AraTable: 基准测试 LLM 对阿拉伯表格数据的推理和理解

Rana Alshaikh[1], Israa Alghanmi[2], Shelan Jeawak[3*]

[1]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, 21911, Saudi Arabia .

[2]Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia.

[3*]School of Computing and Creative Technologies, University of the West of England, Bristol, UK.

*Corresponding author(s). E-mail(s): shelan.jeawak@uwe.ac.uk;
Contributing authors: raalshaikh@kau.edu.sa; iaalghanmi@uj.edu.sa;

Abstract

The cognitive and reasoning abilities of large language models (LLMs) have enabled remarkable progress in natural language processing. However, their performance in interpreting structured data, especially in tabular formats, remains limited. Although benchmarks for English tabular data are widely available, Arabic is still underrepresented because of the limited availability of public resources and its unique language features. To address this gap, we present AraTable, a novel and comprehensive benchmark designed to evaluate the reasoning and understanding capabilities of LLMs when applied to Arabic tabular data. AraTable consists of various evaluation tasks, such as direct question answering, fact verification, and complex reasoning, involving a wide range of Arabic tabular sources. Our methodology follows a hybrid pipeline, where initial content is generated by LLMs and subsequently filtered and verified by human experts to ensure high dataset quality. Initial analyses using AraTable show that, while LLMs perform adequately on simpler tabular tasks such as direct question answering, they continue to face significant cognitive challenges when tasks require deeper reasoning and fact verification. This indicates that there are substantial opportunities for future work to improve performance on complex tabular reasoning tasks. We also propose a fully automated evaluation framework that uses a self-deliberation mechanism and achieves performance nearly identical to that of

human judges. This research provides a valuable, publicly available resource and evaluation framework that can help accelerate the development of foundational models for processing and analysing Arabic structured data.

Keywords: Arabic Benchmark, Tabular Data, Automatic Evaluation, LLMs' Cognitive Capacity

1 引言

大型语言模型（LLM）在广泛的自然语言处理（NLP）任务中展现出卓越的能力，在文本生成 [1, 2] 、翻译 [3, 4] 和问答 [5, 6] 等领域实现了最先进的性能。它们处理和理解复杂文本信息的能力为多种数据驱动应用的自动化和增强开辟了新途径。然而，LLM 的性能在很大程度上依赖于输入数据的性质和结构。尽管在处理非结构化文本方面取得了显著进展，但评估和提高 LLM 在结构化数据，尤其是表格数据上的性能，仍然是一个关键的研究领域。

表格数据广泛应用于金融、医疗保健和公共管理等各个领域，并在行和列中组织了有价值的信息 [7] 。从表格数据中提取洞察和回答问题需要模型不仅理解自然语言查询，还要理解表结构、列与行之间的关系，以及各个单元格内的具体值。评估 LLMs 在涉及表格数据的任务中的表现，如基于表格信息的问题回答、对表格的事实验证以及基于表格信息的复杂推理，提出了独特的挑战 [8–10] 。这些任务需要超越简单文字理解的能力，包括符号推理、数据聚合、比较、数学计算以及基于结构化事实的逻辑推理。

尽管对于评估 LLM 在各种任务上的性能有一些通用基准，最近一些研究也关注了关于英语表格数据的基准，如 TableBench [11] 、DataBench [12] 以及最近的 MMTU [13] ，但在专门设计来评估 LLM 在阿拉伯语表格数据上性能的全面基准方面存在显著缺口。阿拉伯语自身具有一系列语言复杂性，包括丰富的词形变化、各种方言和复杂的语法，这可能会影响 LLM 处理和理解阿拉伯文本的能力，从而影响对阿拉伯语表格数据的处理。现有的阿拉伯语 LLM 评估工作，如 Open Arabic LLM Leaderboard [1] ，通常侧重于一般语言理解，而较少关注处理结构化阿拉伯信息的挑战。

为了解决这一关键差距，本研究引入了 AraTable，这是一个专门用于评估大型语言模型在阿拉伯语表格式数据上表现的新基准。该基准测试主要集中在三个关键任务：直接问答、事实验证和推理，涵盖了一系列复杂性水平，并需要与表结构和内容的不同类型交互。我们考虑了来自各种来源的阿拉伯语表格的多样化数据集，包括维基百科、现实数据和大型语言模型生成的数据。通过使用 AraTable 的综合分析，我们评估了几个著名大型语言模型的表现，具体包括 Llama 3.3 70B [14] 、Mistral Large [15] 、DeepSeek-V3 [16] 和 Jais 70B [17] 。在所有数据集中，我们的评估显示，DeepSeek-V3 始终优于其他模型，紧随其后的是 Llama 3.3 70B 和 Mistral Large，而 Jais 模型的准确率明显较低。所有模型在直接问答（QA）任务上都表现得较为轻松，突显了在处理更复杂的推理问题时存在明显的性能差距，准确率往往低于 60 % 。这表明当前大型语言模型在阿拉伯语表格式数据上进行复杂逻辑推理的能力存在根本限制，而纯阿拉伯语模型，如 Jais，在复杂的表格推理方面是不足的。

此外，我们引入并验证了一个稳健的评估框架，专门用于评估大型语言模型（LLMs）的自由文本回答，并将其与真实答案进行比较。如图 2 所示，我们首先考虑了多轮次的人类评估以确保稳健性。随后，我们提出了一种基于 LLM 的辅助自我审

[1] https://huggingface.co/OALL

议（ASD）机制用于自动评估。目标是为 AraTable 基准测试提供一个可靠、公正的、自动化的模型答案评估过程，使之与人类判断密切对齐。

这项工作为研究界提供了宝贵的资源，能够更有针对性地开发和评估用于阿拉伯语表格数据处理的大规模语言模型，并有助于推动阿拉伯语自然语言理解在结构化领域的进步。本研究的主要贡献可以总结如下：

- 构建一个专门设计用于评估大型语言模型在直接问答、事实验证和推理方面认知能力的阿拉伯语表格数据基准数据集。
- 开发一种混合基准生成和验证过程，通过该过程，LLMs 生成初始内容，然后由母语为阿拉伯语的专家严格过滤和验证，以确保高质量并促进对 LLM 性能的评估。
- 将来自不同来源的阿拉伯表格数据合并到一个数据集，以评估 LLM 在不同领域和阿拉伯表格来源的稳健性。
- 提出 ASD，一种采用自我评议机制的稳健自动评估方法，该方法与人类判断高度一致，以准确衡量 LLM 在 AraTable 上的表现。
- 分析和比较各种 LLM 在开发的基准数据集上的表现，识别它们在处理这些复杂任务时的优缺点。
- 提供一个公开可用的数据集，以促进阿拉伯表格数据处理的进一步研究和开发。

本文的其余部分组织如下：第 2 节描述了与本研究相关的主要工作。第 3 节展示了 AraTable 的构建细节。我们在第 **??** 节中介绍评估 AraTable 实用性的实验设置，介绍全面的评价方法在第 4 节，并在第 5 节中仔细分析和讨论结果。最后，我们在第 6 节中提出了研究的主要结论，并在第 **??** 节中概述了研究的一些局限性。

2 相关工作

鉴于使用大型语言模型（LLMs）进行阿拉伯语表格数据 QA 任务的新颖性，在本节中，我们将回顾主要相关工作，首先介绍与大型语言模型在阿拉伯语任务评估相关的研究，然后讨论其对表格数据理解的研究。最后，我们将讨论那些将大型语言模型作为自动评估工具进行部署的研究。

2.1 阿拉伯语中的 LLM 评估

对大型语言模型（LLM）在阿拉伯语中的评估是一个快速发展但复杂的领域，这主要受到阿拉伯语在全球的重要地位及其独特的语言复杂性驱动，包括丰富的形态学、多样的方言以及复杂的句法结构 [18] ，[19] 。尽管在开发阿拉伯语特定的 LLM 方面取得了显著进展，如 Jais [17] ，但综合评估仍然是一项挑战。现有的基准，如 AraBench [20] ，ALUE [21] ，ARLUE [22] ，及最近的 AraReasoner [23] ，评估了诸如情感分析和摘要等一般的自然语言处理任务。Abdallah et al. [24] 引入了 ArabicaQA，这是一个全面的数据集，旨在评估 LLM 在机器阅读理解和开放域问答中的表现。其他特定于阿拉伯语的关键评估方面仍未充分代表。例如，最近的研究已解决了幻觉和忠实性问题 [25] 。同样，公平性和偏见由 Alghamdi et al. [26] 进行了评估。此外，诸如 BALSAM [2] 和 aiXplain 阿拉伯语 LLM 基准报告 [3] 等全面的倡议旨在标准化多任务评估，强调文化理解。然而，据我们所知，以往的研究中没有包括结构化数据输入或评估 LLM 理解结构化数据的能力。鉴于我们的工作引入了一个名为 AraTable 的阿

[2] https://benchmarks.ksaa.gov.sa/b/balsam/tasks
[3] https://aixplain.com/arabic-llm-benchmark-report/

拉伯语 LLM 评估基准，该基准专门针对表格数据，它旨在通过解决有关结构化输入评估的显著缺口来补充这些先前的努力。

2.2 大规模语言模型与表格数据理解

大语言模型（LLMs）解释和推理结构化表格数据的能力，因其与非结构化文本相比具有独特的挑战性，而引起了越来越多的研究关注。Sui et al. [27] 引入了一个基准来评估 LLMs 对表的结构理解，包括诸如单元格查找和行检索等任务。他们表明输入格式显著影响性能，并提出了一种自增强提示策略以改善性能。同样，Liu et al. [28] 表明即使内容保持不变，LLMs 在处理表的结构变化时也表现出困难。为了应对这些限制，最近的努力已经超越了通用提示。例如，Ziqi and Lu [29] 引入了 Tab-CoT，这是一种在表格格式中组织中间推理步骤以指导对非结构化文本推理的提示策略。基于此，Wang et al. [30] 提出了 Chain-of-Table，这是一种通过迭代应用结构化操作（如过滤和聚合）来逐步转换表格以指导 LLM 推理的方法。模型不是直接生成最终答案，而是构建一系列中间表格修改以导出正确答案。表示模式是影响 LLM 在表格任务上性能的另一个重要因素。Deng et al. [31] 比较了文本线性化、JSON 和基于图像的表格输入，揭示了不同格式下的性能差异。

在从表格数据进行问答的角度评估大型语言模型方面，Chen [32] 调查了大型语言模型在使用少样本上下文学习进行表格推理的能力。他们的研究表明，结合链式思维提示的大型语言模型在表格问答和事实验证任务中表现良好。在相关的工作中，Grijalba et al. [33] 引入了 DataBench，一个覆盖了 65 个涵盖不同领域的现实世界数据集的基准，以支持大规模评估。他们的结果表明，尽管最近取得了一些进展，开源和闭源模型仍然面临挑战，即使是相对简单的问题也是如此。Bhandari et al. [34] 进一步分析了大型语言模型在表格问答任务中的鲁棒性，研究了上下文学习、模型规模、指令微调和领域偏见对多个数据集的影响。他们发现，指令微调可以提高性能，但数据污染和领域特定的复杂性仍然阻碍着模型的可靠性。最近，Xing et al. [13] 介绍了 MMTU，一个包含 25 个现实世界表格任务中超过 30,000 个问题的新基准。MMTU 旨在全面评估模型以专家级别理解、推理和处理表格的能力。使用 MMTU 的初步评估显示，即使是先进的大型语言模型，如 OpenAI 的 GPT-4o-mini 和 DeepSeek R1，准确率也仅达到了约 60 %，突显了在处理结构化数据问答时的显著改进空间。

尽管先前研究中强调了显著的进展，但这些研究的一致发现是，大规模语言模型在结构化表格数据上的表现仍远未达到最佳。此外，几乎所有现有评估都仅集中在英语语言上。据我们所知，以前的研究尚未对大规模语言模型在阿拉伯语环境下的表格问答和事实验证性能进行评估，这揭示了一个关键的研究空白，而本项研究正是要直接针对这点进行解决。

2.3 大型语言模型作为自动评价者

语言生成任务日益增长的复杂性和主观性导致了将 LLM 用作可扩展且具有成本效益的自动评估者的兴趣增加，有时被称为 “LLM 作为裁判” [35] 。基于 LLM 的评估系统已被 Li et al. [35] 分类为三种主要设置：(1) 单 LLM 系统，这种系统仅依赖一个 LLM 评估者，正如 [36] 和 [37] 的工作中所进行的；(2) 多 LLM 系统，结合多个 LLM 来执行评估任务 [38] ，[39] ；以及 (3) 混合系统，结合 LLM 与人类评估者 [40] [41] 。

在其他研究中，Lee et al. [42] 调查了大型语言模型（LLMs）在生成评价分数时的一致性，揭示了在不同提示和采样设置下显著的变化。顺着这一方向，Panickssery et al. [43] 研究了在 LLM 评价者中的自我偏好倾向，即倾向于偏爱自己的输出而非他

人的输出。作者确定了自我识别与自我偏好之间的相关性。微调实验表明这存在因果关系，这引起了对自我评价公平性的担忧。Zhang et al. [44] 探索了使用 LLM 来评估基于真实用户反馈的推荐解释文本的质量。他们提出了一种三级元评价策略，并表明零样本的 LLM 可以媲美或优于传统指标，如 BLEU 和 ROUGE。此外，他们还显示出，结合多个 LLM 或整合人类反馈可以提高评估的准确性和稳定性。

在这项工作中，我们提出了一种新颖的基于 LLM 的自动评估框架，该框架在 AraTable 基准上表现出与人类判断的强一致性。我们的方法称为辅助自我审议（ASD），基于一种自我审议理念，其中两个独立的 LLM 评估答案的正确性。然而，ASD 并不要求完全重新生成答案，而是在发生分歧的情况下激活，促使每个模型重新审视自己的决定，同时也考虑另一个模型为何可能得出不同结论。关键是，这一过程仅依赖于分歧信号，而不透露对方模型的推理，也不进行现有工作中常见的多轮辩论。这种新颖的轻量级策略在发生分歧时只需每位评委进行一次重新评估，每个模型使用共享标准为两种观点提供理由。

3 AraTable：在阿拉伯表格数据上对大语言模型进行基准测试

在本节中，我们介绍了 AraTable，这是一个用于评估 LLMs 在阿拉伯表格数据上的问答任务能力的基准。图 1 提供了基准构建流程的概览。

AraTable 的构建旨在支持多种推理类型，包括对阿拉伯表格进行算术、逻辑和比较推理。除了基于推理的问题外，AraTable 还包含直接答案提取和基于表格的事实验证任务。结合这三种任务能够全面评估模型的能力，从表面理解到更深层次的分析推理和验证。这种多任务结构与先前的工作相一致，例如 WikiTableQuestions [45]、TAT-QA [46] 和 TabFact [47]，这些工作共同展示了结合不同问题类型以评估推理深度和泛化能力的价值。以下各小节介绍了数据来源、QA 生成和人工验证的描述。

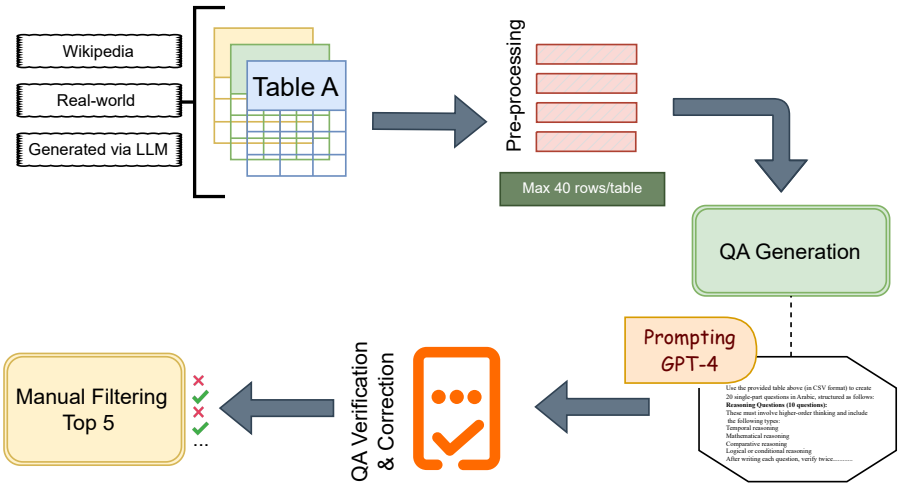


Fig. 1: 每项任务的 AraTable 构建过程概述

	Table Name	Rows	Columns
1	■■■■■■ ■■ ■■■■■ ■■■■■■ ■■■■■ ■■■■ ■■■■■	100	5
2	■■■■■■■ ■■■■■■■ ■■■■■	84	4
3	■■■■■■ ■■■■■ ■■■■■■■ ■■■■■	67	9
4	■■■■ ■■■■■ ■■■ ■■■■■■■■ ■■■■■	125	7
5	■■■■ ■■■ ■■■■■ ■■■■■ ■■■■■ ■■■ ■■■■■ ■■■■■ ■■■	192	4
6	■■■■■ ■■■■■■ ■■■■■■■■ ■■■■■	69	4
7	■■■■■ ■■■ ■■■■■■■ ■■■■■	188	9
8	■■■■■■ ■■■■ ■■■■■■■■ ■■■■■■■■ ■■■■■	422	7
9	■■■■■■■■■ ■■■■■■■ ■■■■■■■■■ ■■■■■ ■■■■■	57	10
10	■■■■■■ ■■■ ■■■■■ ■■■■■	163	5
11	■■■■ ■■■■ ■■■ ■■■■ ■■■ ■■■■■■ ■■■■■	100	5
12	■■■■■■■ ■■■■■ ■■ ■■■■■■■ ■■■■■■ ■■■■■ ■■■■■	112	6
13	■■■■■■ ■■ ■■■■■■■■ ■■■■ ■■■■■	92	6
14	■■■■■ ■■■■■	207	4
15	■■■■■■■ ■■■■■■■■ ■■■■■ ■■■■■	46	6
16	■■■■■■ ■■ ■■■■■■■ ■■■■ ■■■■■	43	8
17	■■■■■■ ■■■■■■■ ■■■■■■■■■ ■■■■■■■	121	13
18	■■■ ■■■ ■■ ■■■■■■ ■■ ■■■■■■■■ ■■■■ ■■■■■ ■■■ ■■■■■■■■■	88	9
19	■■■■■■ ■■ ■■■■■■■ ■■■■■ ■■■■	10	8
20	■■■■■■ ■■■■■■■ ■■■■■■■■■■ ■■■■■	25	3
21	■■■■■■ ■■■■ ■■■■■■■ ■■■■■■■■■ ■■■■■■	22	5
22	■■■■■■■ ■■■■■■■ ■■■■ ■■■■■	18	11
23	■■■■■■■ ■■■■■■■ ■■■■ ■■■■■	34	8
24	■■■■■■ ■■■ ■■■ ■■■■■■■ ■■■■■ ■■■■■	22	10
25	■■■■■■■ ■■■■■ ■■■■■■■■ ■■■■ ■■■■■	62	5
26	■■■■■■ ■■■■ ■■■ ■■■■■■ ■■■■■■ ■■■■ ■■■■■	50	8

Table 1: 本研究中考虑的维基百科阿拉伯语表格列表

3.1 表格数据收集

该基准旨在通过考虑来自旅游、交通、体育、商业、政治、教育、酒店、人口统计、娱乐和政府等多种领域的表格，实现广泛场景的全面评估。这种多样性确保了广泛的领域覆盖，从而便于在阿拉伯语上下文中评估不同的推理能力，并遵循最近的基准设计实践，即强调跨多个领域的覆盖，以支持全面的模型评估 [48] 。

这些表格来自多个平台，包括:

- 维基百科：从阿拉伯语维基百科页面提取的表格提供了跨各个领域的结构化知识的基础来源，并因其一致的格式和多样的主题而广泛用于表格问答基准测试 [45, 49] 。我们在这项研究中考虑的维基百科表格的完整列表，以及它们的维度（行数和列数），在表 1 中展示。
- 真实世界数据：表格来自于 Kaggle 这样的平台、国家开放数据门户和政府机构。这些表格涵盖了各种主题，包括酒店、房地产、汽车、文学、天气和公共事件。完整列表见表 2 。通过引入自然存在的结构、复杂的格式和特定领域的变异性，真实世界的表格补充了维基百科的内容。这与最近的基准测试如 TAT-QA [46] 和 TableEval [50] 相一致，这些基准测试中整合了来自真实世界来源的表格。
- LLM 生成的数据：为了确保涵盖在现有现实世界源码中没有充分代表的场景，使用 GPT-4o 基于结构化提示生成了额外的表格。近期研究表明，利用 LLM 生成的表格增强基准能够有效解决领域差距，并覆盖罕见或边缘案例场景 [51] 。这些补充表格提高了基准的全面性。表 3 展示了表格列表。

	Table Name	Rows	Columns	Source
1	HARD: Hotel Arabic-Reviews Dataset	93700	7	Elnagar et al.[1]
2	Saudi Arabia Used Cars Dataset	8248	14	Kaggle[2]
3	Jamalon Arabic Books Dataset	8986	11	Kaggle[3]
4	Saudi Arabia Real Estate (AQAR)	3718	24	Kaggle [4]
5	■■■■■■ ■■■■■	26	8	SAR [5]
6	■■■■■■■ ■■■■■■■ ■■■ ■■■■■ ■■■■■	14	6	Open Data Platform [6]
7	Events 2024	2637	10	Open Data Platform [6]
8	■■■■■■■■■	45	7	Open Data Platform [6]
9	NUMBER OF THUNDERSTORMS OBSERVED BY PME MET STATIONS-2006	29	13	National Center for Meteorology [7]
10	■■■ ■■■■■■ ■■■■■ ■■■■■ ■■■■ ■■■■■ ■■■■■■■■ ■■■■■■■	14	14	General Authority for Statistics [8]

Table 2: 本研究中考虑的真实世界阿拉伯语表格列表

3.2 表格预处理

鉴于表格来源的多样性，为确保一致性，所有表格均标准化为统一格式。原始表格名称得以保留。对于那些最初不是 CSV 格式或结构不兼容的表格，则进行了手动调整。每个表格还限制为最多 40 行，以适应模型输入大小的限制，特别是针对 Jais 模型的限制，而较短的表格则保持不变。这是在为 LLMs [53] 调整表格数据时的常见考虑。

[1] https://github.com/elnagara/HARD-Arabic-Dataset
[2] https://www.kaggle.com/datasets/turkibintalib/saudi-arabia-used-cars-dataset
[3] https://www.kaggle.com/datasets/dareenalharthi/jamalon-arabic-books-dataset
[4] https://www.kaggle.com/datasets/lama122/saudi-arabia-real-estate-aqar
[5] https://www.sar.com.sa/
[6] https://data.gov.sa/
[7] https://ncm.gov.sa/Ar/MediaCenter/OpenData/
[7] https://www.stats.gov.sa/statistics-tabs?tab=436312&category=1340120

	Table Name	Rows	Columns
1	■■■■■■ ■■■■	20	9
2	■■■■■■■ ■■■■■■■ ■■■■■	25	9
3	■■■■■ ■■■■■■	30	6
4	■■■■■ ■■■■■■ ■■■■	30	7
5	■■■■■■■■ ■■■■■	30	10

Table 3: 在本研究中考虑的由大型语言模型生成的阿拉伯语表格列表

3.3 问答生成

我们采用了一种基于提示的方法，使用 GPT-4 为数据集中的每个表格生成初始的一批 QA 对。QA 生成过程由三个不同的任务组成：直接回答问题、基于推理的问题和事实验证。对于每个任务，GPT-4 每个表格生成 10 个问题，总共为每个表格生成 30 个问题。每个任务的描述如下：

Question	Answer	Task	Reasoning Type
■■■■■■■ ■■■■■■■ ■■■■■ ■■ ١٩٦٠ ■■■■ ■■■■■■■ ■■■■■■■■■	■■■■	Simple QA	-
■■■■■■■ ■■■■■■■■■ ■■■■■■■ ■■■■■ ■■■■ ■■■ ■■■	■■■■■	Reasoning QA	Temporal reasoning
■■■ ■■■■■ ■■■■■■■ ■■■■■■■■■ ١٩٨٠ ■■■	■■■	Fact Verification	-

Table 4: AraTable 中的示例展示了三种问答任务：基于来自维基百科的表格进行简单问答、推理问答和事实验证。

直接问答该任务侧重于从表格中直接检索事实信息，而不需要推理。问题的设计旨在评估模型是否能够准确定位并提取表格内容中的特定值或条目。

推理问答推理问题需要通过多次推理步骤对表格数据进行更深刻的理解。这些问题涵盖各种类型的推理：时间推理，涉及解释和比较与时间相关的数据，例如日期或持续时间；数学推理，需要进行诸如总数和差值的计算；比较推理，侧重于比较条目；以及逻辑或条件推理，涉及评估多个领域的条件或限制。表 5 显示了数据集中每种推理类型出现的统计数据。用于生成直接问答和推理问答的提示提供于 Prompt 3.1 。

Prompt 3.1: Simple and Reasoning QA

Use the table provided above (in CSV format) to create 20 single-part questions in Arabic, structured as follows:
Reasoning Questions (10 questions): These must involve higher-order thinking and include the following types:
* Temporal reasoning
* Mathematical reasoning
* Comparative reasoning
* Logical or conditional reasoning
After writing each question, verify twice that it accurately reflects the intended reasoning type. If it does not, revise the question accordingly.

Non-reasoning Questions (10 questions): These should be straightforward, answerable directly from the table without inference or calculation.
Total Number of Questions: 20

Output Format: Save the output as a CSV file (UTF-8 with BOM encoding) with the following columns:
* Question: Written in Modern Standard Arabic
* Type of Question: Either "Reasoning" or "Non-Reasoning"
* Type of Reasoning: Leave this column blank for non-reasoning questions. For reasoning questions, specify the reasoning type (e.g., "Mathematical reasoning")
* Answer: Provide a short, specific answer derived directly from the table (use numbers only, without extra text, if applicable)
* Description: A brief explanation of how the answer was extracted from the table

Additional Notes: * All questions must be strictly based on information explicitly present in the table (in Arabic). Do not introduce any assumptions, inferred facts, or external knowledge.
* Begin with the 10 reasoning questions, followed by the 10 non-reasoning questions.
* Ensure effective utilisation of the table's structure for each question.

事实验证事实验证任务是根据表格内容评估给定陈述的真假。用来生成事实验证问答对的提示在 Prompt 3.2 中提供。

表 4 显示了来自阿拉伯维基百科奥林匹克主办城市表格的每个任务的一个问答对示例。该表格包含关于城市、国家、洲、年份、开始和结束日期、季节以及历届夏季奥林匹克运动会的信息。每个示例都标记了其对应的任务，以及在适用的情况下，推理类型。

Prompt 3.2: Fact Verification

Use the table provided above (in CSV format) to create 10 truly complex True/False questions that require deep reasoning, such as:
* Combining multiple conditions
* Cross-referencing multiple columns
* Logical deduction based on the provided data
Total Number of Questions: 10 (five with the answer "True" and five with the answer "False")

Output Format: Save the output in a CSV file (UTF-8 with BOM encoding) with the following columns:
* Question: A statement written in Modern Standard Arabic.
* Answer: "True" or "False".
* Description: A concise explanation of how the answer was derived from the table

Additional Notes: * Each statement must be based strictly on information explicitly present in the table (in Arabic). Do not introduce any assumptions, inferred facts, or external context.
* Ensure that the structure of the table is effectively utilised in all questions.

3.4 人工筛选和验证

生成后，所有问题和答案均经过人工审查，以确保清晰性、正确性和相关性。每个答案由三名人工标注者独立验证，结合代码、Excel 功能和人工检查，以确保与相应表格的一致性和事实对齐。如果有必要，问题和答案都进行了修正。对于每个任务，仅保留五个问题，选择这些问题以针对不同列和行的信息。这一过滤步骤确保最终集合是准确且具有代表性的。结果是每个表格包括总共 15 对问答。

3.5 数据集统计

最终的基准测试包含 41 个表格，每个表格配有 15 个问答对：5 个用于直接问答，5 个用于推理问答，5 个用于事实验证任务，总计 615 个人工验证的问答对。整个数据集中每种推理类型的问题数量如表 5 所示。我们的数据集包含 205 个推理问题，其中数学推理占据大多数，达到 79 个问题，几乎占总数的 39%。比较推理包含 67 个问题，逻辑推理有 42 个问题，而时间推理是最小的类型，仅占略高于 8%，有 17 个问题。这种均匀的任务分布允许在问题类型之间进行公平比较。该基准测试可供研究用途公开获取，网址为 [4] 。

我们采用以下实验设置来评估 LLM 在我们提出的基准上的表现。

我们专注于支持阿拉伯语的开源模型。我们选择了四个在阿拉伯语任务中表现优异的模型：Llama 3.3 70B [14] 、Mistral Large [15] 、DeepSeek-V3 [16] 和 Jais 70B [17] 。

[4] https://github.com/rana-alshaikh/AraTable-Benchmark

Reasoning Type	Number of Questions
Mathematical Reasoning	79
Comparative Reasoning	67
Logical Reasoning	42
Temporal Reasoning	17

Table 5: 每种推理类型的问题数量

3.6 零样本上下文学习

我们按照 Grijalba et al. [33] 方法给每个 LLM 提供了三个顺序部分：任务描述包含定义目标的指令：” 仅使用下表中的信息回答以下问题，并且在没有任何解释的情况下返回答案”。

其次，表格数据以带有标题和原始单元格值的 CSV 形式提供。通过以 CSV 格式提供表格，我们确保模型将其处理为结构化文本，而不需要额外的预处理。第三部分是问题集，一系列引用表格内容的问题。每个问题的写法都使得可以使用表格中的一个或多个单元格来找到答案。以这种方式格式化提示是为了尝试限制模型仅使用表格作为回答问题的信息来源，从而最大限度地减少模型内部的外部知识污染。这使我们能够在可再现的方式和一致、受控的条件下评估模型理解用阿拉伯语写成的表格的能力。

在我们的设定中，对答案的格式没有施加任何限制；模型可以自由地以纯文本、表格、列表或任何其他结构进行回答；因此，除了事实验证陈述被视为布尔问题外，任何提示中都没有添加类型后缀。对于这些问题，我们提示模型回答 `"True"` 或 `"False"`。选择不施加任何限制的原因有几个。首先，这有助于减少提示偏差：通过不要求特定的输出结构，我们避免了将模型强制导向特定的表示方式，这使我们能够捕捉它们组织和呈现输出的自然偏好 [54] 。支持这一观点的是，[55] 发现在推理任务中，我们通常可以通过放松格式限制来达到更好的表现。

此外，由于我们的研究仅关注于开源模型，必须强调的是，相比于封闭源代码模型，这些模型通常不太遵循严格的输出格式，正如 [56] ，指出对开源模型施加格式限制通常无法产生预期的优势，从而支持我们的选择。

最后，这种方法反映了实际部署场景，其中用户通常希望获得人类可读的响应，而不是精确的机器可读输出。通过允许模型选择其偏好的响应格式，我们获得了关于它们在灵活和现实环境中提供准确和有效响应能力的洞察 [57] 。这些因素确保我们的评估专注于模型理解阿拉伯表格数据的能力，而不是它们对任意格式限制的行为遵从性。即使模型被指示简短回答，Jais 的回答却往往冗长，充满了不必要的重复和解释。因此，对于 Jais，我们采用了两种收集方法：（1）保存原始的、未修改的输出，以及（2）让两位独立的人类注释者从冗长的回答中提取简明的答案。模型输出以 CSV 格式保存，以供稍后评估。

4 评估

允许模型自由回答而不遵循任何结构，在评估阶段带来了挑战，因为这些模型在冗长性和响应形式上有显著差异：一些模型按照指示给出非常简短、简明的答案，而另一

些则使用问题中单词的各种同义词提供过于详细的解释。例如，模型并没有直接回答“True”，而是使用了诸如“Indeed ■■■■■■ ”或“这确实是一个事实 ■■■■■ ■■■ ”之类的短语。

此外，我们观察到，这些模型中的一些对问题列表、编号和分组的细微变动更加敏感。举例来说，我们发现如果问题没有编号，Jais 和 Mistral 可能会给出无序的答案，甚至遗漏其中一个问题。此外，Jais 偶尔会基于触发词过滤响应或者在回答中途转换为英语，特别是在实体名称中，这可能表明模型并不完全依赖表格中提供的信息来回答问题。另一个挑战是，这些模型经常表现出不一致的标点符号、附加符及文本对齐方式的规范化。

根据这些观察，非约束设置以及由此产生的挑战和不一致表明，使用自动评分方法（如精确匹配或嵌入相似度）可能并不适合我们的设置。即使在我们可以对模型答案施加严格限制的情况下，这些方法也只是部分测试模型遵循给定指令的程度，而不是完全专注于模型对表格数据的理解。因此，在我们的评估框架中，我们采用了一个较为宽松的评估方法，使我们能够专注于衡量模型理解阿拉伯表格数据的能力。
我们设计了一个评估框架，用于评估四个模型的响应，作为整体，一共是五个不同的输出，因为 Jais 有两个版本的响应：一个总结答案和一个完整答案。在评估阶段，所有模型都通过分配数字标识符（例如，模型 1，模型 2）进行了匿名化处理。此外，完整的 Jais 输出及其由人类提取的答案被视为具有不同标识符的独立实体，分别分配为模型 4 和模型 5。这种设计使得（a）评估 Jais 冗长输出中人类提取回应的准确性，并在以简明（模型 4）和冗长（模型 5）形式呈现相同答案时比较人类评估的一致性——理想情况下，两者应产生相同的准确性，因为核心答案未改变；(b) 评估我们的自动评估方法在理解和评估表现出冗长性的 LLM 响应方面的鲁棒性。图 2 提供了评估框架管道的概览。为了建立一个可靠的基准，九名母语为阿拉伯语的人类评估者被分配为评判者。这些评估者被组织成三个小组，以连续三轮的形式评估模型的答案，并使用详细的评分标准来判断正确性。第一组独立地将每个 LLM 生成的答案与真实标注进行评估。第二组随后复审评估结果，并在第三轮中，第三组复审评估并生成最终评估结果。

在使用的评分标准中，其中一条规则是如果答案传达的意义与标准答案相同，那么就被视为正确，即不需要与标准答案完全匹配。这一指导方针允许措辞的变化或描述性词语或单位的添加。例如，“大约 40”，“40”，和“四十年”都被视为等同。在布尔型答案的事实验证问题中，评委被指示接受如 [True , Yes , ¿” ■■■ “ ” ¿■■ “ ” ■■■■ ■■■■ “] 表示的真，以及如 [False , No , ” ¿■■■■ ■■■ “ ” ¿■■■ “ ” ■■■■ ■■■■ “] 表示的假。这一规则也允许接受长句作为真或假，只要其肯定或否定含义明确。关于数值精度比较的指导方针是，在 ±0.005 公差范围内相符的数值

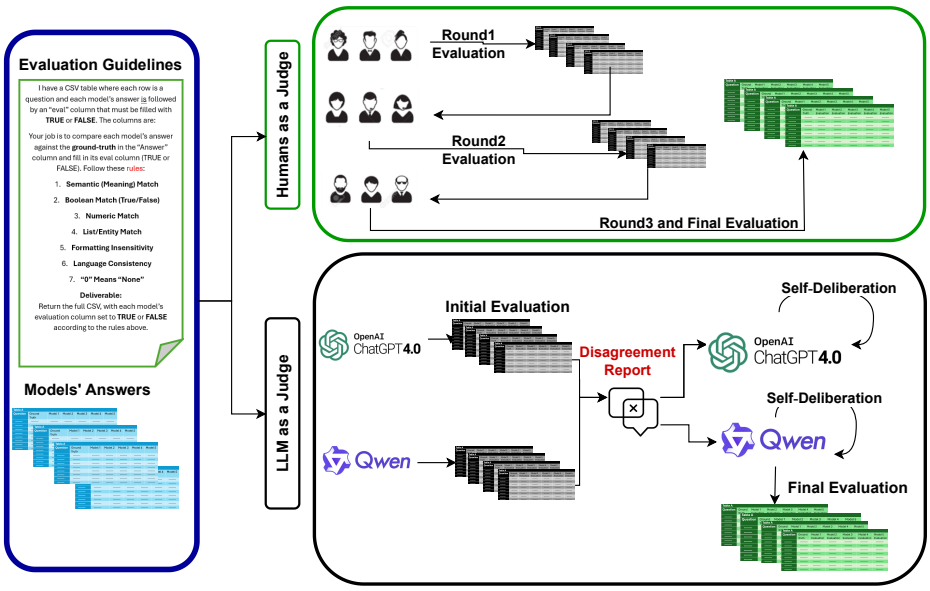


Fig. 2: 使用人工评审和 LLM 评审对 AraTable 进行综合评估框架的示意图。

被视为正确；例如，“1.883”可接受为“1.889”。在数值格式化方面，允许有变化，例如“25 %”等同于“0.25”，数值可以带或不带单位（例如，“38000”，“38 ■■■ ”）。如果标准答案是一个列表，比如酒店名称，评委认为如果两个列表包含相同的项目，则它们是等同的，而不考虑这些列表的顺序、标点符号或语言格式；例如，“■■■■ ■ ■■■■ ■ ■ ”被视为与 [■ ■■■■ ، ■ ■■■■] 相同。此外，如果正确答案包含阿拉伯实体名称，例如大学、城市或街道，则只有用阿拉伯语给出的答案才被视为正确，因为用不同语言回答可能表明模型并未依赖表格作为唯一的信息来源；布尔值或数值答案则不受此规则限制。评分标准还明确说明，标点符号的任何变化，例如引号、括号和逗号；大小写或空格的差异；风格上的细微变化；以及轻微的拼写错误都应被忽略。所有这些指导方针旨在确保我们的评估重点在于评估模型理解阿拉伯表格式数据的能力，而不是简单的格式差异。最后，评审员在每个答案的评估栏中记录其评估结果为“真”或“假”。

作为自动化评估的一部分，使用了两个大型语言模型作为评审——Qwen [58] 和 4O [59] ——它们均未见过原始问题或用于生成答案。这些评审模型被提供了与人工评审一致的事实依据和评判标准，并在两个阶段中对模型的答案进行评估。起初，每个评审分别独立于另一个评审对模型的答案进行评估，并对每个回应判定为真或假。

减少分歧的辅助自我审议（ASD）：为了增加大语言模型（LLMs）作为评判者之间的一致性，并减少与人类评估的分歧，在第二轮中，我们引入了自我审议机制。在这一机制中，(1) 我们为每位评判者匿名，分别命名为 Judge1（4O）和 Judge2（Qwen）；然后（2）我们使用他们的初次评估生成了一份分歧报告，突出显示了他们与另一位评判者的评估不一致的案例。(3) 我们然后将这份报告部分地给予每位评判者，每次三个冲突案例，每个案例包括模型评估和另一位评判者的评估。我们提示每位评判者

（模型）根据评分标准为自己的评估辩护，并通过查找支持或反对每个评估的规则来为另一位评判者的评估辩护。换句话说，另一位的标签充当了失败标记的角色，鼓励反思性的自我批评，因为我们的自我审议协议要求每位评判者（4O 和 Qwen）只在与另一位评判者的标签冲突时才重新审视自己的初始真/假决定，而从不查看另一位评判者的推理。此 ASD 机制使我们能够避免协调成本，并减少可能因两大语言模型交叉审议循环 [60, 61] 而产生的延迟，同时鼓励模型进行自我批评。

4O Self-Deliberation	Qwen Self-Deliberation
Disagreement Analysis (Case 1) [illegible] • 真实值: 4.51 % • 模型答案：4.50 % • 你原来的评估：FALSE • 其他法官：[法官 2 = TRUE]	Disagreement Analysis (Case 1) [illegible] • 真实值: 4.51 % • 模型答案：4.50 % • 您的原始评估：TRUE • 其他法官：[法官 1 = FALSE]
Conflicts with Other Judges Judge 2 marked the answer as TRUE, but the difference is 0.01, which exceeds the ±0.005 tolerance. According to Rule 3 this must be marked as FALSE.	Conflicts with Other Judges [illegible] ٣ [illegible] [illegible] [illegible] [illegible]: [illegible] [illegible]“FALSE”.
Rules Supporting My Evaluation Rule 3 (Numeric Match ± 0.005): 4.50 % is outside the tolerance; match fails.	Conflicts Against My Evaluation [illegible] [illegible] [illegible] ±0.005 [illegible] .[illegible] [illegible] 0.01
Final Decision: Maintain Justification: Although 4.50 % is close to 4.51 % , it lies outside the ±0.005 tolerance, so the answer is FALSE.	Final Decision: Revise Justification: [illegible] [illegible] (±0.005) [illegible] [illegible] [illegible] [illegible] “FALSE”.

Table 6: 4O 和 Qwen 关于数值精度分歧的自我审议输出示例。

5 结果与讨论

5.1 针对 AraTable QA 的 LLMs 表现分析

本节呈现了我们对各种大规模语言模型在理解和提取阿拉伯表格数据方面能力的实证结果，涉及三个问答任务：直接问答、推理和事实验证。具体而言，我们评估了 Llama 3.3 70B、Mistral Large、DeepSeek-V3 和 Jais 70B 在来自三个不同来源的 41 个阿拉伯表格数据集上的表现，如第 3.1 节所述。性能的评估基于这些大型语言模型的答案

与基于人类判断的黄金标准真实答案之间的准确性换句话说，人类手动评估这些大型语言模型的答案是否正确，如第 ?? 节所述。对于不同模型和问题类型的准确性指标，针对维基百科、真实世界和 LLM 生成的数据源，我们在表 7 至 9 中进行了报告。

在所有数据集和问题类型中，模型性能显示出明显的层级。DeepSeek-V3 始终表现出最高的准确率，紧随其后的是 Llama 3.3 70B 和 Mistral Large。Jais 模型的性能显著低于其他模型，特别是在应对更复杂的问题类型时经常表现不佳。Jais 的简明答案和完整答案之间相同的表现表明，完整答案在所有数据集中并没有在准确性上提供明显的优势。所有模型和数据表来源中普遍的模式是对直接问答题的较高表现，表明直接从表中检索信息通常比需要推理能力的任务更容易。

深入查看每个表格，表格 7 展示了在维基百科表格上的表现，显示 DeepSeek-V3 在所有问题类别中领先，直接问答中达到了 96.15 %，推理中为 59.23 %，事实验证中为 81.54 %。Llama 3.3 70B 和 Mistral Large 也表现强劲，尤其是在直接问答问题上。Jais 和 Jais 完整答案模型显示出最低的准确率，推理为 30 %，直接问答为 63.85 %，表明处理维基百科风格的表格数据存在挑战。表格 8 展示了现实世界表格的结果，这些表格通常比维基百科表格更为异质且结构较少。DeepSeek 保持领先，直接问答问题的准确率为 98 %，在事实验证上表现良好，准确率为 80 %，但在推理问题上准确率下降到 48 %。Llama 在该数据集上的推理准确率显著下降至 20 %。Jais 模型继续表现不佳，所有问题类型的准确率都很低，特别是推理问题，仅为 14 %。对于大多数模型而言，现实世界的表格在推理任务上更具挑战性。表格 9 展示了由 LLM 生成的表格结果，提供了有趣的见解。DeepSeek-V3 在直接问答问题上取得了完美的 100 % 分数，显示出从合成表格中提取直接信息的卓越能力。DeepSeek-V3 和 Mistral Large 在推理问题上表现相同（48 %），这相较于现实世界的数据对 Mistral Large 来说是显著进步。Llama 3.3 70B 在直接问答和事实验证上的表现也有所改善。Jais 模型在直接问答的准确率略有提升（56 %），但仍然显著落后于其他模型。

最引人注目的观察是直接问答问题与其他问题类型之间的一贯差距。直接问答问题主要涉及直接提取信息，始终表现出较高的准确率（常常超过顶级模型的 90 %）。相比之下，推理问题始终构成最大的挑战，即使对于表现最好的模型，准确率通常也低于 60 %。这突显了当前大型语言模型在使用表格数据进行复杂逻辑推理时的一个基本局限性，特别是在数据不是完美结构化或包含微妙差异时。事实验证问题则介于两者之间，这表明它们需要比简单提取更深层次的理解，但推理步骤不如推理题复杂。此外，DeepSeek 显著领先，其后紧跟的是 Llama 3.3 70B 和 Mistral Large，这暗示着比起仅仅架构调整，较大的 token 预算和更多样化的预训练语料库（尤其那些包含结构化数据和阿拉伯语来源的语料库）更为重要。尽管 Jais 70B 以阿拉伯语为中心，但在推理方面表现不佳，表明仅仅包含阿拉伯语是不够的；在预训练/微调期间对表格推理模式的接触可能是必需的。这些发现为当前大型语言模型在阿拉伯语表格数据中的表现提供了有价值的见解，并指出了一些未来研究和发展的关键领域，特别是提高推理能力和在多样化现实世界数据集中的稳健性。

5.2 大型语言模型作为评判者：自动评估性能分析

本节分析了大型语言模型（LLM）的性能，特别是 Qwen 和 4O 作为自动评估器（或裁判）来评估其他 LLM 性能时的表现，正如在第 ?? 节中解释的那样。通过考察的所有数据集中表 10 到 12 ，将它们的准确性与人工评估员进行了比较，包括维基百科、真实世界和 LLM 生成的数据集。评估分析了判断准确性在审议阶段之前和之后的表现，重点关注相对于人工判断基准的准确性签名差距。

Question Type	Llama70B	DeepSeek-v3	Mistral-L	Jais70B	Jais Full answer
Direct QA	0.90	0.96	0.92	0.63	0.63
Fact Verification	0.75	0.81	0.70	0.53	0.53
Reasoning	0.45	0.59	0.53	0.30	0.30
Overall	0.71	0.79	0.72	0.49	0.49

Table 7: 每种模型在维基百科表格的问答类型中的准确度

Question Type	Llama70B	DeepSeek-v3	Mistral-L	Jais70B	Jais Full answer
Direct QA	0.86	0.98	0.90	0.40	0.40
Fact Verification	0.74	0.80	0.72	0.38	0.38
Reasoning	0.20	0.48	0.30	0.14	0.14
Overall	0.60	0.75	0.64	0.31	0.31

Table 8: 所考虑的真实世界表格中按问题类型的每个模型的准确性

Question Type	Llama70B	DeepSeek-v3	Mistral-L	Jais70B	Jais Full answer
Direct QA	0.92	1.00	0.96	0.56	0.56
Fact Verification	0.76	0.76	0.60	0.60	0.60
Reasoning	0.32	0.48	0.48	0.20	0.20
Overall	0.67	0.75	0.68	0.45	0.45

Table 9: LLM 生成的表格中每种问题类型的模型准确性

在讨论之前，Qwen 的表现与评估高性能 LLMs（如 Llama 3.3 70B（例如，$\Delta = -0.01$，针对真实世界））、DeepSeek-V3 和 Mistral Large 时的人类评估者非常接近，各个数据集中的精度差异非常小。然而，对于 Jais70B 和 Jais-Full 等模型，Qwen 往往表现出轻微的正偏差，这意味着其倾向于比人类评估者更高地评估这些模型（例如，Wikipedia Jais-Full $\Delta = +0.03$，真实世界 Jais70B $\Delta = +0.07$）。在 LLM 生成的数据集中，Qwen 对某些模型的评分甚至高于人类评审（例如，Mistral Large $\Delta = +0.04$，Jais70B $\Delta = +0.08$）。相比之下，4O 模型在几乎所有目标 LLMs 和数据集中的一致性低于人类评估者。它在准确度上表现出显著的负面差距（例如，Wikipedia Jais-Full $\Delta = -0.15$，真实世界 Jais-Full $\Delta = -0.16$，真实世界 DeepSeek $\Delta = -0.07$）。这表明在最初的、未讨论阶段，4O 的判断往往不如人类评估准确，和人类的评估结果一致性较差。

在讨论阶段之后，观察到 Qwen 作为大语言模型评判者的表现有显著提高。其判断准确性在所有数据集（维基百科、真实世界和大语言模型生成）上稳定地与人类基线趋同，经常与各种目标大语言模型（包括 Llama70B、DeepSeek-V3、Mistral Large、Jais70B 和 Jais-Full）的判断完全一致（例如，真实世界 Llama70B $\Delta = +0.00$，维基

百科 Mistral Large Δ =+0.00）。这种转变效果凸显了讨论机制在优化 Qwen 内部评估标准方面的有效性，使其能够准确反映人类评估。例如，虽然 Qwen 有时在讨论前表现出一些轻微的差异（例如对真实世界数据集中的 Jais 模型有正偏差），但这些差异在讨论后被一致消除，使其自动判断完全符合人类标准。相反，4O 的进步更为有限；尽管显示出一些提升，与人类评估者的对齐度与 Qwen 相比仍然不太一致。它经常保留负差距，例如在维基百科（$\Delta = -0.05$）和真实世界（$\Delta = -0.05$）数据集上的 Jais-Full，表明 4O 要么从讨论过程中受益较少，要么其固有偏差更难以纠正。图 3 显示了在所有三个数据集（a）维基百科，（b）真实世界，和（c）大语言模型生成的问题中的大语言模型评估者与人类基线之间的平均无符号误差（平均 $|\Delta|$），无论是在 ASD 步骤之前（蓝色多边形）还是之后（红色多边形）。自我审议一致且显著地减少了所有模型中 LLM 评估者与人工评估者之间的准确性差距，这可以从每个子图中红色形状很好地嵌套在蓝色形状内得以证实。减少最明显体现在 Jais 的完整答案中，这表明当 LLM 评委处理存在冗长问题的模型时，自我审议最为重要。总的来说，该图证实了 ASD 大约将与人工判断的平均差异减半。

总之，审议阶段被证明是增强作为裁判的 LLM 可靠性和人类对齐性的关键组成部分。在表格 6 中，我们展示了一个案例，其中每个模型在 ASD 阶段后决定修改或保持其评估。尤其是 Qwen，与人类基线相比，在审议后表现出接近完美的表现，突显了其作为高度准确的自动化评估器的强大潜力，显著地推进了使用 AI 进行稳健评估的可行性。虽然 4O 的结果表明审议的有效性因特定 LLM 裁判而有一些可变性，但此阶段的整体影响代表了开发更可靠和人类对齐的自动评估方法论的一大进展。即使我们将最终决定分配给一个模型，咨询第二个模型仍然有价值，因为它的替代输出鼓励多样化的推理，并帮助主要模型通过 ASD 机制进行更批判性地思考。

	Accuracy (↑)					Δ vs Human				
Evaluator	Llama	DeepSeek	Mistral	Jais	Jais-Full	Δ_{L70B}	Δ_{DS}	Δ_{Mis}	Δ_{J70B}	$\Delta_{\text{J-F}}$
Human Raters	0.71	0.79	0.72	0.49	0.49	+0.00	+0.00	+0.00	+0.00	+0.00
Qwen before ASD	0.69	0.77	0.71	0.51	0.52	-0.02	-0.02	-0.01	+0.02	+0.03
Qwen after ASD	0.71	0.79	0.72	0.49	0.49	+0.00	+0.0	+0.00	+0.00	+0.00
4O before ASD	0.69	0.75	0.69	0.46	0.34	-0.02	-0.04	-0.03	-0.03	-0.15
4O after ASD	0.69	0.79	0.72	0.49	0.44	-0.02	+0.00	+0.00	+0.00	-0.05

Table 10: 维基百科数据集的准确性以及在辅助自我审议（ASD）前后相对于人类基准的签名差距。

6 结论与未来工作

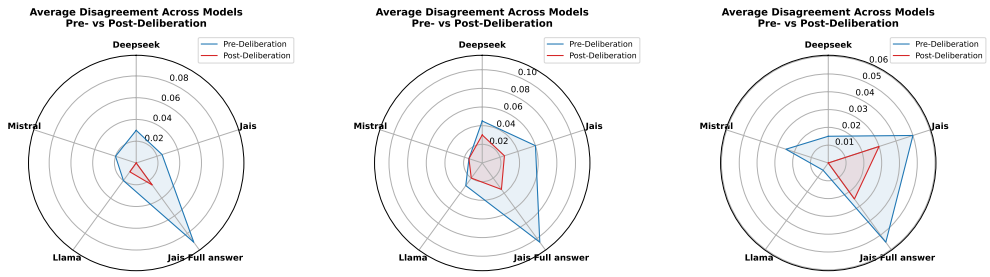
在本文中，我们介绍了 AraTable，这是一个用于阿拉伯语表格数据的问答基准。该基准提供了一个评估 LLMs 在三种不同 QA 任务下表现的框架：直接提取、推理和事实验证。每种问题类型的设计旨在探测模型理解和认知推理能力的不同方面。我们的评估突出显示，尽管一些模型在表面任务上表现良好，但认知推理和验证仍然具有挑战性，揭示了阿拉伯表格理解中的重要差距。尽管本文专注于零样本 QA 评估，AraTable 也可以作为微调或调整模型以适应阿拉伯语表格推理任务的基础。

	Accuracy (↑)					Δ vs Human				
Evaluator	Llama	DeepSeek	Mistral	Jais	Jais-Full	Δ_{L70B}	Δ_{DS}	Δ_{Mis}	Δ_{J70B}	$\Delta_{J\text{-}F}$
Human	0.60	0.75	0.64	0.31	0.31	+0.00	+0.00	+0.00	+0.00	+0.00
Qwen before ASD	0.59	0.73	0.64	0.38	0.36	-0.01	-0.02	+0.00	+0.07	+0.05
Qwen after ASD	0.60	0.75	0.64	0.31	0.31	+0.00	+0.00	+0.00	+0.00	+0.00
4O before ASD	0.55	0.68	0.61	0.26	0.15	-0.05	-0.07	-0.03	-0.05	-0.16
4O after ASD	0.58	0.75	0.64	0.31	0.26	-0.02	+0.00	+0.00	+0.00	-0.05

Table 11: 在进行辅助自我审议 (ASD) 之前和之后报告的真实世界数据集准确性和与人类基线的符号差距。

	Accuracy (↑)					Δ vs Human				
Evaluator	Llama	DeepSeek	Mistral	Jais	Jais-Full	Δ_{L70B}	Δ_{DS}	Δ_{Mis}	Δ_{J70B}	$\Delta_{J\text{-}F}$
Human Raters	0.67	0.75	0.68	0.45	0.45	+0.00	+0.00	+0.00	+0.00	+0.00
Qwen befor ASD	0.68	0.76	0.72	0.53	0.52	+0.01	+0.01	+0.04	+0.08	+0.07
Qwen after ASD	0.67	0.75	0.68	0.49	0.48	+0.00	+0.00	+0.00	+0.04	+0.03
4O before ASD	0.67	0.73	0.69	0.47	0.41	+0.00	-0.02	+0.01	+0.02	-0.04
4O after ASD	0.67	0.75	0.68	0.47	0.47	+0.00	+0.00	+0.00	+0.02	+0.02

Table 12: LLM 生成的数据集准确性和与人类基线的签名差距，报告于辅助自我审议（ASD）之前和之后。



(a) 维基百科数据集　(b) 真实世界数据集　(c) LLM 生成的数据集

Fig. 3: 审议阶段对减少人为评估者与 LLMs 评估者之间分歧的影响。

基于此基础，未来的工作将探讨阿拉伯语表格数据中的少样本和微调设置，特别是针对复杂推理场景。将基准扩展到包含更大和更多样的表格，具有不同的格式，如多表或层次结构，探索处理此类数据的高效方法，包括检索增强生成，是另一个有前景的方向。我们希望这项工作能鼓励未来在大语言模型中推动阿拉伯语表格理解的进步。

我们的评估存在若干局限性。首先，我们仅测试了数量有限的模型，专注于支持阿拉伯语的开源模型。虽然这些模型在阿拉伯语基准的各种任务中表现出色，但可能有些模型在使用表格数据进行这种特定类型的推理时更为优化。其次，在我们的评估中，LLM 仅在零样本环境中进行评估。通过提供更多示例或少样本环境，模型的性能可能

会得到提升，这将是我们未来研究的方向。第三，我们的评估主要集中在小型表格上，每个表格仅包含 40 行。这种集中重点主要是由于某些模型在处理大数据量时的限制。

AraTable 中包含的所有数据均来自于公开可用来源，如阿拉伯语维基百科、政府门户网站和开放数据平台，均在非限制性许可下收集。我们保留了原始表格内容而不进行修改，并相应地注明了每个来源（见表 1 & 2 ）。我们尤其注意排除了包含攻击性、敏感或有争议内容的表格或生成的问答对。虽然在问答生成和评估阶段使用的大型语言模型可能存在偏见，但所有模型输出均经过人工审核，以确保发布的基准中不存在有害、歧视或不当内容。

References

[1] Li, X., Feng, X., Hu, S., Wu, M., Zhang, D., Zhang, J., Huang, K.: Dtllm-vlt: Diverse text generation for visual language tracking based on llm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7283–7292 (2024)

[2] Liu, X., Sun, T., Xu, T., Wu, F., Wang, C., Wang, X., Gao, J.: Shield: Evaluation and defense strategies for copyright compliance in llm text generation. arXiv preprint arXiv:2406.12975 (2024)

[3] Koshkin, R., Sudoh, K., Nakamura, S.: Transllama: Llm-based simultaneous translation system. arXiv preprint arXiv:2402.04636 (2024)

[4] Wang, L., Du, Z., Jiao, W., Lyu, C., Pang, J., Cui, L., Song, K., Wong, D., Shi, S., Tu, Z.: Benchmarking and improving long-text translation with large language models. In: Findings of the Association for Computational Linguistics ACL 2024, pp. 7175–7187 (2024)

[5] Allemang, D., Sequeda, J.: Increasing the llm accuracy for question answering: Ontologies to the rescue! arXiv preprint arXiv:2405.11706 (2024)

[6] Bui, T., Tran, O., Nguyen, P., Ho, B., Nguyen, L., Bui, T., Quan, T.: Cross-data knowledge graph construction for llm-enabled educational question-answering system: a case study at hcmut. In: Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia, pp. 36–43 (2024)

[7] O' Brien Quinn, H., Sedky, M., Francis, J., Streeton, M.: Literature review of explainable tabular data analysis. Electronics **13**(19), 3806 (2024)

[8] Nguyen, G., Brugere, I., Sharma, S., Kariyappa, S., Nguyen, A.T., Lecue, F.: Interpretable llm-based table question answering. arXiv **abs/2412.12386** (2024)

[9] Zhang, Z., Lin, X.V., Agrawal, R., Riedewald, M., Zhang, J.: Normtab: Improving symbolic reasoning in llms through tabular data normalization. In: Findings of the Association for Computational Linguistics: EMNLP 2024 (2024). https://api.semanticscholar.org/CorpusID:272645535

[10] Wang, Y., Zhang, S., Lu, Y., Wang, J., Wang, S., Zhang, X., Li, H.: Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. arXiv **abs/2403.19318** (2025)

[11] Wu, X., Liang, D., Li, Z.: Tablebench: A comprehensive and complex benchmark for table question answering. arXiv **abs/2408.09174** (2024)

[12] Grijalba, J.O., Ureña-López, L.A., Camacho-Collados, J., Cámara, E.M.: Question answering over tabular data with databench: A large-scale empirical evaluation of llms. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 13471–13488 (2024). https://api.semanticscholar.org/CorpusID:266791020

[13] Xing, J., He, Y., Zhou, M., Dong, H., Han, S., Chen, L., Zhang, D., Chaudhuri, S., Jagadish, H.: Mmtu: A massive multi-task table understanding and reasoning benchmark. arXiv preprint arXiv:2506.05587 (2025)

[14] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

[15] AI, M.: Mistral-Large (Version 1.0) [Large language model]. Hugging Face, https://huggingface.co/mistralai/Mistral-Large (2024)

[16] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)

[17] Sengupta, N., Sahu, S.K., Jia, B., Katipomu, S., Li, H., Koto, F., Marshall, W., Gosal, G., Liu, C., Chen, Z., et al.: Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. arXiv preprint arXiv:2308.16149 (2023)

[18] Rhel, H., Roussinov, D.: Large language models and arabic content: A review. arXiv preprint arXiv:2505.08004 (2025)

[19] Mousi, S., al.: Aradice: Benchmarks for dialectal and cultural capabilities in llms. ACL Anthology (2025)

[20] Sajjad, H., al.: Arabench: A multi-task benchmark for arabic natural language processing. Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC) (2020)

[21] Seelawi, H., Tuffaha, I., Gzawi, M., Farhan, W., Talafha, B., Badawi, R., Sober, Z., Al-Dweik, O., Freihat, A.A., Al-Natsheh, H.: ALUE: Arabic language understanding evaluation. In: Habash, N., Bouamor, H., Hajj, H., Magdy,

W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., Touileb, S. (eds.) Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 173–184. Association for Computational Linguistics, Kyiv, Ukraine (Virtual) (2021). https://aclanthology.org/2021.wanlp-1.18/

[22] Abdul-Mageed, M., Elfardy, H., Diab, M.: Arlue: An arabic representation learning benchmark for universal evaluation. arXiv preprint arXiv:2109.00845 (2021)

[23] Hasanaath, A., Alansari, A., Ashraf, A., Salmane, C., Luqman, H., Ezzini, S.: Arareasoner: Evaluating reasoning-based llms for arabic nlp. arXiv preprint arXiv:2506.08768 (2025)

[24] Abdallah, A., Kasem, M., Abdalla, M., Mahmoud, M., Elkasaby, M., Elbendary, Y., Jatowt, A.: Arabicaqa: A comprehensive dataset for arabic question answering. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2049–2059 (2024)

[25] Abdaljalil, S., Kurban, H., Serpedin, E.: Halluverse25: Fine-grained multilingual benchmark dataset for llm hallucinations. arXiv preprint arXiv:2503.07833 (2025)

[26] Alghamdi, E.A., Masoud, R., Alnuhait, D., Alomairi, A.Y., Ashraf, A., Zaytoon, M.: Aratrust: An evaluation of trustworthiness for llms in arabic. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 8664–8679 (2025)

[27] Sui, Y., Zhou, M., Zhou, M., Han, S., Zhang, D.: Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, pp. 645–654 (2024)

[28] Liu, T., Wang, F., Chen, M.: Rethinking tabular data understanding with large language models. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 450–482. Association for Computational Linguistics, Mexico City, Mexico (2024). https://doi.org/10.18653/v1/2024.naacl-long.26 . https://aclanthology.org/2024.naacl-long.26/

[29] Ziqi, J., Lu, W.: Tab-cot: Zero-shot tabular chain of thought. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 10259–10277 (2023)

[30] Wang, Z., Zhang, H., Li, C.-L., Eisenschlos, J.M., Perot, V., Wang, Z., Miculicich, L., Fujii, Y., Shang, J., Lee, C.-Y., et al.: Chain-of-table: Evolving tables in the reasoning chain for table understanding. arXiv preprint arXiv:2401.04398 (2024)

[31] Deng, N., Sun, Z., He, R., Sikka, A., Chen, Y., Ma, L., Zhang, Y., Mihalcea, R.: Tables as texts or images: Evaluating the table reasoning ability of llms and mllms. In: Findings of the Association for Computational Linguistics ACL 2024, pp. 407–426 (2024)

[32] Chen, W.: Large language models are few(1)-shot table reasoners. In: Vlachos, A., Augenstein, I. (eds.) Findings of the Association for Computational Linguistics: EACL 2023, pp. 1120–1130. Association for Computational Linguistics, Dubrovnik, Croatia (2023). https://doi.org/10.18653/v1/2023.findings-eacl.83 . https://aclanthology.org/2023.findings-eacl.83/

[33] Grijalba, J.O., Lopez, L.A.U., Martínez-Cámara, E., Camacho-Collados, J.: Question answering over tabular data with databench: A large-scale empirical evaluation of llms. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 13471–13488 (2024)

[34] Bhandari, K.R., Xing, S., Dan, S., Gao, J.: On the robustness of language models for tabular question answering. arXiv preprint arXiv:2406.12719 (2024)

[35] Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y.: Llms-as-judges: a comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv:2412.05579 (2024)

[36] Lin, Y.-T., Chen, Y.-N.: Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. arXiv preprint arXiv:2305.13711 (2023)

[37] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634 (2023)

[38] Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., Liu, Z.: Chateval: Towards better llm-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201 (2023)

[39] Chu, Z., Ai, Q., Tu, Y., Li, H., Liu, Y.: Pre: A peer review based large language model evaluator. arXiv preprint arXiv:2401.15641 (2024)

[40] Li, Q., Cui, L., Kong, L., Bi, W.: Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. arXiv e-prints, 2310 (2023)

[41] Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., Ma, X.: Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pp. 1–23 (2025)

[42] Lee, N., Hong, J., Thorne, J.: Evaluating the consistency of llm evaluators. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 10650–10659 (2025)

[43] Panickssery, A., Bowman, S., Feng, S.: Llm evaluators recognize and favor their own generations. Advances in Neural Information Processing Systems **37**, 68772–68802 (2024)

[44] Zhang, X., Li, Y., Wang, J., Sun, B., Ma, W., Sun, P., Zhang, M.: Large language models as evaluators for recommendation explanations. In: Proceedings of the 18th ACM Conference on Recommender Systems, pp. 33–42 (2024)

[45] Pasupat, P., Liang, P.: Compositional semantic parsing on semi-structured tables. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1470–1480 (2015)

[46] Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., Chua, T.-S.: Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 3277–3287 (2021)

[47] Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X., Wang, W.Y.: Tabfact: A large-scale dataset for table-based fact verification. arXiv preprint arXiv:1909.02164 (2019)

[48] Wu, X., Yang, J., Chai, L., Zhang, G., Liu, J., Du, X., Liang, D., Shu, D., Cheng, X., Sun, T., *et al.*: Tablebench: A comprehensive and complex benchmark for table question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 25497–25506 (2025)

[49] Kweon, S., Kwon, Y., Cho, S., Jo, Y., Choi, E.: Open-wikitable: Dataset for open domain question answering with complex reasoning over table. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 8285–8297 (2023)

[50] Zhu, J., Wang, J., Yu, B., Wu, X., Li, J., Wang, L., Xu, N.: Tableeval: A real-world benchmark for complex, multilingual, and multi-structured table question answering. arXiv preprint arXiv:2506.03949 (2025)

[51] Berkovitch, Y., Glickman, O., Somech, A., Wolfson, T.: Generating tables from the parametric knowledge of language models. arXiv preprint arXiv:2406.10922 (2024)

[52] Elnagar, A., Khalifa, Y.S., Einea, A.: Hotel arabic-reviews dataset construction for sentiment analysis applications **740**, 35–52 (2018) https://doi.org/10.1007/978-3-319-67056-0_3

[53] Fang, X., Xu, W., Tan, F.A., Zhang, J., Hu, Z., Qi, Y., Nickleach, S., Socolinsky, D., Sengamedu, S., Faloutsos, C.: Large language models (llms) on tabular data: Prediction, generation, and understanding–a survey. arXiv preprint arXiv:2402.17944 (2024)

[54] Nguyen, N.-H., Sim, T., Dao, H., Joty, S., Kawaguchi, K., Chen, N., Kan, M.-Y., *et al.*: Llms are biased towards output formats! systematically evaluating and mitigating output format bias of llms. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 299–330 (2025)

[55] Tam, Z.R., Wu, C.-K., Tsai, Y.-L., Lin, C.-Y., Lee, H.-y., Chen, Y.-N.: Let me speak freely? a study on the impact of format restrictions on large language model performance. In: Dernoncourt, F., Preoţiuc-Pietro, D., Shimorina, A. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 1218–1236. Association for Computational Linguistics, Miami, Florida, US (2024). https://doi.org/10.18653/v1/2024.emnlp-industry.91 . https://aclanthology.org/2024.emnlp-industry.91/

[56] Xia, C., Xing, C., Du, J., Yang, X., Feng, Y., Xu, R., Yin, W., Xiong, C.: Fofo: A benchmark to evaluate llms' format-following capability. arXiv preprint arXiv:2402.18667 (2024)

[57] Lakkaraju, H., Slack, D., Chen, Y., Tan, C., Singh, S.: Rethinking explainability as a dialogue: A practitioner's perspective. arXiv preprint arXiv:2202.01875 (2022)

[58] Qwen: Qwen: A Large Language Model by Tongyi Lab. https://qwenlm.github.io/. Accessed: 2023-10-01 (2023)

[59] OpenAI: GPT-4o (Version gpt-4o-latest). Large language model (2024). https://openai.com/index/hello-gpt-4o

[60] Liu, T., Wang, X., Huang, W., Xu, W., Zeng, Y., Jiang, L., Yang, H., Li, J.: Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. arXiv preprint arXiv:2409.14051 (2024)

[61] Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. In: Forty-first International Conference on Machine Learning (2023)