CamelEval: Advancing Benchmarks for Culturally Aligned Arabic Language Models in Generative Tasks

Anonymous ACL submission

Abstract

Large Language Models (LLMs) serve as the foundation of contemporary artificial intelligence systems. Recently, a diverse range of Arabic-centric LLMs has emerged, designed to align with the values and preferences of Arabic speakers and offering advanced capabilities such as instruction following, open-ended question answering, and information delivery. In this paper, we identify the limitations of existing Arabic LLM benchmarks, which rely exclusively on multiple-choice questions and thereby fails to adequately assess the text generation capabilities of LLMs. To address this shortcoming, we propose a new automated evaluation benchmark, CamelEval, that performs LLM-as-judge evaluation. CamelEval comprises three test suites to evaluate general instruction following, factuality, and cultural alignment. Each test suite contains 805 carefully curated challenging test cases that reflect the nuances of Arabic language and culture. We envision CamelEval as a tool to guide the development of future Arabic LLMs, serving over 400 million Arabic speakers by providing LLMs that not only communicate in their language but also understand their culture.

1 Introduction

A variety of Arabic-capable LLMs have been developed to serve the 400 million Arabic speakers worldwide (Sengupta et al., 2023; Huang et al., 2023; Aryabumi et al., 2024; Penedo et al., 2023; Muennighoff et al., 2022; Ethnologue). However, the evaluation of Arabic LLMs remains preliminary. In fact, existing benchmarks primarily use *multiple choice questions* (MCQs) to assess LLMs in close-ended discriminative tasks, e.g., telling if a statement is true or false (Almazrouei et al., 2023; FreedomIntelligence, 2024; Koto et al., 2024; Elfilali et al., 2024). As a result, Arabic LLMs' capability in open-ended generative tasks such as following instructions and generating responses, remain poorly measured. The lack of such evaluation poses a significant risk to the practical utilization of Arabic LLMs because many important use cases, from chatbots to AI assistants, depends on the text generation capabilities.

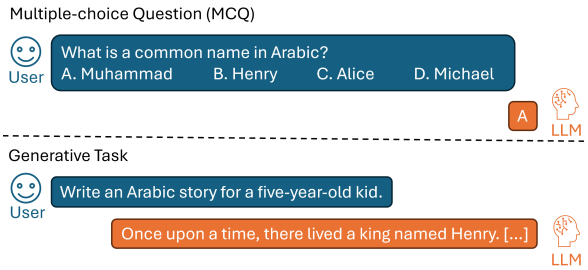

Figure 1: Success in multiple-choice questions does not imply ability to generate culturally aligned output.

Furthermore, although several existing benchmarks evaluate the *knowledge* about Arabic facts and culture (Almazrouei et al., 2023; FreedomIntelligence, 2024; Koto et al., 2024), their MCQ test cases may not capture factuality or cultural alignment in *response generation*. For instance, a LLM may correctly select "Muhammad" as a common Arabic name as opposed to "Henry", but it may still use "Henry" as the name of the main character when asked to generate a Arabic story (Figure 1). Failure to properly evaluate the nuanced factual and cultural aspects may lead to LLMs generating biased, misaligned, or even offensive content.

In summary, there currently lacks a benchmark that gauges the progress towards the following three key goals of Arabic LLMs.

1. Instruction following in the Arabic Language. Whether the LLM understands user instruction in Arabic and generates outputs that are coherent, grammatically correct, and helpful.

2. Factuality in Generation. Whether the LLM incorporates accurate factual information in general (e.g., science) and region-specific fields (e.g., Arabic history) when generating responses.

Benchmark	Task Style	Data Source	Instruction Following	Factuality	Culture
AlGhafa	MCQ	Original	✗	Discriminative	Discriminative
ACVA	MCQ	Original	✗	✗	Discriminative
ArabicMMLU	MCQ	Original	✗	Discriminative	Discriminative
OALL-Trans	MCQ	Translated	✗	Discriminative	✗
CamelEval	Generation	Original	✓	Generative	Generative

Table 1: Comparison of CamelEval and existing benchmarks. CamelEval expands the scope of existing benchmarks which focus on multiple choice questions (MCQ). It can evaluate LLM's generative capabilities in open ended tasks such as instruction following, factuality and culture alignment.

3. Arabic Cultural Alignment in Generation. Whether the LLM generates responses that are appropriate and respectful to the Arabic audience, adhering to the cultural norms of the region.

To close the evaluation gap, we propose CamelEval, a new benchmark for Arabic LLMs, specifically designed to assess their language generation abilities and instruction-following proficiency within Arabic contexts. To achieve the three goals set above, CamelEval contains three sets of challenging test cases ("Arabic Instruction Following", "Factuality", and "Culture") that are grounded in rigorously curated textbook-quality corpses. CamelEval embraces the LLM-as-judge framework, an approach adopted by many state-of-the art LLM benchmarks such as AlpacaEval (Li et al., 2023) and Arena Hard (Li et al., 2024). We have also incorporated bias mitigation techniques to ensure CamelEval reflects LLMs' true capabilities.

We evaluate 20 popular LLMs on CamelEval and confirm that CamelEval preserves the scaling law of LLMs in model size. Furthermore, the performance ranking of LLMs differs significantly between CamelEval and existing benchmarks, suggesting that CamelEval offers an extra dimension of evaluation not previously captured.

Contribution. CamelEval advances existing MCQ-based Arabic LLM benchmarks by capturing LLMs' generative capability with focus on instruction following, factuality, and cultural alignment.

2 Existing Arabic LLM Benchmarks

The Open Arabic LLM Leaderboard (OALL) is by far the most widely adopted Arabic LLM leaderboard (Elfilali et al., 2024). It encompasses three primary benchmarks: AlGhafa (Almazrouei et al., 2023), ACVA (FreedomIntelligence, 2024), and ArabicMMLU (Koto et al., 2024). OALL also include translated versions of standard LLM benchmarks such as EXAMS, ARC, BOOLQ, COPA, HELLASWAG, OPENBOOK-QA, PIQA, RACE, SCIQ, and TOXIGEN. A comparison of these benchmarks and CamelEval is included in Table 1.

It is worth highlighting that all existing benchmarks employ MCQs as the evaluation task. Therefore, they do not directly measure the broad spectrum of LLM capabilities, such as generating responses or following instructions. Furthermore, although existing benchmarks cover factuality and cultural aspects, they do not measure how well LLMs can utilize facts or cultural awareness to generate aligned responses because MCQ primarily captures the discriminative rather than generative capabilities (more details in Appendix A.1). CamelEval bridges this gap by enabling factuality and cultural evaluation in the generative setting.

3 The CamelEval Benchmark

3.1 The Three Test Suites

CamelEval comprises three test suites ("Arabic Instruction Following", "Factuality", and "Culture") to measure the progress towards the three goals we set in the Introduction. Each test suite contains 805 user prompts that serve as the test cases (available in the Data submission). The detailed process of curating the corpus and the test prompts is documented in Appendix A.2.

Arabic Instruction Following (CE-Instruct). This test suite consists of translations of the test cases in the AlpacaEval benchmark (Li et al., 2023), which has gained widespread use for evaluating LLMs in English and has demonstrated a high level of concordance with assessments made by human evaluators (Zheng et al., 2023; Dubois et al., 2024). We have enlisted native Arabic-speaking annotators to translate the set into Arabic.

Factuality (CE-Fact). The factuality suite is based on a corpus of human-curated, textbook-quality content spanning various fields. It includes

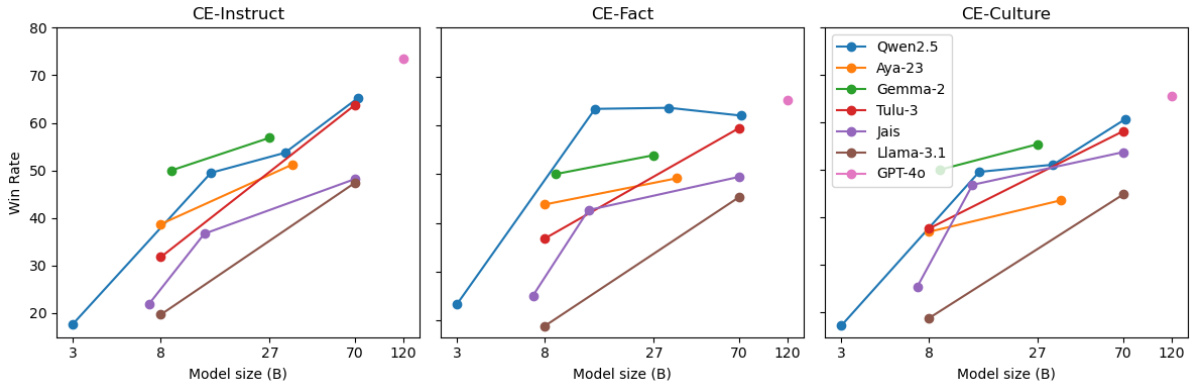

Figure 2: Visualization of family-wise model performance on CamelEval. The "CE-Instruct", "CE-Fact", and "CE-Culture" refer to the instruction following, factuality, and culture suite of CamelEval. The table containing full numerical results is available in Appendix A.6.

both general scientific disciplines such as Physics as well as Arabic-specific topics such as history. We prompted GPT-4o to synthetically generate test cases based on the documents in the corpus (Appendix A.3). Finally, we enlisted annotators to manually select 805 test cases and edit them to ensure relevancy, diversity, and difficulty.

Culture (CE-Culture). The culture suite is curated in a similar manner as the Factuality suite. The test cases are grounded in high-quality corpus covering cultural topics such as social norms, religion, art, and biography. The final test set are also inspected and improved by local Arabic annotators.

3.2 LLM-as-judge Evaluation

LLM-as-judge evaluation has been widely adopted as a cost-effective and scalable approach to evaluate open-ended text generation (Li et al., 2023; Zheng et al., 2023; Li et al., 2024; Dubois et al., 2024). Essentially, two competing LLMs provide responses to a set of test cases (prompts), and they are subsequently evaluated by a "judge" LLM to decide the win rate.

Currently, CamelEval uses GPT-4o by default due to its strong multi-lingual capabilities. We have also investigated Self-taught Llama 3.1 70B as the judge (Wang et al., 2024). However, manual inspection on the annotation reveals that this judge has limited capability in Arabic language despite achieving high agreement with humans on English test cases. Building a tailored LLM judge in Arabic contexts is an interesting venue for future works.

3.3 Bias Correction

Positional bias and response-length bias are well-known biases that could significantly impact the quality and objectivity of the LLM-as-judge benchmarks (Zheng et al., 2023; Koo et al., 2023; Wang et al., 2023; Wu and Aji, 2023). CamelEval incorporate techniques to correct these biases.

Randomization to mitigate positional bias. It has been observed that the order of the two responses presented to the LLM judge impacts the ranking (Li et al., 2023; Zheng et al., 2023). To mitigate this bias, we randomly shuffled the order of two responses during evaluation such that each model have a 50% chance to be presented first.

Regression to correct response length bias. Prior work has pointed out that the LLM judge tends to prefer longer responses. To account for this bias, CamelEval adopts the regression-based length control (Dubois et al., 2024), which has a firm grounding in Causality (Hernán and Robins, 2010). CamelEval fits a generalized linear model (GLM) of the judge preference based on model identity and causal confounders (response length and task difficulty). The GLM then predicts the counterfactual preference if the responses were equally long. The instruction difficulty was annotated during our data curation process using GPT-4o (Appendix A.3).

4 Evaluation Results and Insight

4.1 Evaluation Setup

We included 20 instruction-finetuned LLMs in the evaluation, covering popular model families such as Gemma, Qwen, Llama, Jais, Aya, and GPT-4. These LLMs capture a diverse range of language focuses, model sizes, and development setups, summarized in Appendix A.4.

We used Gemma2-9b-IT as the baseline to calculate the win rate. This is because we found that the

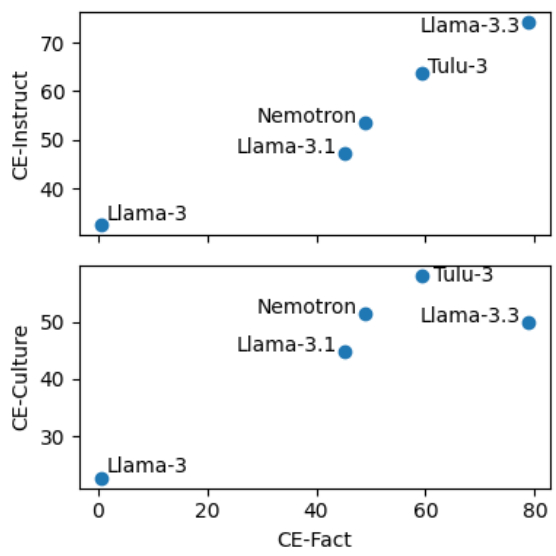

Figure 3: Win rate of 70B Llama-3 releases and finetunes on the three CamelEval test suites.

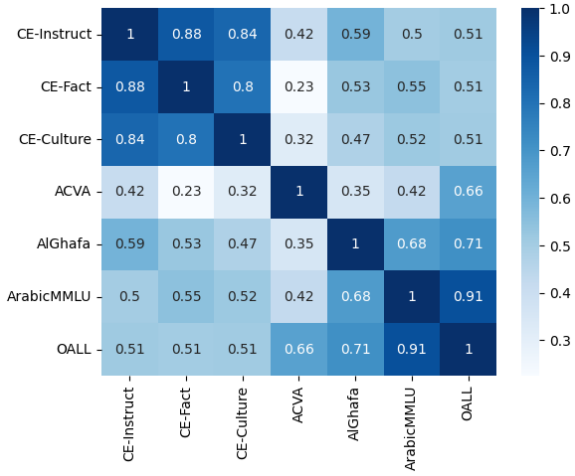

Figure 4: Spearman correlation between different benchmark suites.

response length of Gemma2-9b-IT is close to the average length across all models, making length correction more stable (Appendix A.7). We used the same set of hyperparameters for LLM inference (Appendix A.5).

4.2 CamelEval Evaluation Insights

Family-wise model comparison. In Figure 2, we observe that Qwen-2.5 and Gemma-2 family of LLMs generally perform well on CamelEval across different model sizes. GPT-4o achieves the overall best performance. Jais-70B performed well in the Culture suite but lagged behind on other test suites. Llama-3.1 is lagging behind across the board but Tulu-3, a finetuned version of Llama-3.1, achieves good performance among 70B models.

Preservation of LLM scaling laws. Figure 2 shows that LLMs with more parameters generally perform better on CamelEval than the ones with fewer parameters. This trend is persistent across all three evaluation suites and different model families. The observation is in agreement with the empirical scaling law of LLM parameters and provides a sense check on the validity of CamelEval.

Performance of Llama-3 releases and finetunes. Figure 3 illustrates the performance of the 70B-sized Llama 3 series (3, 3.1, 3.3) and finetunes (Nemotron and Tulu-3). We see a clear trend of improvement on all suites over the three Llama releases. Both finetuned version significantly improved on their base model (Llama 3.1), even matching or exceeding Llama-3.3 on Culture.

4.3 Correlation Analysis

We illustrate the Spearman correlation between the Arabic LLM benchmarks in Figure 4. Spearman correlation is used to capture the rankings of different LLMs, which is often reported in leaderboards.

We observe that the benchmarks form two main clusters. One includes the three test suites in CamelEval, the other includes AlGhafa, ArabicMMLU, and OALL. The ACVA benchmark appears to be uncorrelated with other benchmarks. The cluster structure is reasonable because the three CamelEval suites evaluate LLM's ability to generate responses whereas AlGhafa, ArabicMMLU, and OALL adopt MCQ to measure LLM's discriminative capabilities. We also note that the test suites in CamelEval are not perfectly correlated with each other (Spearman correlation between 0.80 and 0.88) as they capture different aspects of the generation.

5 Discussion

CamelEval is a benchmark for evaluating instruction following, factuality, and culture alignment of Arabic LLMs in generating responses. It serves as a complement to the existing benchmarks which focus on discriminative tasks like MCQ. We aspire that CamelEval will assist the community in advancing the creation of improved and more culturally attuned Arabic LLMs. While this work attempts to reduce the risk of using Arabic LLM in conversational settings, it does not capture all known LLM risk modalities such as toxicity, safety, or adversarial attacks. The community needs to interpret the CamelEval results carefully to avoid those potential risks.

6 Limitations

We note that the typical constraints associated with using LLMs as evaluators also apply to CamelEval. For example, there's a possibility that the judge LLM might show a preference for answers it generates itself. Furthermore, the current version of CamelEval focuses on helpfulness of the LLM and does not cover the harmlessness or safety aspects of LLMs. We aim to tackle these and other unresolved challenges in future updates of CamelEval.

References

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. 2023. AlGhafa evaluation benchmark for Arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275, Singapore (Hybrid). Association for Computational Linguistics.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran AQ Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, et al. 2024. Cidar: Culturally relevant instruction dataset for arabic. *arXiv preprint arXiv:2402.03177*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Ali Elfilali, Hamza Alobeidli, Clémentine Fourrier, Basma El Amel Boussaha, Ruxandra Cojocaru, Nathan Habib, and Hakim Hacid. 2024. Open arabic llm leaderboard. https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard.

Ethnologue. Ethnologue arabic statistics. https://www.ethnologue.com/language/ara/. Accessed: 2024-09-11.

FreedomIntelligence. 2024. Acva arabic cultural value alignment. https://huggingface.co/datasets/FreedomIntelligence/ACVA-Arabic-Cultural-Value-Alignment.

Miguel A Hernán and James M Robins. 2010. Causal inference.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2023. Acegpt, localizing large language models in arabic. *Preprint*, arXiv:2309.12053.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv:2402.12840*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. *arXiv preprint arXiv:2102.00287*.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Appendix

A.1 Limitations of OALL

OALL is instrumental in evaluating text completion capabilities, logical correctness, and factual knowledge across different domains. However, it has several key limitations:

1. Narrow coverage of LLM Capabilities. The benchmark's reliance on multiple-choice questions means it fails to evaluate the broader spectrum of LLM capabilities, such as engaging in conversations or following instructions. It does not measure the helpfullness or the utility of the LLM's replies, which are essential aspects of its performance. In fact, the ability to participate in general conversations is a defining feature of LLMs. Consequently, while this benchmark is effective for assessing the foundational knowledge and reasoning skills of pretrained LLMs, it does not adequately measure the performance of LLMs that have been instruction-finetuned for generating meaningful interactions with users.

2. Oversimplified evaluation metric. The evaluation metric used by OALL, the normalized log-likelihood (NLL), is overly simplistic. NLL calculates the log-probability of producing the "gold response," adjusted for the length of this ideal response. However, the assumption that there's a singular "gold response" is flawed, even in contexts like multiple-choice questions. This inconsistency is apparent in OALL itself, where some correct answers are labeled as A, B, C, or D, and others are identified by the text of the correct option[1]. The variability in defining what constitutes a "gold response" renders NLL an unreliable and imprecise metric for LLMs, which can generate texts in diverse formats and styles.

3. Translation issues. In addition, OALL suffers from translation issues, some examples are listed bellow:

Example One[2]:

انطفأت الشعلة الموجودة على الشمعة.، لأن :

٠) لقد فجرت على الفتيل.

١) لقد وضعت عود ثقاب على الفتيل.

Example Two:

شرب الرجل بكثرة في الحفلة.، لذلك :

٠) كان يعاني من الصداع في اليوم التالي.

١) أصيب بسيلان في الأنف في اليوم التالي.

Example Three:

أخطأ الطالب في كتابة الكلمة.، لذلك :

٠) قام المعلم بتصحيحها.

١) طردتها المعلمة.

Example Four:

امرأة ببدلة حمراء تتحدث خلف مكتب. رجل الاقتراحات:

٠) يأخذ كرات من حامل على الجدار.

١) يقوم بتنظيف كرسي أسود بالمكنسة الكهربائية.

٢) يرش الماء في حوض أبيض صغير.

٣) يقف ويحمل قدمه خلف ظهره.

[1] Adding few-shot examples may help alleviate the arbitrariness of "gold response", but OALL employees zero-shot evaluation in all cases (Elfilali et al., 2024).

[2] We have right-aligned all Arabic text to conform to the language's standard writing style.

A.2 Building A Textbook-quality Corpus

Collecting representative Arabic evaluation datasets was one of the main bottlenecks for developing a better benchmark. Most of the open Arabic datasets are translated from other languages and are subject to translation biases or fail to reflect cultural context appropriately (Vanmassenhove et al., 2021; Stanovsky et al., 2019; Wang et al., 2022; Naous et al., 2023). Moreover, regional datasets are extremely scarce and with limited quality (Singh et al., 2024; Alyafeai et al., 2024).

To systematically collect data, we first identified a comprehensive list of subject categories, such as "Science" and "Humanities." For each subject, we further delineated sub-categories, such as "Physics" and "History." The subject categories are inspired by library classification, specifically Universal Decimal Classification, which is a standard to categorize books and documents. We have prioritized categories that are relevant to the general audience. For each identified category, we conducted a thorough search to locate and collect relevant datasets. Our search methodology encompassed different data sources including open web search, internal datasets, and translatable open-licensed datasets.

A.2.1 Criteria for Data Search

During the search, we considered a variety of criteria as shown in the following list:

1. Relevance to Topic Criteria: The data must be directly related to our subject categories.
2. Timeliness Criteria: The data should be up to date.
3. Completeness Criteria: The dataset should be comprehensive enough to support robust analysis.
4. Granularity Criteria: The data should have the appropriate level of detail.
5. Availability and Accessibility Criteria: The data should be accessible and has an open license.
6. Bias and Objectivity Criteria: The data should be free from bias or, if biased, the bias should be understood and accounted for.
7. Cost: The estimated cost for accessing and curating the data.

We have verified the license of all the contents sourced from internet and we have only retained the content under a permissive license for LLM research and development.

We used a tagging system to annotate various aspects of the data and flagged data with high uncertainty for human review. We prioritized Modern Standard Arabic (MSA) data, ensured sources were reputable, and preferred cleaned data. We collected the resulting dataset into a textbook-like corpus for further processing.

A.2.2 Data Cleaning Checklist

We create a data cleaning checklist in Table 2 and check all the items to ensure the data quality.

Data Quality Checklist
Duplicate or near-duplicate data
Missing data and other artifacts due to web scraping
Artifacts introduced by translation
Data in out-of-scope languages
Ill-formatted code blocks or structured text
Irrelevant system prompts
Unbalanced mix of tasks, categories, or difficulty

Table 2: Common data quality issues with multi-lingual datasets.

A.2.3 Data Cleaning Statistics

In Table 3, we report the number of documents removed by each of the criteria in the checklist. After the cleaning pipeline, we have balanced the dataset by randomly selecting 200 documents for each subject, where each document has between 100 to 1200 words. From each document we generated 10 questions, and sampled 805 from the total number of the generated questions.

Criterion	# Removed (K)
Duplicates	15
Missing data	6
Translation	10
Non-Arabic	9
Code	60
Emoji	4
Invalid prompts	1

Table 3: Number of documents (in thousands) removed for each criterion.

The subject categories in CamelEval CE-Fact

subset are listed in Table 4. The prompts cover a wide range of topics that are science and culture related.

Category	Occurrence
Physics	66
Biographic	65
Economy	62
Philosophy	61
Geology	61
Psychology	60
Nutrition	59
Chemistry	58
Education	56
Arts	54
Medicine	53
Arabic	50
Math	39
Engineering	24
Tech	16
Anthropology	14
Infographic	3
Environment	2
History	2

Table 4: Subject categories in CamelEval Factuality Suite.

Some prompts of the CE-Culutre subset are shown in Listing 5. It is clear that, answering these questions requires a good understanding of the Arabic culture and the regional nuances.

A.3 Prompt Template for Question Generation

The template used to generate the questions is shown in Listing 6. It has four main sections instruction, text, claims, and questions.

A.4 Evaluated models

The list of evaluated models and some of their properties are listed in Table 5.

A.5 Hyperparameter for LLM Inference

We have used Nucleus sampling with temperature $\tau = 0.8$ and top-$p = 0.95$ for all LLMs to generate responses (Holtzman et al., 2019). In our pilot studies, we found that this sampling configuration effectively reduces the chance of generating endless repetitive contents.

Model	Sizes (B)	Arabic Support
Llama 3	70	Unofficial
Llama 3.1	8, 70	Unofficial
Llama 3.3	70	Unofficial
Tulu-3	8, 70	Unofficial
Nemotron	70	Unofficial
Gemma-2	9, 27	Unofficial
Qwen-2	72	Unofficial
Qwen-2.5	3, 14, 32, 72	Official
Jais	7, 13, and 70	Official
Aya	7 and 35	Official
GPT-4	≈ 175	Official

Table 5: Evaluated fine-tuned models, their sizes, and Arabic support.

A.6 Numerical Results

The numerical results of the evaluated models are shown in Table 6.

A.7 Choice of Baseline Model for Response Length Bias Adjustment

In the process of adjusting for response length bias, it's crucial to select a baseline model that closely mirrors the target model. This approach significantly reduces the necessity for extensive correction, thereby enhancing the stability of the adjustment process. Essentially, if the baseline and target models already produce responses of similar lengths, only minimal adjustments are required.

Considering this, it is advantageous to select a baseline model that aligns closely with not just one, but all target models. An effective strategy is to opt for a baseline model that approximates the average response length across the board. To illustrate this, we analyzed the response lengths produced by various models using CamelEval, as depicted in Figure 7. Our analysis revealed that the Gemma2-9B-IT model generates responses that most closely match the average length observed across all models. Consequently, we have chosen Gemma2-9B-IT as our baseline model for adjustments.

ـ ما هو التاريخ الثقافي وراء اعتبار العودة إلى الوطن جزءًا من الاحتفالات العربية؟
ـ كيف يختلف تقديم الهدايا في الثقافة العربية عن الثقافات الأخرى؟
ـ هل توزع الورود على الحاضرين في الجنازات في جميع الديانات الموجودة في العالم العربي؟
ـ هل للألوان الفاتحة في زي الفلاحين التقليدي في المغرب أبعاد دينية أو رمزية؟
ـ اذكر خمس وصفات لإستخدام الحلبة في المأكولات العربية؟
ـ هل تلعب الهدايا اليدوية دورًا في رمز المظهر الاجتماعي في حفلات الزفاف العربية؟
ـ هل هناك أهمية ثقافية لاستخدام الظروف في المناسبات الاجتماعية في المجتمعات العربية؟
ـ كيف تتغير عادات الجنازة بين المجتمعات الريفية والحضرية في الدول العربية؟
ـ ما هي التحديات التي قد يواجهها المنشدون في بدء مسيرتهم في الدول العربية؟
ـ هل هناك اختلافات إقليمية في أسلوب النحت بين مناطق العالم العربي؟
ـ ما اكثر الدول العربية تأثرا بالمطبخ الأوربي؟
ـ اي الدول العربية تشتهر بتحضير المرقوق؟
ـ كيف يمكن مقارنة تفاعل الجنسين في الأماكن العامة في الدول العربية بدول أخرى في العالم؟
ما هو الفرق بين رأي الحنفية والحنابلة فيما يتعلق بجوزة الطيب؟

Translation:

- What is the cultural history behind the return home being part of Arab celebrations?
- How does gift-giving in Arab culture differ from other cultures?
- Are flowers distributed to attendees at funerals in all religions existing in the Arab world?
- Do bright colors in the traditional dress of Moroccan peasants have religious or symbolic connotations?
- Mention five uses of fenugreek in Arab cuisine?
- Do hand-made gifts play a role in the symbolism of social status at Arab weddings?
- Is there a cultural significance to the use of envelopes for social occasions in Arab societies?
- How do burial practices differ between rural and urban communities in Arab countries?
- What challenges may vocalists face at the beginning of their careers in Arab countries?
- Are there regional differences in sculpture styles between regions of the Arab world?
- Which Arab countries are most influenced by European cuisine?
- Which Arab country is famous for preparing "mergouk"?
- How can the interaction of genders in public places in Arab countries be compared to that in other countries in the world?
- What is the difference between the Hanafi and Hanbali views on nutmeg?

Listing 5: Examples of culture-centered prompts from CamelEval.

Instruction
You will be given a text, ANALYZE it and EXTRACT "atomic claims" then WRITE {num_of_questions} questions in Arabic, such that they CANNOT be answered from the text only and require extra knowledge. MAKE sure the questions are grammatically and semantically correct, and USE the provided template. AVOID any questions or claims related to text summarization such as:

ـ ما هو الهدف الرئيسي من هذا النص؟
ـ ما هو السياق العام للنص؟
ـ ما هي النقاط الرئيسية التي تمت مناقشتها في النص؟
ـ ما هي الاساليب الادبية المستخدمة في النص لتحسين فعاليته؟
ـ ما هي الأهمية الثقافية لأحداث النص المفقود؟
ـ من هم الشخصيات الرئيسية في النص المفقود؟
ـ ما هي الرسالة الأساسية للنص المفقود؟
ـ ما هو السياق التاريخي للنص المفقود؟
ـ كيف يمكن مقارنة النص المفقود بأعمال أخرى في نفس المجال؟

Text
"""{input}"""
PUT your response in the following JSON format:
Claims
"claims":[{{ "claim": "claim in your own words in arabic", "reference": "part of the text that supports the claim"}}]
Questions
"questions":[{{ "question": "put the question here", "answer": "your answer of the question in less than 100 words" "difficulty": one of these ["easy", "medium", "hard", "very hard", "extermly hard"], "reason": "reason why you have selected that difficulty"}}]

Table 6: The prompt template used for question generation.

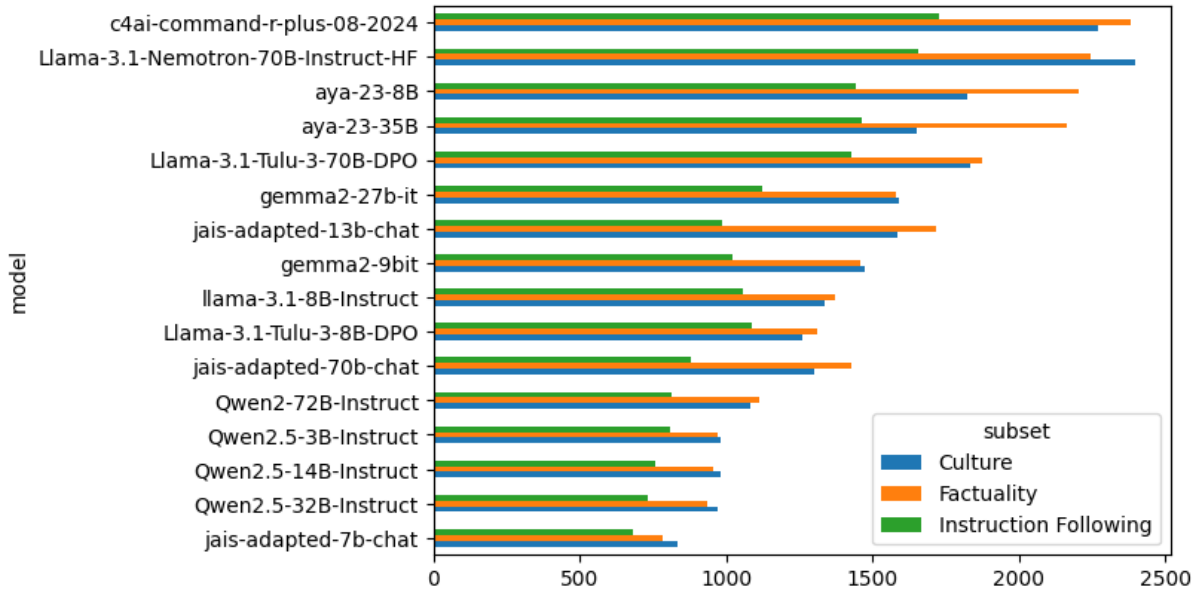

Figure 7: Visualization of response length by different models on CamelEval.

Model	CE-Instruct		CE-Fact		CE-Culture	
	LC-WR%	Std	LC-WR%	Std	LC-WR%	Std
Aya-23-8B	38.72	0.11	43.78	0.10	36.91	0.05
Aya-23-35B	51.29	0.05	49.16	0.09	43.58	0.08
Gemma-2-9B-IT	50.00	0.00	50.00	0.00	50.00	0.00
Gemma-2-27B-IT	56.87	0.01	53.88	0.02	55.43	0.04
GPT-4o-2024-05-13	73.58	0.00	65.12	0.12	65.62	0.04
Jais-adapted-7B-chat	21.89	0.00	25.12	0.01	25.26	0.00
Jais-adapted-13B-chat	36.69	0.00	42.63	0.14	46.84	0.07
Jais-adapted-70B-chat	48.20	0.13	49.44	0.06	53.73	0.08
Llama-3.1-8B-Instruct	19.70	0.14	18.91	0.01	18.70	0.06
Llama-3.1-70B-Instruct	47.40	0.00	45.29	0.12	44.80	0.00
Llama-3-70B-Instruct	32.63	0.14	0.52	0.00	22.67	0.00
Llama-3.3-70B-IT	74.22	0.03	78.91	0.11	50.00	0.08
Llama-3.1-Nemotron-70B-Instruct-HF	53.49	0.12	49.00	0.06	51.58	0.06
Llama-3.1-Tulu-3-70B-DPO	63.76	0.02	59.43	0.09	58.16	0.02
Llama-3.1-Tulu-3-8B-DPO	31.79	0.01	36.82	0.15	37.61	0.03
Qwen2-72B-Instruct	58.11	0.06	65.29	0.10	54.09	0.14
Qwen2.5-14B-Instruct	49.50	0.00	63.37	0.04	49.56	0.05
Qwen2.5-32B-Instruct	53.70	0.00	63.59	0.05	51.09	0.05
Qwen2.5-3B-Instruct	17.75	0.11	23.41	0.07	17.11	0.00
Qwen2.5-72B-Instruct	65.16	0.06	61.97	0.10	60.75	0.06

Table 6: Performance of some notable Arabic-centric or multilingual LLMs on CamelEval. We report the length controlled win-rate (LC-WR) against the Gemma-2-9B-IT model.