ChatGPT

知识与理解维度：阿拉伯语大模型评测数据集

本文档汇总了目前可用于评测阿拉伯语大语言模型（LLM）知识与理解能力的主要数据集，并描述每个数据集的任务内容、常见评判标准及下载链接/论文链接。除了传统指标外，也提供了推荐的扩展评价指标，以便构建更全面的评测面板。

数据集概述

- **ArabicMMLU** – 14,575 题，多学科选择题，覆盖 40 个科目 [1]；评测指标：准确率。
- **AlGhafa** – 多任务选择题基准，包含翻译自多种公开任务（Belebele、COPA-Ar、AraFacts 等） [2]；评测指标：准确率。
- **RACE_Ar** – 阿语版 RACE 阅读理解数据集，长篇文章配多项选择题。
- **ARCD** – Arabic Reading Comprehension Dataset，1,395 条众包问题及对应篇章 [3]；评测指标：EM + F1 [4]。
- **ArabicaQA** – 大规模开放式问答和阅读理解数据集，89,095 条可回答问题和 3,701 条无法回答的问题 [5]；评测指标：EM/F1 和人工评审 [6]。
- **AraSTEM** – 11,637 道多选题，覆盖数学、科学、物理、化学、医学等学科 [7]；评测指标：准确率 [8]。
- **ADMD** – Arabic Depth Mini Dataset，490 道高难度问题跨十个领域 [9]；评测指标：准确率。
- **INCLUDE** – 197,243 道多选题，覆盖 44 种语言的地方考试 [10]；评测指标：准确率 [11]。
- **AAFAQ** – 5,009 条阿语问题，标注 11 类语言属性用于问题分类 [12]；评测指标：分类准确率、宏/微 F1 等 [13]。
- **ORCA** – 综合基准，整合 60 个数据集组成 29 项任务及七大任务簇 [14]；评测指标：统一的 ORCA Score。

各数据集详解

ArabicMMLU

ArabicMMLU 是首个全面评测阿语多任务理解能力的多项选择基准，包含 40 个科目共 **14,575 道** 题目 [1]。题目覆盖从中学到大学的通识知识，如理工、社科、语言学等。评测通常采用**准确率**为主要指标；另可按学科统计子准确率。
- **数据集页面**：`MBZUAI/ArabicMMLU`（Hugging Face）
- **论文链接**：https://arxiv.org/abs/2402.12840

AlGhafa

AlGhafa 是由阿联酋科技创新院（TII）推出的多任务选择题基准，用于评估阿语大模型在零样本和少样本条件下的综合能力。它集合了多个现有任务并加入自建的 HandMade 语料 [15]。任务包括 Belebele 阅读理解、COPA-Ar 因果推理、AraFacts 真/假判断、MCQ Exams、OpenBookQA 等 [2]。主要评判指标为**准确率**，允许细分各子任务。
- **数据集页面**：`OALL/AlGhafa-Arabic-LLM-Benchmark-Native`
- **论文链接**：https://aclanthology.org/2023.arabicnlp-1.21

RACE_Ar

RACE_Ar 是英文 RACE 数据集的阿拉伯语翻译版，包含学生阅读考试中的长篇文章与选择题。它测试模型对长文章的理解能力。评测指标包括 **准确率** 和 **EM**（完全匹配率）。
- **数据集页面**：`Hennara/race_ar`（Hugging Face）
- **原始 RACE 论文**：https://arxiv.org/abs/1704.04683

ARCD – Arabic Reading Comprehension Dataset

ARCD 提供 **1,395 条** 众包问题与相应维基百科篇章 [3]。它是阿拉伯语版的 SQuAD 风格数据集，用于抽取式问答。评测采用 **EM** 和 **F1** 统计答案与标注的匹配程度 [3]。
- **数据集链接**：https://github.com/husseinmozannar/SOQAL
- **论文链接**：https://aclanthology.org/W19-4612

ArabicaQA

ArabicaQA 是 2024 年发布的开放域问答与阅读理解集合，包含 **89,095 条可回答问题**和 **3,701 条无法回答的问题** [5]。大部分问题由众包工人基于维基百科生成，兼顾短答案和长答案 [16]。评测采用 **EM/F1**（抽取式）和 **人工评审**（生成式），并可配合 AraDPR 检索器使用 [6]。
- **数据集仓库**：https://github.com/DataScienceUIBK/ArabicaQA
- **论文链接**：https://arxiv.org/abs/2403.17848

AraSTEM

AraSTEM 由 11,637 道多项选择题组成，涵盖数学、物理、化学、生物、医学、信息技术等科目 [7]。数据分为小学、中学和大学三个层次。论文使用**思维链提示**（chain-of-thought prompts）引导模型逐步推理，并以 **准确率**衡量各学科表现；Jais-30B-Chat 在该基准上约 56% 准确率 [8]。
- **数据集/代码链接**：https://github.com/arabic-nlp/ara-stem
- **论文链接**：https://arxiv.org/abs/2501.00559

ADMD – Arabic Depth Mini Dataset

ADMD 是一套 **490 道** 高难度问题，跨越十个主要领域（包括数学、阿语语言、伊斯兰学、法律等） [9]。设计目的是衡量 LLM 在深层知识和文化理解方面的能力，题目多为开放式或复杂选择题。常用的评判标准是 **准确率**，但由于题目难度大，各模型表现普遍较低（约 30%） [9]。
- **数据集链接**：暂未公开；可通过论文中提供的联系获取
- **论文链接**：https://arxiv.org/abs/2405.10957

INCLUDE – Multilingual Evaluation with Regional Knowledge

INCLUDE 是 2025 年公布的多语言评测基准，包含 **197,243** 道多选题，来自不同国家/地区考试 [10]。其中约 6,000 道属于阿拉伯语（具体数目未单独披露）。它主要用于测试模型在地方知识和推理上的能力。评测指标为 **准确率**；论文测试表明 5-shot 提示可显著提升性能，GPT-4o 在全体语言上达到 77.1% [11]。
- **数据集链接**：https://huggingface.co/datasets/KnowledgeLab/Include
- **论文链接**：https://arxiv.org/abs/2503.04437

AAFAQ Question Classification Dataset

AAFAQ 收录 **5,009 条** 阿语问题，并为每条标注 11 个认知与语言属性（疑问词类型、意图、回答类型、认知层次等） [12]。它既可用于多标签分类任务（评估指标为 **准确率**、**宏/微 F1**），也可作为辅助特征提升生成式问答质量 [13]。

- **数据集链接**：https://github.com/UBC-NLP/AAFAQ
- **论文链接**：https://www.nature.com/articles/s41597-025-02790-x

ORCA Benchmark

ORCA 是一个综合性阿语 NLU 基准，覆盖 **60 个公开数据集**并划分为 **29 个任务和 7 大任务簇** [14]。它提供统一的 **ORCA Score** 用于评价模型在多任务上的总体性能，是衡量阿语模型综合理解能力的黄金指标。
- **数据集链接**：https://huggingface.co/datasets/UBC-NLP/orca
- **论文链接**：https://aclanthology.org/2023.findings-acl.609

推荐扩展评价指标

为了对模型的知识与理解能力进行更细致的分析，建议引入以下评估指标：

- **平衡准确率**：适用于答案分布不均的数据集，避免模型倾向多数类别。
- **加权 F1**：结合类别权重评估模型在所有类别上的综合性能。
- **BLEU/ROUGE/BERTScore**：用于生成式问答，衡量回答与参考答案的相似度。
- **模型置信度分析**：记录多选预测的概率分布，以分析模型犹豫程度 [17]。
- **事实一致性检验**：利用事实核查模型验证生成答案的正确性，补充准确率指标。
- **难度分层分析**：将题目按难度分层，分别计算指标，观察模型在简单与复杂问题上的表现差异。

该文档旨在为阿拉伯语 LLM 研究者提供快速查阅的评测数据集索引，帮助你在 GitHub 项目中随时调取相应资源并开展知识与理解维度的系统评测。

---

[1] [2402.12840] ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic
https://arxiv.org/abs/2402.12840

[2] OALL/AlGhafa-Arabic-LLM-Benchmark-Native · Datasets at Hugging Face
https://huggingface.co/datasets/OALL/AlGhafa-Arabic-LLM-Benchmark-Native

[3] [4] Neural Arabic Question Answering
https://aclanthology.org/W19-4612.pdf

[5] [6] [16] ArabicaQA: A Comprehensive Dataset for Arabic Question Answering
https://arxiv.org/html/2403.17848v1

[7] [8] [17] 2501.00559.pdf
https://arxiv.org/pdf/2501.00559.pdf

[9] From Guidelines to Practice: A New Paradigm for Arabic Language Model Evaluation
https://arxiv.org/html/2506.01920v1

[10] [11] Include: Evaluating Multilingual Language Understanding with Regional Knowledge
https://arxiv.org/html/2411.19799v1

[12] [13] A Benchmark Arabic Dataset for Arabic Question Classification using AAFAQ Framework | Scientific Data
https://www.nature.com/articles/s41597-025-05688-0

[14] ORCA: A Challenging Benchmark for Arabic Language Understanding - ACL Anthology
https://aclanthology.org/2023.findings-acl.609/

15 2023.arabicnlp-1.21.pdf
https://aclanthology.org/2023.arabicnlp-1.21.pdf