Towards Inclusive Arabic LLMs: A Culturally Aligned Benchmark in Arabic Large Language Model Evaluation

Omer Nacar[1], Serry Sibaee[1], Samar Ahmed[2], Safa Ben Atitallah[1], Adel Ammar[1],
Yasser Alhabashi[1], Abdulrahman S. Al-Batati[1], Arwa Alsehibani[1], Nour Qandos[2], Omar Elshehy[3],
Mohamed Abdelkader[1], Anis Koubaa[1]
[1]Robotics and Internet-of-Things Lab, Prince Sultan University, Riyadh 12435, Saudi Arabia
[2]Independent Researcher [3]Universität des Saarlandes, Saarbrücken, Germany
{onajar, ssibaee, satitallah, aammar, yalhabashi, aalbatati, Aalsehibani, mabdelkader, akoubaa}@psu.edu.sa
Samar.sass6@gmail.com, nooramerq0@gmail.com, omar.elshehy@physik.uni-saarland.de

Abstract

Arabic Large Language Models are usually evaluated using Western-centric benchmarks that overlook essential cultural contexts, making them less effective and culturally misaligned for Arabic-speaking communities. This study addresses this gap by evaluating the Arabic Massive Multitask Language Understanding (MMLU) Benchmark to assess its cultural alignment and relevance for Arabic Large Language Models (LLMs) across culturally sensitive topics. A team of eleven experts annotated over 2,500 questions, evaluating them based on fluency, adequacy, cultural appropriateness, bias detection, religious sensitivity, and adherence to social norms. Through human assessment, the study highlights significant cultural misalignment and biases, particularly in sensitive areas like religion and morality. In response to these findings, we propose annotation guidelines and integrate culturally enriched data sources to enhance the benchmark's reliability and relevance. The research highlights the importance of cultural sensitivity in evaluating inclusive Arabic LLMs, fostering more widely accepted LLMs for Arabic-speaking communities.

1 Introduction

Arabic, spoken by over 400 million people, ranks among the world's most widely used languages UNESCO. Despite its global prominence, Arabic has received limited attention in NLP research, classifying it as a low-resource language Magueresse et al. (2020). Consequently, Arabic NLP models, particularly large language models, are often evaluated on translated datasets that fail to capture the language's rich cultural context Guellil et al. (2021). This reliance on culturally detached benchmarks has led Arabic LLMs to frequently exhibit biases and misalignment, diminishing their effectiveness and cultural adequacy, especially in areas that require cultural sensitivity. Given that culture fundamentally shapes communication and social norms Masoud et al. (2023), it is essential for LLMs to authentically reflect these nuances to better serve Arabic-speaking communities.

The reliance on culturally misaligned benchmarks creates a problematic feedback loop: models trained and evaluated on such data are less likely to handle culturally sensitive or nuanced topics, as they are never adequately assessed for these capabilities. Consequently, Arabic LLMs may perform well on technical metrics yet fail to resonate with the cultural values and expectations of their target audience Cao et al. (2023); Navigli et al. (2023). This disconnect reduces trust in the model's outputs, limiting its usefulness for Arabic-speaking users and decreasing wider acceptance of Arabic LLMs Blasi et al. (2021). Bridging this benchmarking gap is essential for creating linguistically accurate and culturally relevant Arabic resources.

To address these challenges, our study undertakes a comprehensive evaluation of the Arabic Massive Multitask Language Understanding (MMLU) Benchmark Hendrycks et al. (2020), a widely recognized benchmark with multiple Arabic versions, including machine-translated using GPT-3.5-Turbo Model Huang et al. (2023) and human-translated provided by Openai[1]. The MMLU Benchmark has gained popularity for evaluating LLMs due to its extensive coverage of 57 topics across various fields, providing a robust framework for assessing a model's general knowledge and adaptability across domains.

This study emphasizes the critical need to prioritize cultural alignment in the development and evaluation of Arabic LLMs. By focusing on benchmarks and methodologies that reflect the linguistic and cultural intricacies of Arabic-speaking communities, our work aims to advance the creation of more inclusive and contextually accurate language

[1]https://huggingface.co/datasets/openai/MMMLU

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 387–401
January 20, 2025.

technologies. This approach underscores the importance of moving beyond technical performance metrics to ensure that Arabic LLMs are both culturally resonant and widely trusted by their users.

2 Related Work

Research on cultural values in AI emphasizes designing systems that respect user cultural contexts for improved social acceptability and effectiveness. Studies highlight challenges in culturally aligning language models (LLMs) trained on English language datasets, which may overlook the values of other cultural contexts.

Jinnai (2024) explores Japanese LLMs aligned with English datasets, finding limitations in capturing Japanese moral frameworks and calling for culturally tailored Japanese data. Yuan et al. compares AI responses between Chinese and English, revealing biases that underscore the need for culturally aware AI design with continuous monitoring. Tao et al. (2024) evaluates cultural bias across major LLMs, noting they often reflect Protestant European cultural norms and proposing "cultural prompting" to enhance alignment with diverse regions, though scarce language data remains a challenge.

Koto et al. (2024) introduced ArabicMMLU, an Arabic dataset with 14,575 questions across 40 tasks to evaluate Arabic language models, enhancing comprehension in North African and Levantine contexts. Qian et al. (2024) presents Juhaina, an Arabic-English bilingual LLM, paired with CamelEval, a benchmark for assessing cultural relevance in Arabic LLM responses. Zhu et al. describes AceGPT-v1.5, which improves Arabic vocabulary handling through progressive vocabulary expansion, enhancing text comprehension and cultural alignment for Arabic users.

Our study focuses on six culturally misaligned topics—human sexuality, moral disputes, moral scenarios, philosophy, world religions, and professional psychology—where cultural sensitivity is particularly critical. To further enhance the benchmark's cultural relevance, we introduced five additional topics uniquely significant to Arabic-speaking communities: Islamic religion, Old Arab history, Islamic history, Arabic ethics, and Arabic educational methodologies. A team of eleven experts reviewed over 2,500 questions across these domains, applying detailed criteria covering fluency, adequacy, cultural appropriateness, bias detection, religious sensitivity, and adherence to social norms. This comprehensive evaluation highlights significant cultural misalignments and biases, prompting the development of annotation guidelines and the incorporation of culturally enriched data sources to improve the benchmark's reliability.

3 Methodology

In this work, we critically examine the Arabic MMLU Benchmark, focusing on its cultural alignment and relevance for evaluating Arabic Large Language Models (LLMs). The original MMLU Benchmark is in English, has since been translated into Arabic in two versions: one by GPT-3.5 Turbo and another by Arabic native human translators, both of which are widely used to assess the capabilities of Arabic LLMs. Figure 1 presents the various topics included in the MMLU benchmark, categorized by their level of cultural alignment sensitivity. The identified Critical Misalignment topics frequently lack alignment with Arabic cultural norms and values, potentially leading to inaccurate or culturally insensitive outputs in Arabic language models.

The Arabic MMLU Benchmark includes over 700 questions on Western-centric topics, such as European and U.S. History and U.S. Foreign Policy, which lack cultural relevance for Arabic-speaking communities, rendering them unsuitable for cultural alignment assessments. To address this, we implemented a comprehensive evaluation framework encompassing linguistic and cultural dimensions. Linguistic metrics include Fluency (naturalness and grammatical correctness) and Adequacy (faithfulness in conveying the source text's meaning), both rated on a 1–5 scale. For cultural alignment, we introduced four metrics: Cultural Appropriateness (sensitivity to cultural nuances), Bias Detection (presence of various bias types), Religious Sensitivity (respect for religious beliefs), and Social Norms (adherence to societal values), each carefully scored or annotated.

Alongside human evaluation metrics, we employed several automated metrics to quantify translation quality and similarity. These include BLEU Papineni et al. (2002), ROUGE Lin (2004), METEOR Banerjee and Lavie (2005), chrF Popović (2015), BERTScore Zhang et al. (2019), and COMET Rei et al. (2020), which provide insights into linguistic accuracy and fluency. By combining these automated metrics with hu-

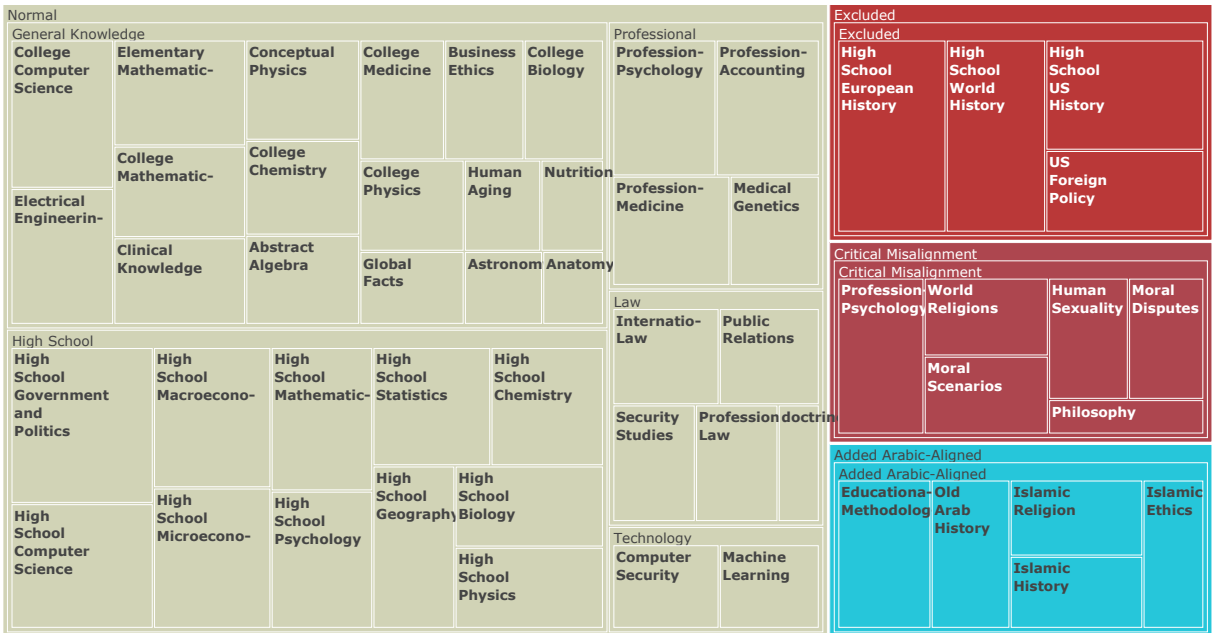


Figure 1: Arabic MMLU Benchmark Topics: General, Excluded, Added, and Culturally misalignment Topics

man evaluations, we established a rigorous and multidimensional framework to support a comprehensive analysis of the benchmark's cultural and linguistic suitability for Arabic-speaking communities. This approach allows us to identify key areas of misalignment and provides valuable insights for enhancing Arabic NLP models' cultural sensitivity and reliability.

Lastly, to facilitate a standardized evaluation for Arabic LLMs, we created the Index for Language Models for Arabic Assessment on Multitasks (ILMAAM)[2], a dedicated leaderboard that benchmarks performance on the refined Arabic MMLU, excluding culturally sensitive topics assessed for alignment. ILMAAM serves as a reliable measure of model accuracy across non-critical topics, providing transparency and consistency in Arabic LLM evaluation.

The refined dataset addresses linguistic and cultural misalignments identified in the Arabic MMLU Benchmark. The updated version, which includes culturally enriched questions, is publicly available on Hugging Face[3].

[2] https://huggingface.co/spaces/Omartificial-Intelligence-Space/Arabic-MMMLU-Leaderborad

[3] https://huggingface.co/datasets/Omartificial-Intelligence-Space/ILMAAM-Arabic-Culturally-Aligned-MMLU

4 Annotation Process

The annotation methodology involved eleven trained Arabic-language experts who independently assessed question subsets to ensure coverage and consistency. Three annotators evaluated each topic, with quality checks by researchers to uphold accuracy and guideline adherence. This approach promoted high inter-annotator reliability, minimizing subjectivity for robust evaluations.

The cultural alignment assessment was structured to identify subtle and overt cultural misalignments through a multi-step procedure. Annotators evaluated fluency, adequacy, cultural appropriateness, and sensitivity using predefined metrics, with regular consensus meetings to refine interpretations. This framework systematically captured cultural biases, offering a comprehensive cultural assessment. For detailed guidelines, see Appendix A.

5 Results

Our evaluation of the Arabic MMLU Benchmark identifies key issues in three areas: Cultural, Methodological and Structural, and Linguistic (Figure 5). Cultural Issues include deficiencies in representing Philosophical and Ethical Foundations and Language and Expressions, leading to content that may feel culturally misaligned or insensitive for Arabic-speaking users. Methodological and Structural Issues reveal inadequacies in structural design and source relevance, affecting content clarity and

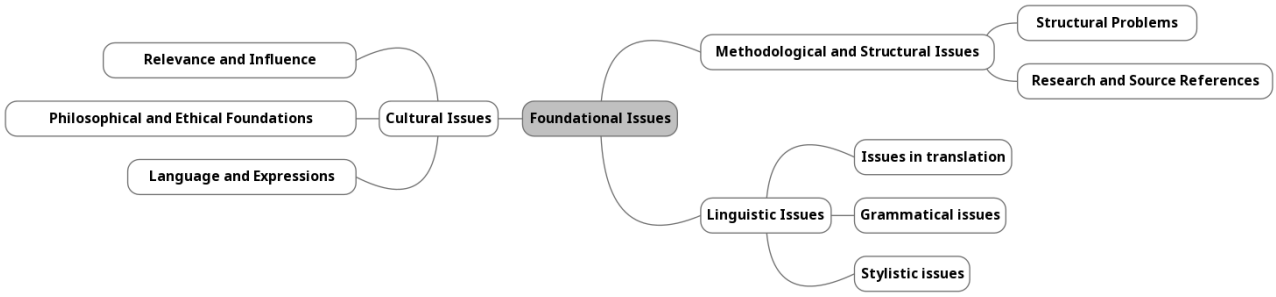


Figure 2: Foundational Issues of Cultural Misalignment in the Arabic MMLU Benchmark

coherence. Linguistic Issues highlight translation problems, including grammatical and stylistic errors that reduce readability and authenticity.

These findings emphasize the need for a culturally aligned evaluation framework and refined translation methods. Subsequent sections provide detailed analyses, including translation scores, similarity metrics, and reviewer assessments.

5.1 Translation Quality Metrics

To evaluate the translation quality between human-translated and GPT-translated versions of the Arabic MMLU Benchmark, we used a range of automated metrics, including BLEU, ROUGE, METEOR, chrF, BERTScore, and COMET. Table 1 presents the results for culturally critical topics, such as human sexuality, moral disputes, and philosophy, alongside excluded Western-centric topics like U.S. history and European history, which lack cultural alignment for Arabic-speaking audiences.

As shown in Table 1, the critical cultural topics generally scored higher, with philosophy achieving a notable BERTScore of 0.884 and COMET score of 0.861, reflecting strong semantic alignment between human and GPT translations. In contrast, topics like High School U.S. History and European History displayed lower performance, with near-zero BLEU scores and lower scores across other metrics, suggesting challenges in achieving accurate and contextually relevant translations for these subjects.

Metrics such as ROUGE, METEOR, and chrF further reinforced these findings, showing consistently higher scores for topics involving complex ethical or psychological content (e.g., moral scenarios, professional psychology), while historically Western-centric subjects tended to score lower across metrics. These results highlight the variability in translation quality across different subject areas, underscoring the importance of topic-specific evaluation metrics to accurately gauge translation fidelity in Arabic-language LLM benchmarks.

5.2 Human Evaluation Metrics

To assess the translation quality and cultural sensitivity of the Arabic MMLU Benchmark, we conducted a comprehensive human evaluation across six essential metrics: fluency, adequacy, cultural appropriateness, bias detection, religious sensitivity, and social norms. The evaluation was applied to six culturally sensitive topics, including human sexuality, moral disputes, moral scenarios, philosophy, world religions, and professional psychology. Figure 3 presents these findings, highlighting key areas of cultural alignment and misalignment across topics.

As shown in Figure 3, there are significant cultural challenges in certain areas. For example, human sexuality shows moderate scores in fluency at 3.78 and adequacy at 4.21, but it significantly lags in cultural appropriateness at 3.26 and religious sensitivity at 2.18. This topic also has a high bias detection rate of 65.5 percent, underscoring substantial cultural misalignment. Similarly, world religions, while achieving high scores in fluency at 4.82 and adequacy at 4.85, reveal major issues with cultural appropriateness at 2.71 and have the highest bias detection rate at 78.62 percent, indicating strong cultural dissonance.

In contrast, some topics demonstrate better cultural alignment. Moral scenarios score well in both fluency at 4.32 and adequacy at 4.34 and have a balanced cultural appropriateness score of 3.05, with a relatively low bias detection rate of 10.05%, reflecting minimal cultural bias. Professional psychology performs better with cultural appropriateness at 4.75 and religious sensitivity at 4.85 and a low bias detection rate of 7.51 percent, indicating better

Topic	BLEU	ROUGE	METEOR	chrF	BERTScore	COMET
high_school_european_history	0.0000024	0.144	0.018	4.461	0.669	0.532
high_school_us_history	0.0000021	0.180	0.023	4.530	0.665	0.505
high_school_world_history	0.0000138	0.240	0.029	5.612	0.678	0.544
human_sexuality	0.222	0.035	0.376	46.695	0.841	0.816
moral_disputes	0.250	0.008	0.440	55.401	0.868	0.837
moral_scenarios	0.356	0.937	0.578	61.078	0.853	0.769
philosophy	0.329	0.000	0.497	56.236	0.884	0.861
professional_psychology	0.234	0.089	0.400	49.992	0.849	0.823
us_foreign_policy	0.314	0.060	0.533	64.544	0.882	0.899
world_religions	0.199	0.019	0.398	54.489	0.867	0.853

Table 1: Translation Metrics for Arabic MMLU Comparing Human Translations to GPT MMLU on Culturally Critical and Excluded Misaligned Topics

alignment with Arabic cultural expectations.

In addition to culturally sensitive topics identified within the original Arabic MMLU benchmark, our study introduced five new topics specifically relevant to Arabic-speaking communities: Islamic religion, Old Arab history, Islamic history, Islamic ethics, and educational methodologies. Figure 4 displays the number of questions added across five culturally significant topics. These additions ensure a more comprehensive cultural representation and allow for a nuanced evaluation of Arabic LLMs in areas central to the Arabic-speaking world. The distribution of questions within these topics varies, with Islamic ethics containing the highest number of questions at 188, followed by Old Arab history with 168 and Islamic history with 160. Islamic religion and educational methodologies have 136 and 114 questions, respectively. By incorporating these culturally significant areas, the evaluation framework is better equipped to assess the cultural alignment and sensitivity of Arabic language models, addressing gaps that were previously overlooked in standard Western-oriented benchmarks.

5.3 ILMAAM Leaderboard Results

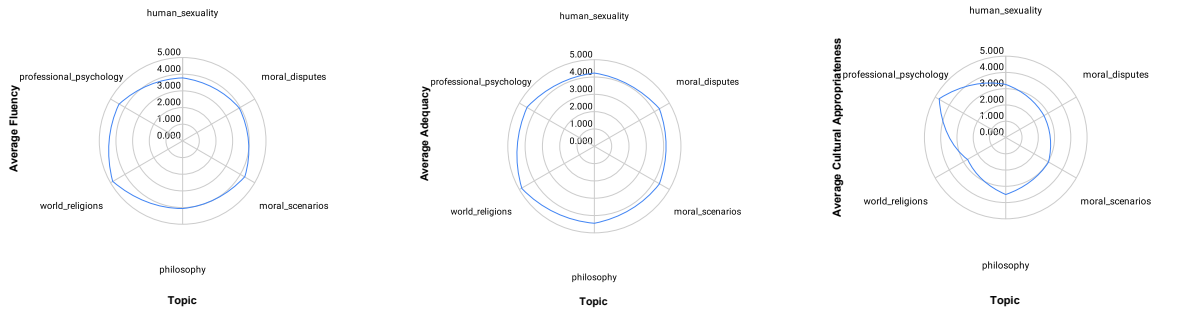
The ILMAAM leaderboard offers a comprehensive performance overview of 31 Arabic LLMs on the refined Arabic MMLU Benchmark, showcasing each model's strengths and weaknesses through average accuracy scores. Table 2 presents the results for the top-performing models, averaged across various topics, excluding culturally sensitive ones. For a comprehensive view of ILMAAM results, see Appendix D, which lists the performance of 30 Arabic LLMs on the culturally refined benchmark.

As shown in Table 2, the ILMAAM leaderboard results highlight significant variation in performance across Arabic LLMs, emphasizing the impact of model size and tuning approach on accuracy. Larger models, such as *Qwen/Qwen2.5-72B-Instruct* and *CohereForAI/aya-expanse-32b*, lead with the highest average scores of 73.45 and 63.87, respectively, indicating that increased parameters often correlate with improved accuracy on the Arabic MMLU benchmark. Instruction-tuned models generally perform better, with Qwen models occupying multiple top spots, suggesting that instruction tuning enhances cultural and linguistic understanding in Arabic tasks. Pretrained models, while generally strong, show slightly lower scores, such as *CohereForAI/aya-expanse-8b* at 51.79. This variation underscores the importance of model customization for optimal performance in culturally nuanced evaluations, affirming ILMAAM's value in benchmarking Arabic LLM capabilities.

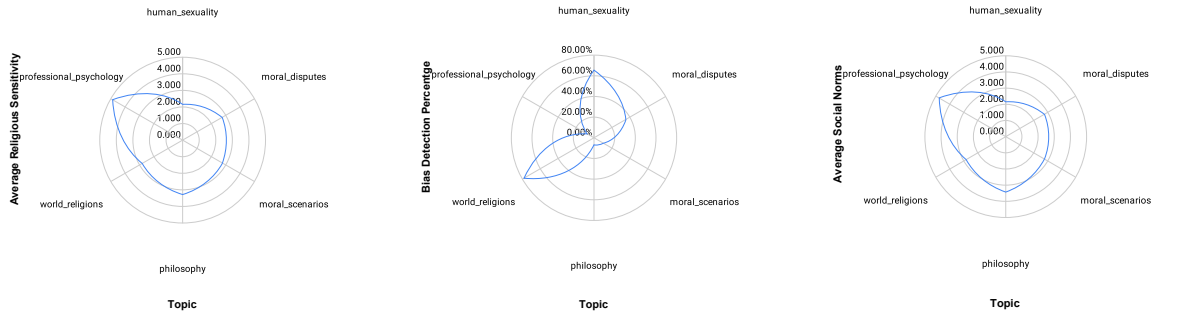
6 Discussion

The evaluation of the Arabic MMLU Benchmark highlights foundational challenges across three key areas: linguistic, cultural, and methodological/structural issues. These challenges underscore the limitations of directly translating Western-centric benchmarks for Arabic-speaking audiences, emphasizing the urgent need for a more culturally aligned and linguistically coherent approach to developing NLP resources for Arabic LLMs. Figure 5 summarizes the primary issues identified, serving as a basis for the discussions and recommendations presented in this study.

Linguistic Issues were prevalent throughout the corpus, impacting clarity and coherence. Transla-



(a) Average Fluency (b) Average Adequacy (c) Average Cultural Appropriateness



(d) Average Religious Sensitivity (e) Bias Detection Percentage (f) Average Social Norms

Figure 3: Radar Charts of Human Evaluation Metrics for Culturally Sensitive Topics in the Arabic MMLU Benchmark

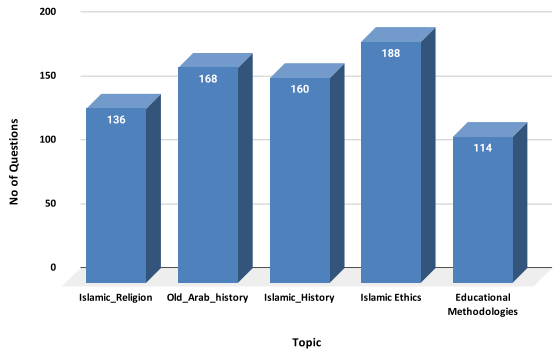


Figure 4: Distribution of Questions in Newly Added Culturally Relevant Topics

tion inconsistencies, such as the variable treatment of key terms and inconsistent handling of certain letters, detract from readability and comprehension. For example, some terms remain untranslated or inconsistently Arabized, even when well-established Arabic equivalents exist. This inconsistency disrupts the flow of the text, making it harder for readers to engage with the material. Additionally, grammatical errors and stylistic misalignments—such as overly literal translations—fail to adapt English sentence structures to Arabic, resulting in awkward or unnatural phrasing. These issues not only impact grammatical accuracy but also diminish the text's fluidity and clarity, making it feel less accessible and authentic to Arabic-speaking users.

Cultural Issues are evident where Western concepts, values, and figures are presented without adaptation, assuming universality and disregarding their relevance to Arabic-speaking communities. The corpus includes frequent references to Western laws, systems, and historical figures, while notable Arab figures and culturally significant examples are notably absent. This lack of cultural resonance weakens the benchmark's relevance for Arabic users, as it fails to reflect the linguistic and cultural heritage central to the Arabic-speaking world. Moreover, the reliance on Western terms, examples, and expressions, without providing classical or colloquial Arabic alternatives, distances the corpus from Arabic cultural and linguistic authenticity. Additionally, the inclusion of references to foreign ethnic groups and lineages, as well as cul-

Main issues	Sub problems	Details	examples
Linguistics	translation	Some terms are left untranslated.	وفقًا لنموذج القيادة الطارئة الذي وضعه فيدلر، فإن القادة بنقاط LPC عالية Professional psychology Q538 Did not translate the meaning of the abbreviation
		Lack of consistency in translation (e.g., use of letters in responses).	In moral_disputes Q303 used Arabic letters in the answers and in Q279 it used the English letters.
		Arabization of certain terms despite having Arabic equivalents.	"أخطر المضاعفات الفسيولوجية للشره المرضي العصبي هو" ← "علم وظائف الأعضاء" أو الجسمورية (وهو تحت) Professional psychology Q151
	Grammatical issues.	-	وفقًا لكارل روجرز، يكون المعالج "متوافقًا" عندما: "أصيل وغير دفاعي" ← "أصيلا وغير دفاعي" لأنها منصوبة بـ"يكون" مضمرة من السؤال. Professional psychology Q300
	Stylistic issues	Literal translation without adapting the sentence structure.	بعد ثلاث جلسات معهم أحد عملاء العلاج، يدرك الدكتور ليونارد ليكوفسكي أنه يشعر بالعدالية إلى حد ما تجاه العميلة لأنها تذكره بزوجته التي يطلقها حاليًا. أفضل مسار عمل للدكتور ليكوفسكي هو (D* استشارة أخصائي نفسي آخر لتحديد ما إذا كنت ستستمر في رؤية العميلة في العلاج أم لا" ← "إذا كان سيستمر" Professional psychology Q93
		Linguistically weak phrasing, even if grammatically correct.	"لتقليل العداء بين مجموعات طلاب المدارس الإعدادية، يُنصح بما يلي" ← "لتقليل العدالية ..." Professional psychology Q91
Cultural Issues	Relevance and Influence	Reliance on Western laws and regulations.	فيما يتعلق بالمبادئ التوجيهية للجنة لكافة فرص العمل، فإن معدل الاختيار أو الترقية المختلف بشكل كبير والذي يؤدي إلى الإضرار بأشخاص من جنس أو عرق أو إثنية معينة هو دليل على Professional psychology Q162 This is an american institute
		Questions lack relevance to Arabic culture overall.	الدكتور جولز، وهو أخصائي علم نفس سريري، كان يرى ليزا إل. في العلاج لمدة شهر واحد. يعتبر الدكتور جولز أن ليزا امرأة جذابة للغاية ويجد لديه تخيلات جنسية عنها. بصفته أخصائي علم نفس أخلاقي، يجب على الدكتور جولز. Professional psychology Q24
		Frequent mention of Western figures, with no reference to Arab figures.	All of the moral_disputes questions are from Western figures
		Inclusion of foreign ethnicities unrelated to Arab regions.	-
	Philosophical and Ethical Foundations	Western concepts are presented as universally accepted.	"لقد ثبت أن الروحانية تعمل كعامل وقائي ضد المرض، ومن الضروري بذل الجهد نحو فهم الروحانية ودمجها في الممارسة السريرية" Professional psychology Q143
		Differences in philosophical and ethical foundations are overlooked.	القيمة الأساسية التي يقوم عليها وجود المبادئ الأخلاقية هي: النهوض وحماية رفاهية عملاء الأخصائيين النفسيين Professional psychology Q107
	Language and Expressions	Use of Western terms, examples, and expressions without classical or colloquial Arabic alternatives.	"أنا أكره الجميع وكل شيء!" "لا، إنه غير ممكن!" "أنا أستسلم؛ أنا عاجزا" "إذا ساعدتني يا الله، سأصلح حياتي!" Professional psychology Q161 التنبه إلى اللفظ الصحيح من اللهجة العربية (ولو كان ذلك نقصًا في المكتوب ولكنّهُ أكثر مولوقية) وقد تضاف له جملٌ من الثقافة مثل : "الحمد لله على كل حال"، "إنا لله وإنا إليه راجعون"، "لا حول ولا قوة إلا بالله"، "كله خير ان شاء الله".
		Terminology is altered to terms that conflict with Arabic cultural norms.	من عام 1988 إلى عام 1990 بين المغايرين جنسيًا في الولايات المتحدة. human sexuality Q108 هذا مصطلح الطبيعيين وليس المغايرين
Methodological and Structural Issues	Structural Problems	Misplacement of questions in incorrect sections, leading to a lack of logical organization.	"أي مما يلي هو أفضل مثال على الخطأ المستمر" Professional psychology Q344 Question in statistics in psychology questions
	Research and Source References	Knowledge sources should be properly attributed, such as referencing original sayings of ancient philosophers from Arabic first translations	"يعرّف أرسطو الفضيلة بأنها" الميل إلى تجنب التطرف في المشاعر والأفعال. بينما في كتاب الأخلاق بترجمة ابن حنين فهناك صياغات أخرى أفصح وأبلغ moral disputes Q337 Instead of translating the english translation of Arabic translation of the original language
		Insufficient reliance on Arabic references and books in writing and translating texts.	-
		Absence of Arabic statistical research and studies in the questions, reducing cultural and contextual relevance.	No mention of any Arabic region research or papers الاضطراب النفسي الوظيفي الأكثر شيوعًا في وقت لاحق من الحياة هو "الاكتئاب" Professional psychology Q268

Figure 5: Foundational Issues in the Arabic MMLU Benchmark Dataset

Model Name	Parameters	Average Score	Model Type
Qwen/Qwen2.5-72B-Instruct	72B	**73.45**	Instruction-tuned
CohereForAI/aya-expanse-32b	32B	63.87	Pretrained
Qwen/Qwen2.5-32B-Instruct	32B	60.27	Instruction-tuned
CohereForAI/c4ai-command-r-08-2024	32.2B	59.85	Pretrained
google/gemma-2-9b-it	9B	57.73	Pretrained
Qwen/Qwen2.5-7B-Instruct	7B	55.57	Instruction-tuned

Table 2: ILMAAM Leaderboard: Top Performing Arabic LLMs

turally inappropriate terminological choices, can create dissonance with Arabic norms and values, further reducing the corpus's applicability and cultural accuracy.

Methodological and Structural Issues were also observed, indicating a lack of organization and clear source attribution within the corpus. Misplaced questions and a lack of references to Arabic sources and statistical research limit the benchmark's relevance and accuracy in Arabic contexts. Without properly cited sources or organized content, the text may feel less credible, as it does not ground its questions or assumptions in resources or research relevant to the Arabic-speaking world. This lack of structural coherence undermines the benchmark's utility, as it risks presenting information or perspectives that may not be applicable or accurate in an Arabic cultural framework.

7 Conclusion

This study provides a comprehensive evaluation of the Arabic MMLU Benchmark, highlighting critical issues in linguistic coherence and cultural alignment that hinder its effectiveness for Arabic. Results reveal cultural misalignments stemming from an over-reliance on Western concepts and a lack of clear Arabic source references, all of which reduce the benchmark's cultural relevance and usability. Furthermore, the large volume of questions across varied topics poses a challenge for thorough cultural review, as addressing this comprehensively requires a larger team and extended time commitment. These insights underscore the need for a refined benchmark with culturally aligned topics. Future work should focus on evaluating Arabic LLMs on culturally tailored benchmarks to assess their performance when engaging with content that resonates with Arabic social, historical, and ethical perspectives.

Acknowledgments

The authors thank Prince Sultan University for their support.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *arXiv preprint arXiv:2110.06733*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, et al. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

Yuu Jinnai. 2024. Does cross-cultural alignment change the commonsense morality of language models? *arXiv preprint arXiv:2406.16316*.

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*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Zhaozhi Qian, Faroq Altam, Muhammad Saleh Saeed Alqurishi, and Riad Souissi. 2024. Cameleval: Advancing culturally aligned arabic language models and benchmarks. *arXiv preprint arXiv:2409.12623*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

UNESCO. World arabic language day. https://www.unesco.org/en/world-arabic-language-day. Accessed: October 29, 2024.

Ximen Yuan, Jinshan Hu, and Qian Zhang. A comparative analysis of cultural alignment in large language models in bilingual contexts.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Jianqing Zhu, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Abdulmohsen Alharthi, Bang An, Juncai He, Xiangbo Wu, et al. Second language (arabic) acquisition of llms via progressive vocabulary expansion.

Appendices

A Annotation Guidelines

Guidelines for Culture Alignment and Translation Evaluation for Arabic MMLU Benchmark

A.1 Evaluation Criteria

Evaluators assess translations based on two primary categories: Translation Metrics and Culture Alignment Metrics. The Culture Alignment Metrics apply only to topics requiring additional cultural sensitivity (CA-marked topics).

A.1.1 Translation Metrics

These metrics evaluate the linguistic quality of the translation to ensure accuracy and naturalness in the target language.

- **Fluency:** Measures grammatical accuracy and ease of reading. Ratings range from:
  - **1** – Incomprehensible
  - **2** – Poor fluency with many grammatical errors and unnatural phrasing
  - **3** – Understandable but contains some awkward language
  - **4** – Good fluency with minor errors and natural phrasing
  - **5** – Native-level fluency, flawless grammar, and exceptionally natural language
- **Adequacy:** Assesses how accurately the translation conveys the meaning, intent, and nuances of the source text. Ratings are:
  - **1** – None of the meaning is conveyed; translation is irrelevant
  - **2** – Little meaning is conveyed; major information missing or incorrect
  - **3** – Some meaning is conveyed; partial information is accurately translated
  - **4** – Most meaning is conveyed; minor details may be missing or slightly inaccurate
  - **5** – Complete and precise meaning conveyed without loss or distortion

A.1.2 Culture Alignment Metrics (CA Topics Only)

These metrics evaluate the cultural appropriateness and sensitivity of the translation to ensure alignment with the target audience's cultural norms and values.

- **Cultural Appropriateness:** Evaluates respect for cultural norms, values, and sensitivities. Ratings are:
  - **1** – Highly inappropriate or offensive

  - **2** – Contains inappropriate elements
  - **3** – Neutral but lacks cultural adaptation
  - **4** – Appropriate with minor issues
  - **5** – Highly appropriate and culturally adapted

- **Bias Detection:** Identifies any biases or stereotypes in the translation. Evaluators mark:
  - **Yes** – Bias is present
  - **No** – No bias detected

  If bias is detected, specify the type:
  - Gender Bias, Cultural Bias, Religious Bias, Socioeconomic Bias, Age-related Bias, or Other (specify)

- **Religious Sensitivity:** Assesses alignment with religious beliefs and practices. Ratings are:
  - **1** – Highly Offensive or Blasphemous: Disrespectful or blasphemous towards religious beliefs
  - **2** – Inappropriate or Disrespectful: Uses sacred symbols or references inaccurately
  - **3** – Neutral but lacks sensitivity: Does not demonstrate awareness of religious nuances
  - **4** – Appropriate with minor issues: Mostly respectful with minor inaccuracies
  - **5** – Highly Respectful and Aligned: Fully respects religious beliefs, with accurate references

- **Social Norms:** Determines acceptability within societal context, respecting cultural traditions and values. Ratings are:
  - **1** – Highly inappropriate or taboo: Violates societal norms or includes taboo content
  - **2** – Inappropriate or insensitive: Contains elements that may cause discomfort
  - **3** – Acceptable but lacks cultural adaptation: Generally acceptable but culturally neutral
  - **4** – Appropriate with minor misalignments: Mostly aligns with social norms
  - **5** – Highly appropriate and culturally adapted: Fully aligns with cultural values and traditions

A.2 Evaluation Procedure

- **Preparation:** Review the source and translated text to understand the context. For CA-marked topics, ensure familiarity with relevant cultural and religious norms.
- **Rating Process:** First evaluate Fluency and Adequacy. For CA topics, proceed with Culture Alignment metrics.
- **Documentation:** Record scores for each metric, specifying any detected bias type and providing constructive feedback.

A.3 Best Practices

- **Consistency:** Apply criteria uniformly across all translations.
- **Objectivity:** Base evaluations strictly on defined criteria, minimizing personal bias.
- **Cultural Sensitivity:** Approach each translation with respect for cultural differences.

A.4 Quality Assurance

- **Calibration Sessions:** Conduct training sessions to align understanding of evaluation criteria.
- **Inter-Rater Reliability:** Compare evaluations to ensure consistency among evaluators.

A.5 Ethical Considerations

- **Respect and Sensitivity:** Handle all content respectfully, particularly sensitive cultural or religious topics.
- **Impartiality:** Evaluate objectively, without cultural biases.

Adhering to these guidelines ensures that translations are not only accurate and fluent but also culturally resonant and sensitive, supporting the development of high-quality, reliable, and respectful translations.

B Statistics

This section provides additional statistics and evaluation results relevant to the Arabic MMLU Benchmark.

B.1 Topic Distribution

Figure 6 shows the distribution of all topics included in the Arabic MMLU Benchmark, detailing the number of questions per topic. This includes reviewed, excluded, and newly added culturally relevant topics, providing an overview of the breadth of content evaluated in this study.

As shown in Figure 6, there is substantial variability in question coverage across different subjects, totaling 14,808 questions. Among these, 2,466 questions are in topics requiring Cultural Alignment (CA), such as Human Sexuality (131 questions), Moral Disputes (346 questions), World Religions (171 questions), and Professional Psychology (612 questions). This focus on culturally sensitive topics aims to ensure that Arabic language models can handle nuanced cultural content effectively. In addition, 706 questions are allocated to topics marked as excluded, including High School European History (165 questions), High School U.S. History (204 questions), and U.S. Foreign Policy (100 questions), as these topics lack cultural relevance for Arabic-speaking communities. To address gaps in cultural representation, 766 questions were added in newly introduced topics that are culturally significant for Arabic speakers, such as Islamic Religion (136 questions), Old Arab History (168 questions), Islamic Ethics (188 questions), and Educational Methodologies (114 questions). These numbers underscore the benchmark's attempt to balance general, culturally aligned, and excluded topics, though certain areas like professional law (1,534 questions) and moral scenarios (895 questions) have a disproportionately high representation. This uneven distribution highlights areas for potential improvement, emphasizing the need for a more balanced approach to ensure comprehensive cultural and linguistic evaluation in Arabic NLP models.

B.2 Automated Metrics for All Topics

Table 3 presents the results of automated evaluation metrics across all topics in the Arabic MMLU Benchmark, including BLEU, ROUGE, METEOR, chrF, BERTScore, and COMET scores for each topic.

The automated metrics for the Arabic MMLU Benchmark reveal significant variations in translation quality across topics. Technical and structured subjects like abstract algebra and international law achieve relatively high BLEU scores (0.442 and 0.353, respectively), indicating effective alignment with source material. Mathematical and scientific topics such as elementary mathematics and high school mathematics also perform well, benefiting from consistent terminology that translates effectively. In contrast, culturally sensitive topics like high school European history and high school US history display extremely low BLEU scores, highlighting the difficulty of adapting Western-centric content to an Arabic cultural context, which supports their designation as excluded topics.

Semantic metrics such as BERTScore and COMET provide a more consistent evaluation across topics, with scores generally above 0.85 for areas like sociology and world religions, indicating successful semantic preservation even when literal translations vary. However, fields requiring precise language, such as professional medicine (with a chrF score of 19.561), show lower performance, reflecting challenges in maintaining accuracy and clarity in complex professional contexts. These results emphasize the need for targeted adaptation in culturally sensitive areas and specialized refinement in technically demanding domains to improve translation quality and cultural relevance.

C Examples of New Added Topics

We provide some examples of the newly added topics, such as Islamic Religion, Old Arab History, Islamic History, Islamic Ethics, and Educational Methodologies, which represent the new refined Arabic MMLU benchmark. Figure 7 shows some of the examples of different topics. The examples in Figure 7 illustrate the depth and relevance of the newly added topics, focusing on culturally and contextually significant themes for Arabic-speaking audiences. Topics such as Islamic religion and Islamic ethics address core principles like honesty and the pillars of faith, which are fundamental to understanding the cultural and religious values prevalent in Arabic-speaking societies. Meanwhile, Old Arab History and Islamic History provide historical insights that are crucial for a well-rounded knowledge base within the Arabic context, such as significant events and geographical knowledge like the conquest of Constantinople and notable locations in Yemen. Educational methodologies emphasize Islamic perspectives on social and academic development, offering culturally aligned educational insights. Together, these examples demonstrate the enhanced cultural specificity and educa-

tional depth of the refined Arabic MMLU Benchmark, ensuring more accurate and culturally relevant assessments for Arabic NLP models.

D Comprehensive View of ILMAAM Results

The ILMAAM leaderboard provides a cohesive overview of how various Arabic-focused large language models (LLMs) perform across diverse academic and professional topics, revealing both strengths and limitations, as shown in Table 4. The top-performing models, such as *Qwen/Qwen2.5-72B-Instruct* and *CohereForAI/aya-expanse-32b*, excel in specific areas like college biology and high school US history, showcasing the benefits of larger parameter sizes and instruction tuning for handling nuanced questions. However, even high-performing models demonstrate variability, indicating the complexity of aligning language models with Arabic culturally specific content.

Notably, pretrained models tend to lag behind instruction-tuned counterparts, suggesting that additional fine-tuning is essential for capturing the subtleties of Arabic language and cultural context. The best-performing topics often center around Western historical and legal concepts, indicating a need for enhanced cultural and contextual training within Arabic-speaking contexts. This analysis underscores the importance of dedicated Arabic NLP resources and culturally aligned benchmarks, like ILMAAM, to foster Arabic LLMs that are both accurate and culturally relevant, promoting their utility and acceptance in Arabic-speaking communities.

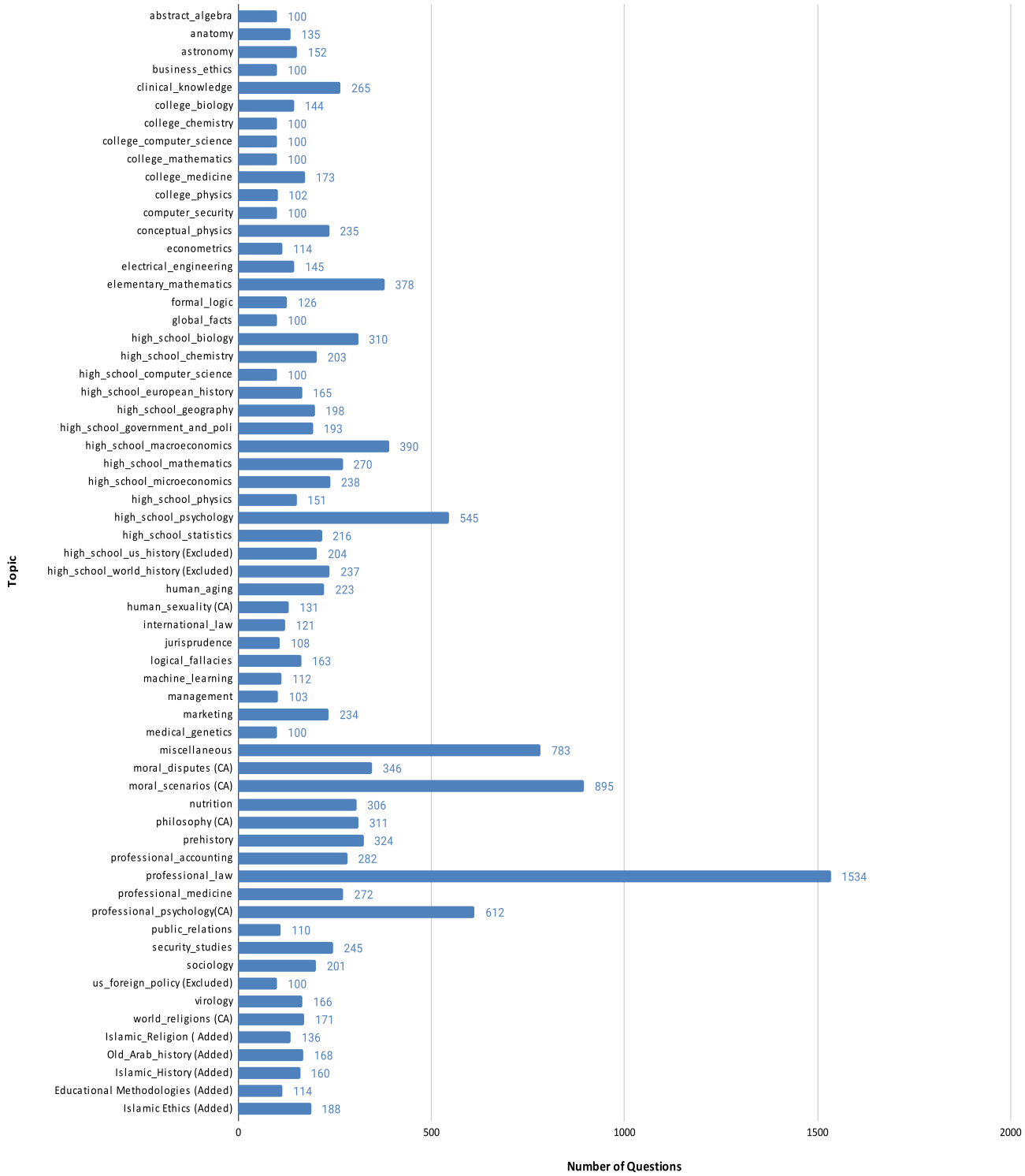


Figure 6: Distribution of Topics with Number of Questions in the Arabic MMLU Benchmark

Topic	BLEU	ROUGE	METEOR	chrF	BERTScore	COMET
abstract_algebra	0.442	0.908	0.597	53.876	0.888	0.793
anatomy	0.172	0.006	0.310	47.994	0.838	0.779
astronomy	0.244	0.116	0.455	53.058	0.858	0.838
business_ethics	0.381	0.060	0.436	56.939	0.864	0.848
clinical_knowledge	0.238	0.084	0.385	53.841	0.853	0.827
college_biology	0.136	0.078	0.297	43.896	0.821	0.771
college_chemistry	0.206	0.472	0.392	46.386	0.845	0.791
college_computer_science	0.174	0.473	0.372	38.579	0.836	0.774
college_mathematics	0.313	0.814	0.480	46.355	0.855	0.807
college_medicine	0.050	0.216	0.351	24.422	0.841	0.815
college_physics	0.172	0.594	0.357	43.853	0.842	0.796
computer_security	0.136	0.237	0.313	38.050	0.817	0.785
conceptual_physics	0.226	0.130	0.399	52.547	0.842	0.802
econometrics	0.221	0.332	0.413	46.531	0.839	0.785
electrical_engineering	0.186	0.252	0.367	49.892	0.833	0.789
elementary_mathematics	0.255	0.774	0.487	51.349	0.865	0.840
formal_logic	0.432	0.571	0.575	59.415	0.875	0.796
global_facts	0.239	0.684	0.449	56.924	0.857	0.868
high_school_biology	0.162	0.133	0.333	45.205	0.836	0.803
high_school_chemistry	0.253	0.363	0.409	52.655	0.853	0.825
high_school_computer_science	0.347	0.500	0.496	53.532	0.868	0.837
high_school_european_history	0.0000024	0.144	0.018	4.461	0.669	0.532
high_school_geography	0.237	0.027	0.432	58.121	0.853	0.868
high_school_government_and_poli	0.192	0.092	0.351	52.461	0.838	0.836
high_school_macroeconomics	0.248	0.135	0.410	58.315	0.859	0.848
high_school_mathematics	0.458	0.817	0.579	55.292	0.885	0.846
high_school_microeconomics	0.216	0.103	0.361	52.805	0.841	0.826
high_school_physics	0.164	0.456	0.336	40.815	0.835	0.791
high_school_psychology	0.175	0.065	0.359	46.895	0.841	0.823
high_school_statistics	0.200	0.537	0.385	47.783	0.847	0.818
high_school_us_history	0.0000021	0.180	0.023	4.530	0.665	0.505
high_school_world_history	0.0000138	0.240	0.029	5.612	0.678	0.544
human_aging	0.206	0.027	0.370	50.559	0.834	0.833
human_sexuality	0.222	0.035	0.376	46.695	0.841	0.816
international_law	0.353	0.060	0.530	63.608	0.889	0.896
jurisprudence	0.212	0.052	0.393	50.554	0.846	0.852
logical_fallacies	0.213	0.068	0.302	45.712	0.811	0.773
machine_learning	0.264	0.515	0.384	47.302	0.843	0.769
management	0.185	0.034	0.367	51.505	0.860	0.866
marketing	0.259	0.053	0.421	56.000	0.844	0.854
medical_genetics	0.194	0.165	0.293	46.570	0.814	0.769
miscellaneous	0.217	0.107	0.444	50.127	0.867	0.859
moral_disputes	0.250	0.008	0.440	55.401	0.868	0.837
moral_scenarios	0.356	0.937	0.578	61.078	0.853	0.769
nutrition	0.238	0.077	0.441	56.045	0.866	0.861
philosophy	0.329	0.000	0.497	56.236	0.884	0.861
prehistory	0.254	0.053	0.405	52.457	0.851	0.826
professional_accounting	0.192	0.310	0.395	47.588	0.844	0.820
professional_law	0.192	0.306	0.395	47.588	0.833	0.796
professional_medicine	0.026	0.600	0.177	19.561	0.802	0.722
professional_psychology	0.234	0.089	0.400	49.992	0.849	0.823
public_relations	0.248	0.088	0.447	54.326	0.859	0.869
security_studies	0.230	0.016	0.433	56.244	0.869	0.889
sociology	0.218	0.132	0.418	51.580	0.840	0.850
us_foreign_policy	0.314	0.060	0.533	64.544	0.882	0.899
virology	0.239	0.073	0.442	52.197	0.858	0.862
world_religions	0.199	0.019	0.398	54.489	0.867	0.853

Table 3: Automated Metrics Results for All Topics

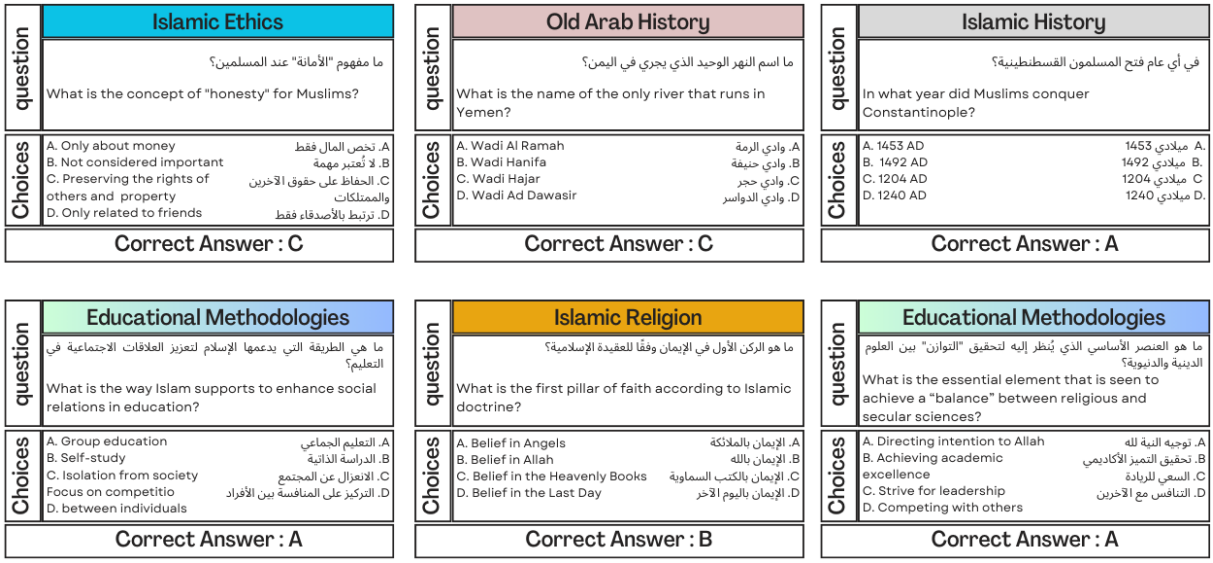


Figure 7: Examples from Islamic Ethics and Educational Methods Topics

Model Name	Parameters (in Billion)	Model Type	Average Score	Best Performing Topic	Best Topic Score
Qwen/Qwen2.5-72B-Instruct	72.7	instruction-tuned	73.455	college_biology	91
CohereForAI/aya-expanse-32b	32.3	pretrained	63.873	high_school_us_history	88
Qwen/Qwen2.5-32B-Instruct	32.764	instruction-tuned	60.272	international_law	79
CohereForAI/c4ai-command-r-08-2024	32.296	pretrained	59.852	high_school_us_history	86
google/gemma-2-9b-it	2.61	pretrained	57.732	high_school_world_history	79
Qwen/Qwen2.5-7B-Instruct	7.616	instruction-tuned	55.571	high_school_world_history	76
FreedomIntelligence/AceGPT-v2-32B	32.5	pretrained	54.851	high_school_world_history	79
silma-ai/SILMA-9B-Instruct-v1.0	9.24	fine-tuned	53.331	us_foreign_policy	74
CohereForAI/aya-expanse-8b	8.03	pretrained	51.790	us_foreign_policy	76
Qwen/Qwen2.5-3B-Instruct	3.086	instruction-tuned	48.450	high_school_world_history	71
FreedomIntelligence/AceGPT-v1.5-13B-Chat	13.147	pretrained	47.810	marketing	76
CohereForAI/aya-23-8B	8.028	pretrained	43.069	security_studies	69
google/gemma-2-9b-it	9.24	pretrained	40.288	sociology	66
Qwen/Qwen2.5-1.5B	1.544	pretrained	39.468	us_foreign_policy	63
FreedomIntelligence/AceGPT-v2-8B-Chat	8.03	instruction-tuned	39.068	international_law	64
Qwen/Qwen2.5-0.5B-Instruct	0.494	instruction-tuned	33.287	international_law	56
inceptionai/jais-family-13b-chat	13.5	instruction-tuned	32.587	high_school_european_history	51
Qwen/Qwen2.5-0.5B	0.494	pretrained	31.906	us_foreign_policy	52
meta-llama/Llama-3.2-3B-Instruct	3.21	instruction-tuned	31.806	sociology	53
meta-llama/Llama-3.2-3B	3.21	pretrained	28.906	high_school_world_history	43
inceptionai/jais-family-2p7b-chat	2.95	instruction-tuned	28.806	high_school_statistics	44
meta-llama/Llama-3.2-1B	1.24	pretrained	26.785	high_school_computer_science	36
inceptionai/jais-family-30b-8k	30	pretrained	26.545	business_ethics	39
meta-llama/Llama-3.2-1B-Instruct	1.24	instruction-tuned	25.705	us_foreign_policy	41
inceptionai/jais-family-2p7b	2.95	pretrained	22.985	business_ethics	32
inceptionai/jais-family-1p3b-chat	1.56	instruction-tuned	22.765	machine_learning	33
inceptionai/jais-family-13b	13.5	pretrained	22.565	marketing	32
arcee-ai/Meraj-Mini	7.62	pretrained	22.424	high_school_world_history	36
inceptionai/jais-family-590m-chat	771	instruction-tuned	22.404	business_ethics	30
inceptionai/jais-family-590m	771	pretrained	22.364	machine_learning	33
inceptionai/jais-family-1p3b	1.56	pretrained	22.204	professional_accounting	31

Table 4: ILMAAM Leaderboard: Performance Overview of Arabic LLMs