ChatGPT

阿拉伯语大模型评测综述

近年来随着大语言模型在阿拉伯语领域的不断发展，针对这些模型的评测基准也逐渐丰富。本文首先汇总了最近两年内发布的**权威阿拉伯语大模型评测数据集**，按照评测目标分为“知识与理解”、“文化常识与价值观”和“推理能力”三个方面。随后，我们总结各评测维度的目标和标准，并**构建评测面板**以展示不同数据集如何覆盖上述维度。最后，我们提供一个**大语言模型部署与评测的教程**，一步步指导如何在服务器上部署阿拉伯语大模型并利用这些数据集进行评测。

1. 最新阿拉伯语大模型评测数据集

过去两年中，研究者发布了多项面向阿拉伯语大模型的评测基准，涵盖了从知识问答、学术测验到文化常识、价值观 alignment 以及推理能力等多个方面。以下按类别列出主要的评测数据集：

1.1 知识与理解评测数据集

- **ArabicMMLU (2024)** – 由 Koto 等人在 ACL 2024 提出的首个阿拉伯语多任务语言理解基准 [1]。该数据集源自中东和北非多国的**学校考试题目**，包含 **40 个任务、14,575 道** 多项选择题，全为现代标准阿拉伯语（MSA）[2]。题目覆盖 STEM（理工）、社科、人文、阿拉伯语言等领域，用于全面评估模型在各科知识上的理解和**常识问答**能力。研究表明当前最好的开源模型在该基准上**准确率不足50%**，最好的阿拉伯语定向模型也仅约 62% [3]，显示仍有很大提升空间。

- **AlGhafa (2023)** – 由阿联酋 TII 团队推出的**多项选择评测基准** [4]。AlGhafa 汇总了多种公开数据集，并新增了一个含 80 亿词的新手工数据集，用于评测阿拉伯语大模型的综合NLP能力 [5]。作为展示，作者训练了包括14B参数在内的一系列阿拉伯语模型，并将其与现有模型比较。AlGhafa 涵盖了**从阅读理解、常识判断到专业领域知识**等多任务，弥补了阿拉伯语评测资源的不足 [4]。

- **阿拉伯语学术考试/阅读理解** – 除了上述综合基准，一些特定任务数据也常用于评测模型的知识理解能力：

- RACE_Ar：由英文高中阅读理解数据集 RACE 翻译/改编而成的阿拉伯语版，用于测评**长文阅读理解**能力 [6]。
- Arabic Reading Comprehension Dataset (ARCD)：约含 1300 多道问答对 [7]（Mozannar 等，2019），测试模型对于篇章的**细节理解**和**问题回答**能力。
- ArabicaQA (2024)：一个包含约 6万多问答对的**开放域问答**数据集（Abdallah 等，2024），涵盖各类常识和百科知识问答，用于评估模型的**知识检索与回答**能力。

上述数据集主要侧重考察模型对**客观知识、文本理解**的掌握程度。例如 ArabicMMLU 通过多学科考试题评估模型在各领域的知识水平，而 RACE_Ar/ARCD 则关注模型理解文章和回答问题的能力。

1.2 文化常识与价值观评测数据集

- **ACVA (Arabic Cultural and Value Alignment, 2023)** – 由 AceGPT 模型作者提出的**阿拉伯文化与价值观对齐**基准 [8]。该数据集用于评估模型对阿拉伯文化细节和社会价值观的理解程度 [8]。ACVA 通常以问答或选择题形式考察模型在具有文化背景的问题上的表现，例如对于风俗习惯、伦理道德的判断是否符合阿拉伯社会规范。在后续研究中，ACVA 常被用来检验模型的**文化敏感性**，例如 LlamAr 模型就使用 ACVA 来评估其对阿拉伯文化细微差异的适应程度 [8]。

- **ArabCulture (2025)** – 一个针对**阿拉伯文化常识推理**的新数据集，由 Huang 等人在 ACL 2025 引入 [9]。该数据集完全由母语者原创题目（非翻译），包含 **3,482 道** MSA 提问，覆盖中东和北非 **13 个国家**，跨越 **12 个主要领域和54个子话题** [10]。内容涉及阿拉伯世界的**社会规范、传统、日常生活**等各方面常识。ArabCulture 可用于评估模型在**文化情境下的常识推理**能力。实验表明，即使规模高达32B的模型，在此数据集上的**文化常识推理**仍然表现不佳，不同地区问题的表现差异明显 [11] [12]。

- **SaudiCulture (2025)** – 专注于**沙特阿拉伯**区域文化的评测基准 [13]。由 Ayash 等人在2025年提出，SaudiCulture 提供了覆盖沙特 **西部、东部、南部、北部、中部五大区域**以及全国通用问题的综合题库 [14]。题目形式多样，包括**开放问答、单选、多选**（部分需要多项正确答案） [15]。内容涵盖**饮食、服饰、娱乐、庆典、手工艺**等文化领域，并区分了全国共知的常识和各地区的特有知识 [16]。研究对 GPT-4、Jais、Fanar、AceGPT 等6个模型进行了测试，结果显示模型在**高度本地化**的问题上表现显著下降，如 GPT-4 在沙特西部问题上准确率66%，但 Jais 在北部问题仅16% [17]。这凸显了模型需要融入**区域知识**以提升文化胜任力。

- **CamelEval (2024)** – 这是一个面向**文化对齐和指令遵循**的综合自动评测框架，由 Qian 等人在2024年提出 [18]。CamelEval 的特点是采用 LLM-as-judge（以大模型作为评审）的自动化评测方案，包含三个测试子集，每个子集 **805** 个案例，分别评估模型的**指令遵循、事实准确性**和**文化贴合度** [18]。尤其是文化子集，精心设计了涉及**阿拉伯语言和文化细微差别**的开放式生成任务 [18]。例如测试模型对不同**方言**的理解、**文化引用**的识别、现代阿拉伯网络用语的掌握等 [19]。CamelEval 力图弥补仅依赖选择题评测的不足 [20] [21]，通过开放问答和让模型进行生成，再由评审模型判断输出是否符合文化预期，从而更全面地评估模型在**阿拉伯文化语境**下是否产生恰当、无偏见的回答 [21] [22]。该框架也伴随推出了一个名为 **Juhaina** 的阿拉伯-English双语LLM，用CamelEval评测证明其在文化相关任务上的优势 [23] [24]。

- **Palm (2024)** – 由 Alwajih 等人发布的一个用于**指令微调和文化评测**的大规模数据集 [25]。Palm 的独特之处在于：它是**完全人工标注**的阿拉伯语指令数据集，覆盖**所有22个阿拉伯国家**，涉及 **20 个文化相关话题** [25]。数据既包括 MSA 也包含各地**本地方言**，每条都是本地专家撰写的指令-回答对 [25]。Palm 数据集既可用于**文化化的模型指令微调**，也可用作评测基准，以检验模型对各国文化背景指令的理解和响应能力 [26]。作者使用 Palm 对多个开源和前沿模型进行了评测分析，包括对比模型在 MSA vs 方言上的表现差异，以及借助 GPT-4 等作为评审来自动评价输出质量 [27] [28]。Palm 的发布填补了**阿拉伯国家粒度**文化内容在指令数据和评测中的空白。

上述数据集与基准专注于模型对**阿拉伯文化**的了解程度，从普遍的文化常识（ArabCulture）、价值观偏好（ACVA），到特定国家/地区文化（SaudiCulture）以及语言多样性（Palm）。这些评测有助于确保大模型的输出符合阿拉伯世界的文化语境和社会期望。

1.3 推理能力评测数据集

- **COPA_Ar** – 由英文常识因果推理数据集 COPA 翻译而来的**阿拉伯语常识因果推理**任务，用于测试模型在给定情境下选择可能原因/结果的能力，属于**常识推理**范畴 [6]。COPA_Ar 提供两句子，要求模型判断因果关系，在阿拉伯语环境下检验模型的**逻辑推断**和**常识理解**。

- **AraLogic / ArBench (待发布)** – 社区中也有提议专门构建阿拉伯语的逻辑推理评测基准（暂称 ARBench），涵盖数字推理、真假判断等，但截至目前主要通过翻译现有英语任务（如 **BoolQ** 判断句子真假、**ReCORD**共指推理等）来评估。这方面的工作仍在进行中，尚未形成统一公开的数据集，但体现了对**阿拉伯语严格推理**能力评测的关注。

- **数学推理与分析** – 对于数学和算术推理能力，Alghamdi 等人在 2022 年构建了 Ar_Math 数据集，包含约 6000 道数学题及解答步骤 [7]。虽然略早于两年范围，但它经常用于评测模型的**算术和逻辑推导**能力。此外，一些研究也通过将 GSM8K 等数学问答翻译为阿拉伯语，来考察模型在母语环境下解决数学问题的表现。

- **AraTable (2025)** – Myung 等人提出的针对**阿拉伯语表格数据推理**的评测框架 [29]。AraTable 包含从维基百科和现实数据源提取的阿拉伯语表格，并自动生成了多种问题以评估模型对表格的**推理问答**能力 [30]。题型包括算术运算、逻辑比较、表格事实核对等 [31]，综合考察模型从**简单查询**到**复杂推理**的水平。这是**多模态**（结构化数据）推理的一个特定方向，丰富了阿拉伯语推理评测的维度。

综上，阿拉伯语大模型的推理能力评测涵盖**常识因果推断**(COPA_Ar)、**逻辑判断**、**数学推理**(Ar_Math)以及**结构化数据推理**(AraTable)等多个方面。这些基准可以让我们发现模型在中文本之外的**深层推理**和**分析能力**上的不足之处，从而有针对性地改进。

1.4 数据集链接与来源索引

为方便进一步查阅，这里列出上述主要数据集/基准的来源链接：

- **ArabicMMLU** – [GitHub: mbzuai-nlp/ArabicMMLU (包含数据下载及论文链接)] [1] [32]；论文预印本 [33]。
- **AlGhafa** – [ACL Anthology: ArabicNLP 2023 论文] [4]。
- **ACVA** – 由 AceGPT 提出的文化价值评测，见 AceGPT 论文 (Huang et al. 2023) [8]。
- **ArabCulture** – [ACL 2025 论文: Commonsense Reasoning in Arab Culture] [9]。
- **SaudiCulture** – [Springer 期刊论文 2025，开放获取] [13]。
- **CamelEval & Juhaina** – [arXiv 2024 论文: Advancing Culturally Aligned Arabic LLMs] [18]；见摘要对 CamelEval 的描述 [18]。
- **Palm 数据集** – [arXiv 2025 预印本: Palm: A Culturally Inclusive Arabic LLM Dataset] [25]。
- **COPA_Ar/RACE_Ar** – 翻译自原始COPA和RACE，由社区评测使用 [6]。
- **AraTable** – [arXiv 2025 论文: AraTable: QA over Arabic Tables] [30]。

读者可通过以上链接获取更多细节（如完整数据、下载方法等）。接下来，我们将基于上述数据构建评测维度面板，明确不同数据集所评测的侧重点。

2. 评测维度与标准面板

大模型评测通常围绕多个能力维度展开。针对阿拉伯语模型，我们特别关注以下三大方面，每一方面都对应若干评测数据集和指标：

评测维度	代表数据集 / 基准	评估重点
知识与理解 （语言理解与领域知识）	- **ArabicMMLU** [1]：40项多学科考试题，测试**各学科常识**和**专业知识** - **AlGhafa** [4]：多任务选择题，覆盖**广泛NLP任务** - **RACE_Ar** [6] / **ARCD**：长文阅读理解，考察**段落理解** - **开放QA**（ArabicaQA等）：百科常识问答，考验**知识检索**能力	- **语言理解**：准确理解问句和篇章内容 - **常识/百科知识**：掌握历史、地理、科学等常识 - **领域知识**：专业领域（理工、人文等）的知识储备 - **阅读理解**：综合文章信息作答的能力
文化常识与价值观 （本地化知识与文化适应）	- **ACVA** [8]：文化价值观对齐测试，检查**价值倾向**是否契合 - **ArabCulture** [9]：阿拉伯文化常识问答，涉及**社会习俗**和**地域知识** - **SaudiCulture** [14]：沙特区域文化知识，评估**地区差异**理解 - **CamelEval（文化子集）** [18]：生成任务评文化敏感度，如**方言**、**俗语**运用 - **Palm 数据集** [25]：多国多方言指令，评测模型**对各国文化背景指令的响应**	- **文化常识**：理解风俗传统、历史典故等本地常识 - **价值观 alignment**：输出是否符合社会伦理和主流价值 - **地域多样性**：区分不同国家/地区的文化差异 - **语言多样性**：对MSA与各地口语**方言**的掌握程度 - **文化敏感性**：避免冒犯性或西方偏见的内容

评测维度	代表数据集 / 基准	评估重点
推理与推断 （逻辑、常识和分析）	- **COPA_Ar** [6]：因果常识推理，考察**日常逻辑** - **Ar_Math**：数学推理题，考验**数字计算与逻辑** - **AraTable** [30]：表格数据推理，涉及**多步推理** - **推理挑战**（BoolQ、WIQA等翻译版）：判断句子真假、因果问答等 - **CamelEval（指令/事实子集）** [18]：开放问答，由模型评审**推理正确性**	- **常识推理**：基于常识判断因果、目的等（COPA类） - **逻辑演绎**：遵循逻辑规则推断结论的能力 - **数学推理**：算术、代数和数学词题的求解 - **多步推理**：需要跨句整合信息得出答案 - **事实一致性**：生成回答中事实是否正确、一致

表：阿拉伯语大模型评测的主要维度、对应数据集及其评估重点。

上述面板展示了评测框架如何覆盖模型能力的方方面面。在实际评测时，我们会根据模型用途选择相应的数据集组合，如面向通用助手型模型，会综合考虑知识问答、文化敏感性、推理等；若是垂直领域模型，则侧重该领域相关的理解和推理测试。

值得注意的是，当前不少阿拉伯语评测仍在增长和完善中。例如**ArabicMMLU**和**AlGhafa**主要以选择题形式评估客观问题，对开放生成任务（如对话中的礼貌程度、文化语气等）的考察不足 [20] [21]。针对这点，CamelEval 等引入了生成式评估，使我们能检测模型输出中的**微妙偏差**（如用词不符合本地文化） [21] [22]。因此，在构建评测体系时，**选择题+生成任务**相结合是更全面的方案。

3. 大模型部署与评测教程

下面我们提供在服务器上部署开源阿拉伯语大语言模型并进行评测的具体步骤指南。假定读者已有一台装有 NVIDIA GPU 的服务器环境（如Ubuntu系统），并具备基本的命令行操作能力。步骤如下：

3.1 环境准备

1. **硬件要求**：由于大模型参数量大，建议使用具有 **高显存GPU** 的服务器（如 NVIDIA A100 40GB 或至少 24GB 显存的卡）。模型参数7B以上一般需要≥16GB显存，13B模型建议≥24GB。如显存不足，可考虑多卡或采用8-bit量化加载。
2. **操作系统**：Linux 环境（Ubuntu 20.04+）或其他兼容POSIX的系统。确保安装了 NVIDIA 驱动及 CUDA 工具包，使 PyTorch 能正常调用 GPU。
3. **Python 环境**：推荐使用 Python 3.8+，可通过 Anaconda 或 `venv` 创建隔离环境：

```
# 创建并激活虚拟环境
conda create -n llm_env python=3.9
conda activate llm_env
```

4. **依赖安装**：安装 Hugging Face Transformers、Datasets 库，以及评测所需工具：

```
pip install torch transformers[torch] datasets evaluate bitsandbytes
```

5. `torch` 可根据CUDA版本选择对应版本安装；Transformers用于模型加载；Datasets用于加载评测数据集；`evaluate` 是HuggingFace评测库；`bitsandbytes` 用于后续8-bit量化加载模型（可选）。

3.2 模型获取与部署

1. **下载开源模型**：选择需要部署的阿拉伯语大模型并下载权重。常用开源阿拉伯语或多语言模型有：

2. **Jais-13B** (Inception / Cerebras发布)：13亿参数的**阿拉伯语-英语**双语模型，在72B阿拉伯语标记上训练 [34] 。HuggingFace模型仓库为 `inceptionai/jais-13b` [34] 。还有微调后的聊天版 `inceptionai/jais-13b-chat` 可用。
3. **Juhaina** (ELM研发布署)：一个对阿拉伯文化对齐的**7B级别**模型（双语），在多样本地语料上预训练 [23] 。可通过 HuggingFace (`elmresearchcenter/juhaina`) 获取。
4. **Qwen-7B / 14B** (阿里巴巴达摩院)：虽然主要面向中英，但作为多语种基础模型，在阿拉伯语上也有良好理解度 [35] 。HuggingFace上模型名为 `Qwen/Qwen-7B` 等。需要注意 Qwen 模型需遵守其开源协议。
5. **ARBERT / MARBERT**：早期的阿拉伯语BERT系列（推理问答和分类用），参数较小（~0.2B），可作为对比基线。但不支持生成，仅用于理解任务。

6. 其他如 **Fanar**、**Noor**, **BLOOMZ**（多语种）, **LLaMA2 (Arabic fine-tune)** 等也可选择。根据评测需求和硬件资源决定加载哪个模型。

7. **模型加载**：使用 Transformers 提供的 `AutoModelForCausalLM` 等接口加载模型。例如加载 Jais-13B 模型：

```
import torch
from transformers import AutoTokenizer, AutoModelForCausalLM
model_name = "inceptionai/jais-13b"  # 模型仓库名称
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    device_map="auto",          # 自动分配到GPU
    torch_dtype=torch.float16 # 使用16-bit半精度节省显存
)
```

如果显存有限，可进一步设置 `load_in_8bit=True` 并安装 BitsAndBytes 实现8-bit量化加载，这将大幅降低显存占用：

```
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    device_map="auto",
    load_in_8bit=True,     # 启用8-bit量化
    torch_dtype=torch.float16,
    trust_remote_code=True  # 某些模型需要这个参数加载自定义代码
)
```

> **提示**: 部分模型（如 Jais）在加载时需要 `trust_remote_code=True` 以允许自定义模型类 [36] 。另外，请确保登录 HuggingFace账户并接受模型协议（如果有的话），否则在 `from_pretrained` 时可能报错要求认证。

8. **模型推理测试**：加载完成后，可先用简单示例测试模型生成效果。例如：

```
prompt = "首都的问题: 阿拉伯联合酋长国的首都是哪个城市？"
inputs = tokenizer(prompt, return_tensors="pt").to("cuda")
outputs = model.generate(**inputs, max_new_tokens=50)
```

```
answer = tokenizer.decode(outputs[0], skip_special_tokens=True)
print(answer)
```

模型应输出类似“阿拉伯联合酋长国的首都是阿布扎比。”的答案。可以尝试中英提示，检查模型双语能力[37]。

3.3 评测数据准备

1. **获取评测数据**：利用 Datasets 可以方便地加载前述评测基准的数据。例如加载 ArabicMMLU 全部任务：

```
from datasets import load_dataset
mmlu = load_dataset('MBZUAI/ArabicMMLU', 'All')  # 加载全部任务
print(mmlu['train'][0])  # 查看第一道题示例
```

类似地，许多数据集可通过 HuggingFace Hub 获取（如 `MBZUAI/ArabicMMLU`[38]，`arabicNLP/ArabCultureQA` 等）。对于没有直接开放的数据（如 ACVA、CamelEval），可能需要从论文附录或作者GitHub获取，我们可以手动准备若干示例进行测试。

2. **选择评测任务**：根据关注的维度选择数据子集。例如：

3. 知识问答：从 ArabicMMLU 中挑选某个主题，或使用 **开放域问答** 数据（如 TyDiQA-Ar）。
4. 文化常识：使用 ArabCulture 数据集中某国的问题，或自定义几道 ACVA 风格的问题。

5. 推理能力：选择 COPA_Ar 的若干题，或简单数学题。
   确定测试集后，可以将问题与标准答案列表准备好。

6. **数据格式**：确保模型能处理输入格式。如果是选择题类，可以把问题和选项拼成一个 Prompt，让模型回答选项字母或内容。如果是开放问答，则直接提问。必要时在 Prompt 前加上系统指令（如 "你是一个博学的助手..."）以调整模型语气，使其以所需形式作答。

3.4 执行评测与收集结果

1. **推理生成**：针对测试集中每个问题，将其送入模型生成回答：

```
def ask_model(question):
    inputs = tokenizer(question, return_tensors="pt").to("cuda")
    output_ids = model.generate(**inputs, max_new_tokens=100)
    answer = tokenizer.decode(output_ids[0], skip_special_tokens=True)
    return answer

# 示例：对前3道MMLU题目让模型作答
for i in range(3):
    q = mmlu['train'][i]['question']
    choices = mmlu['train'][i]['choices']
    formatted_q = q + "\n选项: " + " ".join([f"({c})" for c in choices])
    print("Q:", formatted_q)
    print("A:", ask_model(formatted_q))
```

对于有标准答案的数据，可以直接让模型选择（如回答 `(A)` 或输出选项内容）。生成结果需要进行**后处理**以提取模型最终答案。

2. **评估准确性**：将模型输出与标准答案进行比对。以选择题为例：

```
predictions = []
references = []
for item in mmlu['validation']:  # 假设有validation或test split
    q = format_question(item)    # 自定义函数格式化问题+选项
    model_ans = ask_model(q)
    pred_choice = extract_choice(model_ans)  # 提取字母或选项文本
    true_choice = item['answer']             # 标准答案
    predictions.append(pred_choice)
    references.append(true_choice)
# 计算准确率
accuracy = sum(1 for p, t in zip(predictions, references) if p == t) / len(references)
print("Accuracy:", accuracy)
```

如果是开放问答，则可能需要计算 BLEU、ROUGE 等评分，或采用人工评估正确性。也可以借助 HuggingFace `evaluate` 库的现成指标：

```
import evaluate
rouge = evaluate.load("rouge")
results = rouge.compute(predictions=predictions, references=references)
print(results)
```

对于文化价值观等生成任务，可考虑引入**自动评审模型**（如 GPT-4）来判分，或者根据预定义规范规则打分。

3. **多维度记录**：最好针对每类任务分别统计指标。例如知识问答用准确率，文化题可以根据回答符合度打分，推理题看正确率等。将结果整理成表格，便于和其它模型横向比较。在 SaudiCulture 论文中，不同模型在各区域题目的准确率差异明显 [17]；在 ArabicMMLU 中，开源模型整体低于闭源模型 [3]。这些对比有助于了解被测模型的相对水平。

4. **结果分析**：通过结果可以分析模型的强项和弱项。例如，如果发现模型在**宗教习俗类**问题上错误较多，可能预示训练语料中文化内容不足，需要加强。如果数学题全错，则表明模型**算术推理**能力薄弱，可以考虑补充相应数据微调。分析时也可以参考论文中的基准表现，例如 ArabCulture 数据集的研究指出模型在某些地区问题上表现尤为糟糕 [11]，这些都是值得关注的改进方向。

3.5 后续改进与部署

1. **模型微调**：若评测发现模型在特定类别任务上效果不佳，可以考虑收集相关数据对模型进行指令微调或继续预训练。例如利用 Palm 数据集中某国数据细粒度微调，以提升模型对该国文化的掌握。在微调前后再次运行上述评测，验证提升幅度。
2. **部署API**：当模型在测试集上达到满意表现后，可将其通过 API 服务化。例如使用 FastAPI 或 Flask 编写一个简单的 Web 服务，将 `ask_model` 封装为接口，方便前端调用。HuggingFace 也提供 Inference Endpoint（需付费）或者使用 `transformers` 的 `TextGenerationPipeline` 快速构建推理管道：

```
from transformers import pipeline
text_gen = pipeline("text-generation", model=model, tokenizer=tokenizer, device=0)
result = text_gen("用户: 你好，用阿拉伯语问好\n助手:", max_new_tokens=50)
print(result[0]['generated_text'])
```

这样可以集成到聊天机器人等应用中。

3. **持续评测**：部署后仍需持续评测监控。如果模型将来更新版本或经过微调，需要再次运行先前测试确保性能没有回退，并观察在真实用户提问中是否出现新问题。例如监测模型是否对某些敏感文化话题产生不当回答，从而及时调整。

最后，将上述教程内容整理发布到项目仓库（如提供的 GitHub 链接：`Ruoyu1106/arabic-llm-tuning`），便于团队协作和他人参考学习。通过**系统的评测**，我们能全面了解阿拉伯语大模型的能力边界，从而不断改进模型表现，打造更强大、更符合本地需求的开源阿拉伯语大模型。

参考文献：

1. Koto, F. et al. (2024). ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic. [1] [3]
2. Almazrouei, E. et al. (2023). AlGhafa Evaluation Benchmark for Arabic Language Models. [4]
3. Huang, X. et al. (2024). AceGPT: Introducing Arabic Cultural and Value Alignment (ACVA) Benchmark. [8]
4. Huang, X. et al. (2025). Commonsense Reasoning in Arab Culture (ArabCulture Dataset) [9]
5. Ayash, L. et al. (2025). SaudiCulture: Evaluating LLMs' Cultural Competence in Saudi Arabia. [14] [17]
6. Qian, Z. et al. (2024). CamelEval: Benchmarking Arabic LLMs for Cultural Alignment [18]
7. Alwajih, F. et al. (2025). Palm: A Culturally Inclusive and Linguistically Diverse Dataset for Arabic LLMs. [25]
8. Zhou, H. et al. (2023). LlamAr & GemmAr: Enhancing LLMs Through Arabic Instruction-Tuning [35] [6]
9. Myung, C. et al. (2025). AraTable: Benchmarking LLMs on Arabic Tabular Data. [30]

---

[1] [2] [32] [38] GitHub - mbzuai-nlp/ArabicMMLU
https://github.com/mbzuai-nlp/ArabicMMLU

[3] [33] [2402.12840] ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic
https://arxiv.org/abs/2402.12840

[4] [5] AlGhafa Evaluation Benchmark for Arabic Language Models | OpenReview
https://openreview.net/forum?id=8yz3laymKl&referrer=%5Bthe%20profile%20of%20Ruxandra%20Cojocaru%5D(%2Fprofile%3Fid%3D~Ruxandra_Cojocaru1)

[6] [7] [8] [35] LlamAr & GemmAr: Enhancing LLMs Through Arabic Instruction-Tuning
https://arxiv.org/html/2407.02147v1

[9] [10] [11] [12] aclanthology.org
https://aclanthology.org/2025.acl-long.380.pdf

[13] [14] [15] [16] [17] SaudiCulture: A benchmark for evaluating large language models' cultural competence within Saudi Arabia | Journal of King Saud University Computer and Information Sciences
https://link.springer.com/article/10.1007/s44443-025-00137-9

18 20 21 22 openreview.net
https://openreview.net/pdf?id=hYLNm07w7c

19 23 24 CamelEval: Advancing Culturally Aligned Arabic Language Models and Benchmarks | AI Research Paper Details
https://www.aimodels.fyi/papers/arxiv/cameleval-advancing-culturally-aligned-arabic-language-models

25 26 27 28 Palm: A Culturally Inclusive and Linguistically Diverse Dataset for Arabic LLMs
https://arxiv.org/html/2503.00151v1

29 30 31 AraTable: Benchmarking LLMs' Reasoning and Understanding of Arabic Tabular Data
https://arxiv.org/html/2507.18442

34 36 37 inceptionai/jais-13b · Hugging Face
https://huggingface.co/inceptionai/jais-13b