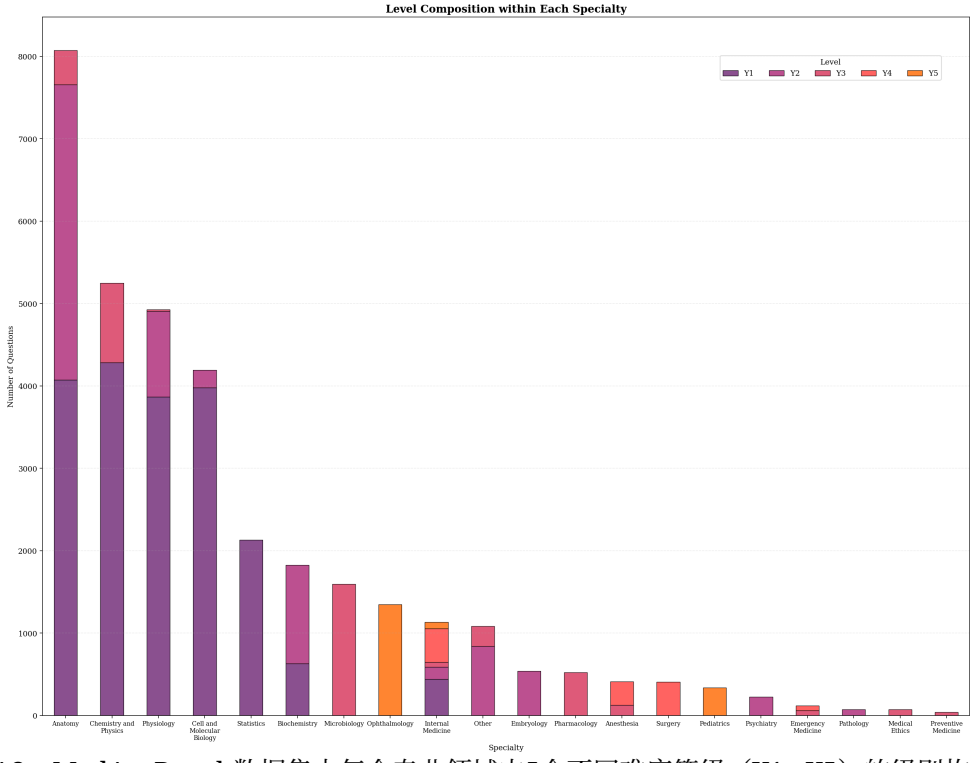
另一方面，从难度级别来看，数据集中最大的子集属于 Y1（15095 道题，占数据集的 60.89%），而最小的子集属于 Y4（1033 道题，占 4.17%）。表 A2 和上方的图 A1 (a) 展示了按级别划分的问题分布详情。此外，图A2展示了 MedAraBench 所含 16 个专业领域中五个级别的构成。

表 A1：各医学专科的问题分布。

医学专科	问题数量	百分比
解剖学	6100	24.61%
麻醉学	405	1.63%
生物化学	1387	5.59%
细胞与分子生物学	3155	12.73%
化学与物理学	3894	15.71%
胚胎学	184	0.74%
急诊医学	120	0.48%
内科学	762	3.07%
微生物学	750	3.03%
眼科学	1318	5.32%
病理学	56	0.23%
药理学	329	1.33%
生理学	3302	13.32%
统计学	1795	7.24%
外科学	357	1.44%

表 A2：不同难度级别的问题分布。

难度级别	问题数量	百分比
Y1	15095	60.89%
Y2	4954	19.98%
Y3	2313	9.33%
Y4	1033	4.17%
Y5	1396	5.63%



图A2：MedAraBench数据集中每个专业领域内5个不同难度等级（Y1 - Y5）的级别构成。

A.2 TokenLength 分布

我们的数据集包含 24,791 个问题。问题长度适中，平均总长度为 37.86 个字符。答案的格式各不相同，其中 18,554 个问题有四个答案选项（A、B、C 和 D），6,228 个问题有五个答案选项（A、B、C、D 和 E）。各数据集的答案平均长度为 171.09 个字符。下方的图 A3 详细展示了整个数据集中各类文本长度的分布。

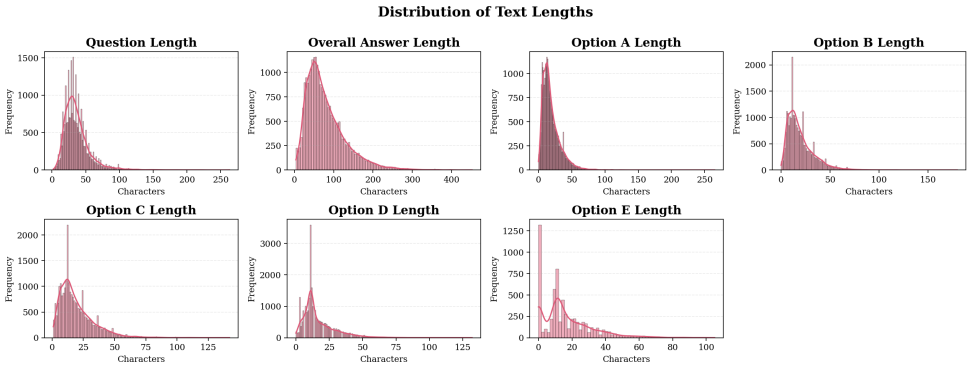


图 A3：MedAraBench 数据集中文本长度的分布：(a) 问题长度分布；(b) 答案长度分布；(c) 选项 A 长度分布；(d) 选项 B 长度分布；(e) 选项 C 长度分布；(f) 选项 D 长度分布；以及 (g) 选项 E 长度分布。

B 基准测试与少样本学习提示

B.1 基准测试提示

清单 1：医学选择题评估的基准测试提示

```
"You are an expert medical virtual assistant.
Please provide the correct answer letter (A, B, C, or D)
for the following Arabic medical multiple-choice question.
Question:
{question_text_in_Arabic}

Options:
A:{option_A_in_Arabic}
B:{option_B_in_Arabic}
C:{option_C_in Arabic}
D:{option_D_in Arabic}
Answer:"
```

B.2 少样本学习提示

清单 2：医学选择题评估的少样本提示

```
"You are an expert medical assistant. You will be provided with
    a few Arabic
medical sample questions and their correct answers, then a new
   question
that you must answer.

Sample Questions:
{sample_questions}

Now, the new question:
Question: {question}
Options:
{options_text}
(Please answer with the letter choice {allowed_letters} ONLY):"
```

样例问题如下：

السؤال: المفصل الحلقي الطرجهالي من المفاصل؟
الخيارات: أ: الليفي, ب: الزليلي, ج: الزجاجي, د: المرن
الإجابة: ب

السؤال: فيما يخص العصب الزندي كل مما يلي صحيح ماعدا
الخيارات: أ: ينشأ من الحبل الأنسي للضفيرة العضدية, ب: يدخل الساعد بين رأسي الكابة المدورة, ج: يخترق الحاجز الأنسي للعضد برفقة الشريان الجانبي الزندي العلوي, د: يكون وحشي وتر قابضة الرسغ الزندية عند المعصم, هـ: يعصب العضلة قابضة الرسغ الزندية
الإجابة: ب

السؤال: من أسباب نقص سكر الدم الارتكاسي:
الخيارات: أ: نقص سكر الدم الهضمي, ب: الحساسية للوسين, ج: الغالاكتوزيمية, د: كل ما سبق صحيح
الإجابة: د

C 专家质量评估详情

在本附录中，我们详细列出了第 2 节介绍的专家评估结果。正文总结了所有专业领域的整体平均值和一致性水平，而此处我们报告各专业领域的结果，以便更细致地了解模型性能和标注者一致性。

图 C1–C4 展示了四项评估指标（医学准确性、临床相关性、问题难度和问题质量）中各专业领域的标注者评分分布。相应的表 C1–C4 列出了各专业领域的平均分、已评估问题数量、一致率和科恩卡帕系数。这些结果共同突显了不同领域之间的差异，并为解读正文中展示的汇总质量指标提供了背景。

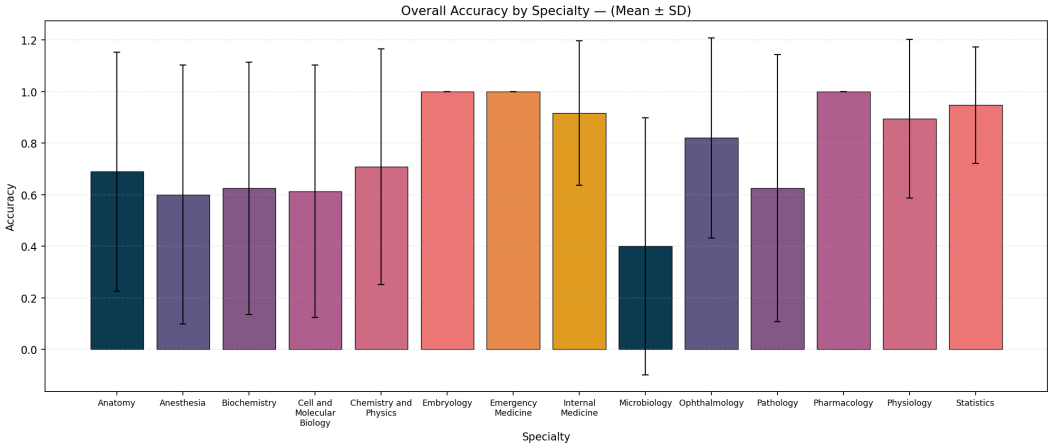


图 C1：各专业领域标注者的总体准确率评分。

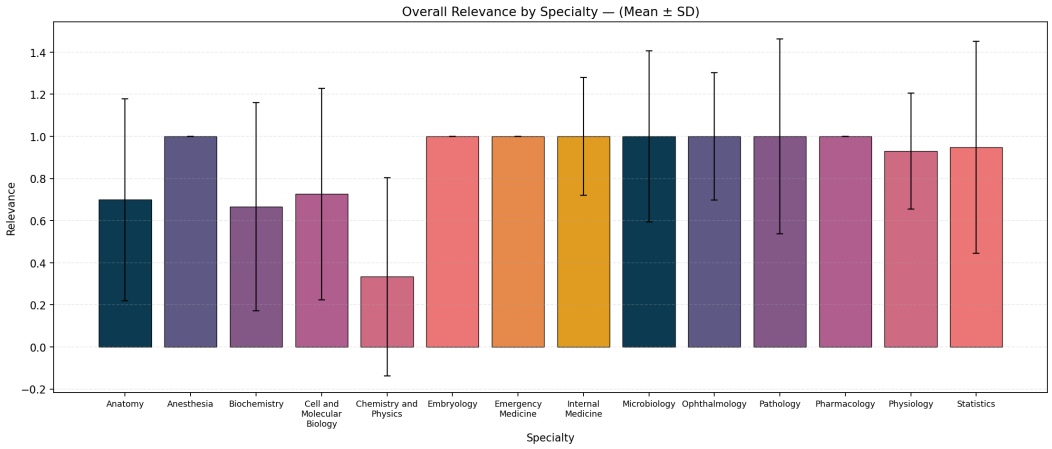


图 C2：各专业领域标注者的总体相关性评分。

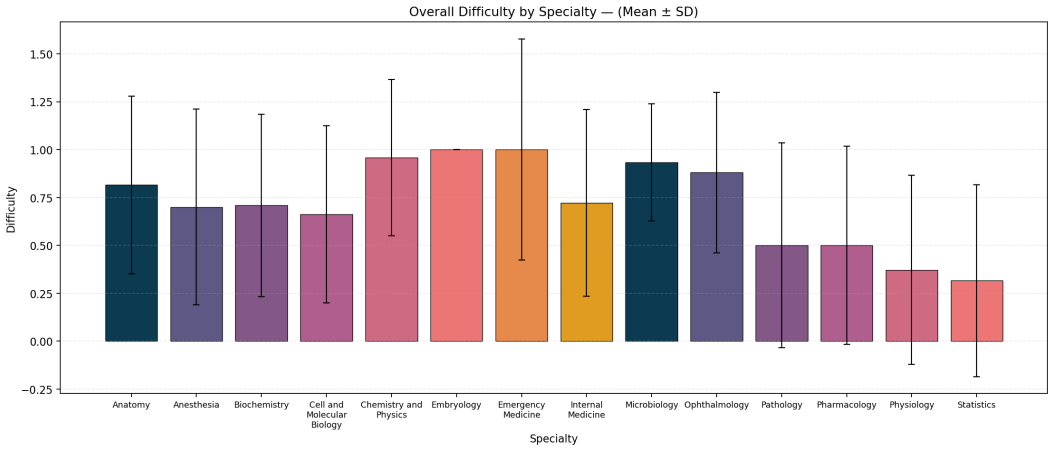


图 C3：各专业领域的整体标注者难度评分。

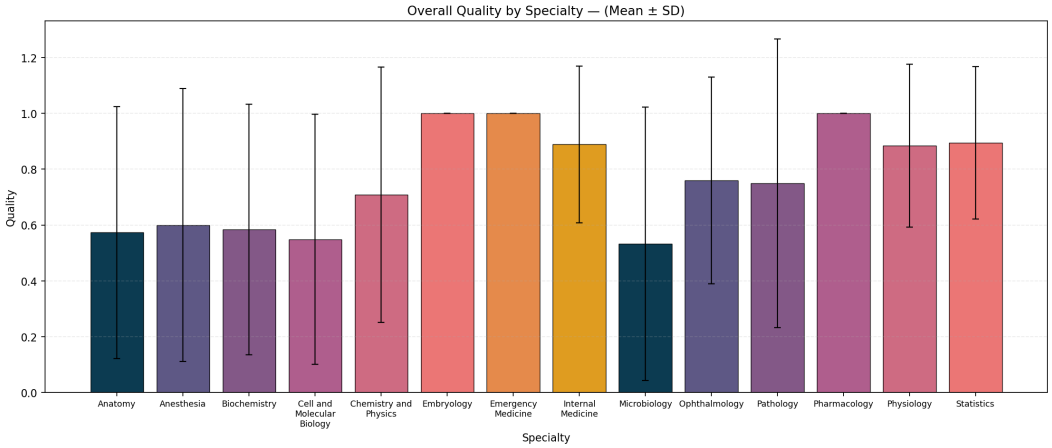


图 C4：各专业领域的整体标注员质量评分。

表 C1：各专业领域的标注者准确率评分。

专业领域	平均值 [标准差]	问题数量	一致率	Cohen's Kappa
解剖学	0.689 [0.464]	103	0.748	0.431
麻醉学	0.600 [0.503]	10	0.600	0.167
生物化学	0.625 [0.489]	24	0.750	0.471
细胞与分子生物学	0.613 [0.489]	62	0.774	0.529
化学与物理学	0.708 [0.457]	48	0.917	0.798
胚胎学	1.000 [0.000]	1	1.000	-
急诊医学	1.000 [0.000]	2	1.000	-
内科学	0.917 [0.280]	18	0.944	0.640
微生物学	0.400 [0.498]	15	0.733	0.455
眼科学	0.820 [0.388]	25	0.880	0.603
病理学	0.625 [0.518]	4	0.750	-
药理学	1.000 [0.000]	4	1.000	-
生理学	0.895 [0.308]	43	0.884	0.380
统计学	0.947 [0.226]	19	1.000	1.000

表 C2：各专业领域的标注者相关性评分。

专业领域	平均值 [标准差]	问题数量	百分比一致性	Cohen's Kappa
解剖学	0.699 [0.479]	103	0.641	0.225
麻醉学	1.000 [0.000]	10	1.000	-
生物化学	0.667 [0.494]	24	0.708	0.400
细胞与分子生物学	0.726 [0.501]	62	0.435	0.095
化学与物理学	0.333 [0.470]	48	0.688	0.286
胚胎学	1.000 [0.000]	1	1.000	-
急诊医学	1.000 [0.000]	2	1.000	-
内科学	1.000 [0.280]	18	0.833	0.000
微生物学	1.000 [0.407]	15	0.600	0.000
眼科学	1.000 [0.303]	25	0.800	0.000
病理学	1.000 [0.463]	4	0.500	-
药理学	1.000 [0.000]	4	1.000	-
生理学	0.930 [0.275]	43	0.884	0.224
统计学	0.947 [0.504]	19	0.211	0.021

表 C3：各专业领域的标注者难度评分。

专业领域	平均值 [标准差]	问题数量	一致率	Cohen's Kappa
解剖学	0.816 [0.464]	103	0.650	0.240
麻醉学	0.700 [0.510]	10	0.500	0.074
生物化学	0.708 [0.476]	24	0.750	0.442
细胞与分子生物学	0.661 [0.463]	62	0.710	0.320
化学与物理学	0.958 [0.408]	48	0.625	0.027
胚胎学	1.000 [0.000]	1	1.000	-
急诊医学	1.000 [0.577]	2	0.000	-
内科学	0.722 [0.487]	18	0.611	0.182
微生物学	0.933 [0.305]	15	0.800	-0.098
眼科学	0.880 [0.418]	25	0.720	0.229
病理学	0.500 [0.535]	4	1.000	-
药理学	0.500 [0.518]	4	0.750	-
生理学	0.372 [0.494]	43	0.651	0.281
统计学	0.316 [0.500]	19	0.368	-0.009

表 C4：各专业领域的标注员质量评分。

专业领域	平均值 [标准差]	问题数量	一致率	Cohen's Kappa
解剖学	0.573 [0.451]	103	0.573	0.044
麻醉学	0.600 [0.489]	10	0.700	0.348
生物化学	0.583 [0.449]	24	0.625	0.143
细胞与分子生物学	0.548 [0.448]	62	0.484	-0.120
化学与物理学	0.708 [0.457]	48	0.792	0.496
胚胎学	1.000 [0.000]	1	1.000	-
急诊医学	1.000 [0.000]	2	1.000	-
内科学	0.889 [0.280]	18	0.944	0.640
微生物学	0.533 [0.490]	15	0.533	0.037
眼科学	0.760 [0.370]	25	0.840	0.432
病理学	0.750 [0.518]	4	0.750	-
药理学	1.000 [0.000]	4	1.000	-
生理学	0.884 [0.292]	43	0.860	0.178
统计学	0.895 [0.273]	19	0.842	-0.075

D 大语言模型评审

向模型提供了完整的选择题测试集（题干、选项和正确答案），并指示其仅返回有效的 JSON 输出。具体提示如下：

代码清单 3：提供给大语言模型的提示

```
You are a medical education expert. Evaluate the following
```

此设置确保大语言模型输出经过标准化且可由机器读取。通过汇总数千道测试题的得分，我们得到了描述性统计和模型分布，从而能够进行比单纯二元人工评分更细致的分析。

为考察模型与专家评估者之间的一致性，我们按逐题计算了皮尔逊相关系数。分析结果见下方表 D1。皮尔逊相关结果表明，在与专家评估结果的一致程度方面，GPT-o3 的表现最佳。因此，我们选择使用 GPT-o3 对 MedAraBench 的整个训练集进行大语言模型即裁判评估，从而全面评估整个数据集。此外，我们进行了 t 检验，以比较 GPT-o3 在训练集和测试集上的得分；结果发现，在全部 4 项评估指标上均无显著差异，表明不同划分的质量一致。该实验结果见下方表 D2。

表 D1：模型与专家评分之间的皮尔逊相关系数。

	专家 A				**专家 B**			
模型	GPT-o3	Claude-4-Sonnet	Gemini-2.0-Flash	GPT-5	GPT-o3	Claude-4-Sonnet	Gemini-2.0-Flash	GPT-5
医学准确性	**0.577**	0.065	0.023	0.043	**0.505**	0.053	0.053	0.068
临床相关性	**0.252**	0.165	0.176	0.071	**0.377**	0.131	0.131	0.104
问题质量	**0.407**	0.007	0.023	0.044	**0.336**	0.062	0.062	0.033
问题难度	-0.114	-0.070	**0.019**	-0.116	-0.018	**0.039**	**0.039**	-0.049

表 D2：GPT-o3 大语言模型评审在训练集和测试集上的评分（均值 [标准差]），以及t-检验的 p 值。

数据集	**医学准确性**	**临床相关性**	**问题质量**	**问题难度**
训练集	0.638 [0.481]	0.821 [0.383]	0.839 [0.367]	0.561 [0.496]
测试集	0.673 [0.469]	0.827 [0.378]	0.588 [0.492]	0.841 [0.366]
P值	0.0001	0.0362	0.0001	0.0001

E 各专业领域和级别的表现

为补充整体评估，我们分析了模型在不同医学专科和难度等级上的准确率。这种细分分析揭示了模型在特定领域的优势与不足，以及模型表现如何随五个课程级别（Y1—Y5）变化。下文表E1和表E2详细列出了各模型在不同专业领域和级别上的准确率。

就专业领域表现而言，我们的分析显示各模型之间存在明显差异：GPT-5 和 GPT-o3 在大多数领域表现出色，而 apollo-7b 和 med42-8b 的表现始终较差。值得注意的是，规模较小的 medgemma-4b-it 在病理学领域表现出色（0.574），这表明专门化训练有时能够弥补规模上的不足。

就难度级别表现而言，所有模型的准确率都从 Y1 到 Y3 逐渐下降，随后在 Y4—Y5 回升。GPT-5 在五个级别中的四个级别保持最高表现，而较弱模型在中间级别的表现下降幅度更大。这一稳定趋势表明，Y3 题目可能需要更复杂的临床推理，因此对所有模型架构而言都是最具挑战性的。

不同难度等级之间显著的性能差异，凸显了当前大语言模型在医学应用中的关键局限。从 Y1 到 Y3，性能平均下降 15%–20%，这表明模型虽然擅长在课程早期阶段进行事实回忆，却难以胜任中级阶段所需的综合性临床推理。这种难度扩展效应在较小模型中最为明显，说明模型规模和精细的训练对于处理复杂的医学推理任务至关重要。这些发现表明，与复杂的临床决策相比，大语言模型或许更适用于基础知识和专业领域。

Physiology	Statistics	Surgery
0.484	0.624	0.430
0.183	0.253	0.109
0.442	0.571	0.391
0.747	0.794	0.703
0.693	0.750	0.562
0.496	0.621	0.391
0.702	0.772	0.648
0.757	0.794	0.656
0.822	0.810	**0.766**
0.823	**0.835**	0.727
0.349	0.478	0.328
0.598	0.695	0.539
0.321	0.475	0.297
0.424	0.549	0.320
0.682	0.766	0.648

med42-8b	medgemma-4b-it	qwen-plus
0.339	0.416	0.644
0.322	0.357	0.581
0.244	0.297	0.490
0.260	0.321	0.638
0.226	0.354	0.549

F 答案选项分布平衡

在构建数据集之后，我们观察到答案分布存在虽小但显著的不平衡。

这可能导致模型偏差，使其偏向任何特定的答案位置（例如，总是选择“A”）。因此，为解决这一问题，我们分析并调整了所有子集和划分中正确答案选项的分布，通过进行最小限度的针对性调整（例如，在可能时重新排序选项），使正确答案频率更接近均匀性。这一优化是针对每个答案格式组（ABCD、ABCDE、ABCDEF）的训练集和测试集分别独立完成的。该调整确保了正确答案的均衡表示，并有助于降低模型学习基于位置的启发式方法的可能性。

图 F1–F3 展示了所得分布。

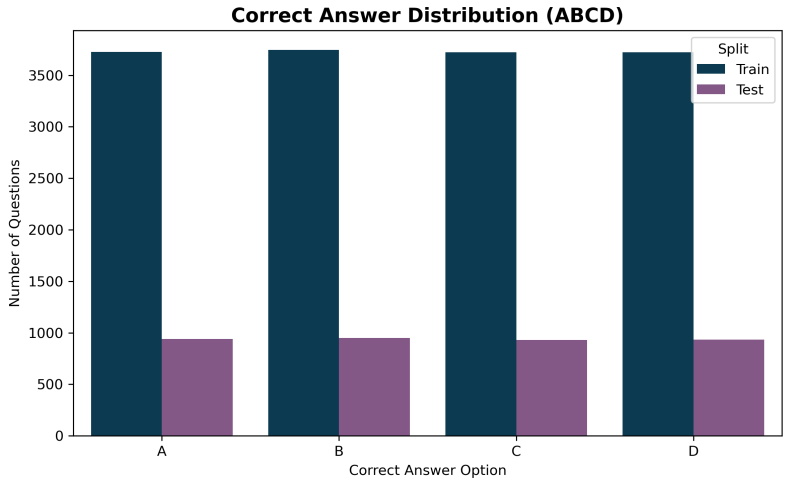


图 F1：ABCD 格式的答案选项分布（训练集/测试集）

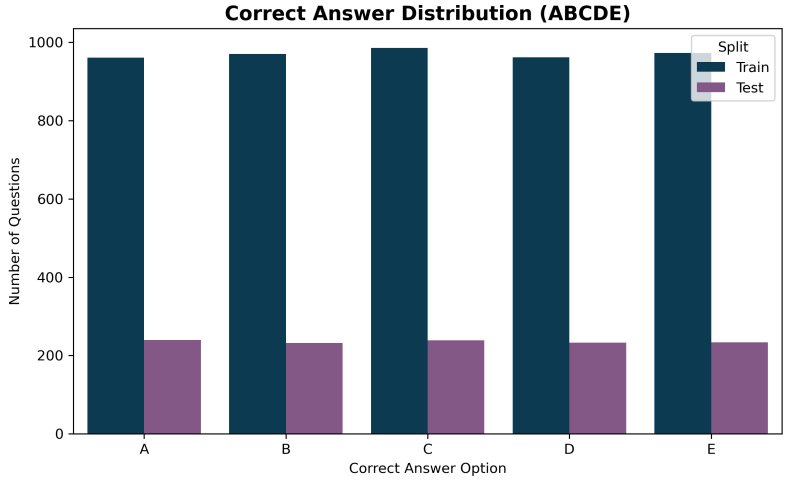


图 F2：ABCDE 格式的答案选项分布（训练集/测试集）

我们进一步使用卡方拟合优度检验验证了这种平衡。检验结果确认，ABCD 和 ABCDE 划分中的均匀性不存在显著偏差（$p > 0.97$），这表明观察到的答案分布与均匀分布不存在显著差异。由于

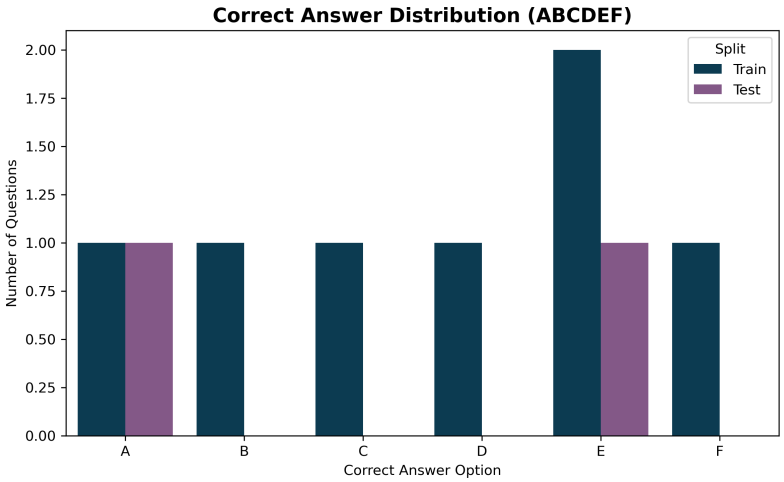


图 F3：ABCDEF 格式的答案选项分布（训练集/测试集）

样本量不足（$\mathrm{n} = 2$），因此未将 ABCDEF 格式纳入测试集中的这项分析。所有检验均支持这样的结论：这些分布与均匀分布不存在显著差异。
所得的 χ^2 值和 p值如下：

- **ABCD（训练集）**： $\chi^2 = 0.099, \quad p = 0.992$
- **ABCD（测试集）**： $\chi^2 = 0.212, \quad p = 0.976$
- **ABCDE（训练集）**： $\chi^2 = 0.422, \quad p = 0.981$
- **ABCDE（测试集）**： $\chi^2 = 0.226, \; p = 0.994$
- **ABCDEF（训练集）**： $\chi^2 = 0.714, \; p = 0.982$
- **ABCDEF（测试集）**： $\chi^2 = 6.02, \; p = 0.304$

这一调整确保了正确答案的均衡呈现，并有助于降低模型学习基于位置的启发式方法的可能性。

G 当代模型与传统模型评估

在本节中，我们通过在 MedArabiQ Abu Daoud 等人（2025）数据集上评估当代模型及其旧版对应模型，并比较它们的准确率分数，从而评估当代模型相对于旧版对应模型的演进。此外，该分析使我们能够比较模型在 MedArabiQ 和 MedAraBench 基准测试上的分数，从而为每个基准测试的价值提供有价值的洞见。

结果如下表 G1 所示。我们在 MedArabiQ 基准测试上的实验结果表明，所有当代模型在选择题任务上的表现均优于旧版模型。此外，我们的结果显示，gemini-2.0-flash、gpt-4.1、GPT-5、GPT-o3 和 qwen-plus 在 MedArabiQ 基准测试上表现更好，而 claude-sonnet-4-20250514 和 llama-3.3-70b-instruct 在 MedAraBench 上表现更好。这表明 MedAraBench 基准测试总体上更具挑战性，因此对医学阿拉伯语自然语言处理的发展具有更大价值。需要注意的是，旧版模型 gemini-1.5-pro 和 claude-3.5-sonnet 已被弃用，这使我们无法在 MedAraBench 上对其进行评估，从而无法在各基准测试之间进行更好的比较。

表 G1：传统模型与当代模型在 MedArabiQ 和 MedAraBench 上的模型性能

MedArabiQ				**MedAraBench**	
传统模型	**准确率**	**当代模型**	**准确率**	**当代模型**	**准确率**
claude-sonnet-3.5	0.535	claude-sonnet-4-20250514	0.6869	claude-sonnet-4-20250514	**0.6937**
gemini-1.5-pro	0.575	gemini-2.0-flash	**0.7273**	gemini-2.0-flash	0.6539
		gpt-4.1	**0.8081**	gpt-4.1	0.6733
gpt-4	0.535	GPT-5	**0.8586**	GPT-5	0.7642
		GPT-o3	**0.8384**	GPT-o3	0.7652
Llama-3.1-8B	0.262	llama-3.3-70b-instruct	0.4949	llama-3.3-70b-instruct	**0.5466**
通义千问2.5-7B	0.380	qwen-plus	**0.6566**	qwen-plus	0.6177