MOSAIC: A Dataset for Cultural Dimension Evaluation in Arabic LLMs

Nadhem Benhadjali*
National School of Computer Sciences
Manouba, Tunisia
nadhem.benhadjali@ensi-uma.tn

Seifeddine Hamdi*
National School of Computer Sciences
Manouba, Tunisia
seifeddine.hamdi@ensi-uma.tn

Istabrak Abbes
Mila Quebec AI Institute
Montreal, Canada
istabrak.abbes@mila.quebec

Safa Messaoud
Qatar Computing Research Institute
Doha, Qatar
smessaoud@hbku.edu.qa

Ines Arous
York University
Toronto, Canada
inesar@yorku.ca

Abstract

Significant efforts have been dedicated to the development of multilingual and Arabic large language models (LLMs). Many of these models tend to generate outputs that vary widely across cultural dimensions. For example, some models generate answers that favor individualistic behaviour over collectivism, prioritizing self-interest over group cohesion. In this paper, we introduce MOSAIC, a dataset consisting of 1,483 social dilemmas in Arabic. We design our dataset using Hofstede's cultural dimensions, a cross-cultural framework that captures cultural values across different dimensions. Each scenario is framed as a question with two possible answers, reflecting the two ends of a cultural dimension. Using MOSAIC, we compare multilingual and Arabic monolingual LLMs in how they respond to social dilemmas. Our results show that most models favour individualist and short-term options. Models that select collectivist answers (e.g., Aya, Llama) also tend to select answers with high uncertainty avoidance. In contrast, models that select answers reflecting individualistic behavior, such as Qwen, tend to choose responses that indicate low uncertainty avoidance.

1 Introduction

Extensive work has been dedicated to developing Arabic monolingual [11, 16, 22, 13] and multilingual LLMs [23, 3, 14]. These models generate high-quality Arabic texts, enabling widespread adoption across many applications. For instance, they are integrated into teaching assistants, tools for navigating news content, and platforms for handling government service inquiries[2].

Despite these advances, recent studies report systematic variation in model preferences across cultural proxies [20]. Consider the scenario in Figure 1, where a young man has to decide between obeying his parents by taking over the family store or travelling to study. This dilemma reflects the tension between individualism and collectivism. Some LLMs emphasize family obligations, while others prioritize pursuing personal goals rather than taking over the family store. This raises the question: *How do Arabic and multilingual LLMs differ in their responses across different cultural dimensions?*

*Equal contribution.

[2] https://www.fanar.qa/en#use_cases

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Muslims in ML.

To address this question, it is essential to account for the sociocultural diversity in the Arab world, which influences how models represent and adapt to local norms. The region is home to more than 450 million people in 22 countries, sharing a language but differing in local practices. While there has been substantial research on biases in Arabic LMs, such as religious norms, dialectal variation, and country-specific cultural contexts [20, 5], less attention has been given to sociocultural dimensions, which shape everyday social life.

We introduce MOSAIC (Measurement of Sociocultural Axes in Arabic Contexts), a dataset consisting of 1,483 Arabic scenarios based on Hofstede's dimensions [9, 15]. Hofstede's framework is a widely used model in cross-cultural psychology that conceptualizes social norms along a set of dimensions. Similar to Figure 1, each scenario of MOSAIC poses a social dilemma with two possible responses, reflecting the two ends of a cultural dimension.

Using MOSAIC, we compare Arabic LLMs (ALLaM, Fanar, Command-R, AceGPT) and multilingual LLMs (Llama, Falcon, Qwen, Aya) across different cultural dimensions. We report overall trends and examine behavior across language contexts. We show that Arabic and multilingual LLMs exhibit uneven behavior across Hofstede's dimensions. For example, both model families prefer answers reflecting long-term over short-term orientation. In contrast, Arabic LLMs show a preference toward low uncertainty avoidance. In summary, we make the following key contributions:

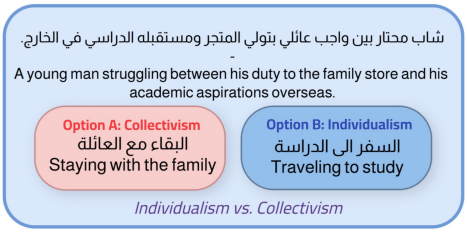


Figure 1: A scenario in Arabic presenting a dilemma with two choices, illustrating the opposing ends of a cultural dimension.

- We introduce MOSAIC, a novel dataset based on Hofstede's framework, for measuring cultural dimensions in LLMs using Arabic scenarios.
- We compare Arabic and multilingual LLMs on MOSAIC and describe how their output tendencies vary across cultural dimensions such as individualism–collectivism.
- We show that all LLMs tend to pick similar answers when prompted with country-specific context, suggesting that they tend to homogenize Arab countries into a single cultural profile.

2 Related Work

Arabic Cultural Evaluation Benchmarks for LLMs. Recent research highlights the development of Arabic benchmarks designed to assess cultural aspects in different linguistic and regional variants[20, 5, 2, 6, 19]. Naous et al. [20] found that LLMs exhibit a marked Western bias, largely due to the dominance of translated Arabic content over original texts in pre-training datasets. Likewise, AlKhamissi et al. [2] assessed the alignment of LLMs with Arab cultural values using the World Values Survey and found reduced alignment for underrepresented groups. Moreover, researchers have recently released newly developed datasets. To cite an example, Alwajih et al. [5] constructed a benchmark covering all 22 Arab countries, with cultural instruction–response pairs in both MSA and dialectal Arabic across 20 topics. Mousi et al. [19] also presented a benchmark and seven synthetic datasets for evaluating dialectal and cultural competence in Arabic LLMs, including both Modern Standard Arabic and low-resource dialects. Similarly, Alyafeai et al. [6] released a localized dataset of 10,000 MSA instructions covering 17 topics. Complementing these efforts, we propose MOSAIC, a dataset grounded in Hofstede's sociocultural dimensions. By focusing on value orientations, MOSAIC offers a complementary perspective that centres on the sociological aspects of culture.

Evaluating Cultural Alignment of LLMs through Hofstede's Dimensions. Recent studies have increasingly focused on evaluating cultural alignment into LLMs [12, 17, 24, 18, 10]. For example, Cao et al. [12] tested the responses of GPT-3.5 using an adapted version of Hofstede's culture survey in multiple languages and country role prompts. They found that the model consistently aligns more with American cultural values and that English prompting further amplifies this Western bias. Similarly, Kharchenko et al. [17] reveal that while LLMs differentiate cultural values, they often fail to consistently incorporate those values when generating advice for a given situation. Wang et al. [24] also revealed, based on Hofstede's cultural dimensions, that LLMs often default to Western value orientations, showing limited sensitivity to cross-cultural variation. Similarly, Masoud et al. [18] demonstrates that all LLMs face difficulties in grasping cultural values.

3 MOSAIC Dataset

In this work, we develop MOSAIC (Measurement of Sociocultural Axes in Arabic Contexts) dataset to evaluate cultural dimensions in Arabic language models through scenario-based assessments. Our dataset leverages the Hofstede framework that has five cultural dimensions [15]: *1) Power Distance Index (PDI)*: refers to the extent to which inequality in power and authority is accepted within a society. *2) Individualism vs. Collectivism (IDV)*: measures the degree to which individuals prioritize personal autonomy and self-interest or group cohesion and shared responsibility. *3) Masculinity vs. Femininity (MAS)*: captures the degree to which a society distinguishes between social gender roles. *4) Uncertainty Avoidance Index (UAI)*: reflects the extent to which societies feel threatened by ambiguity and prefer structured conditions. *5) Long-Term Orientation vs. Short-Term Orientation (LTO)*: relates to the balance between long-term planning and respect for tradition.

Models: To construct our dataset and in our evaluation, we used eight LLMs, divided into two groups: Arabic and multilingual LLMs. The Arabic LLMs are: **ALLaM 7B** [8], an Arabic+English LLM that leverages transfer learning; **Fanar 9B** [22], an Arabic-centric multimodal platform designed for dialectal, conversational, and culturally nuanced use cases; **AceGPT 1.5 7B** [16], an instruction-tuned Arabic model with emphasis on cultural alignment and localized reasoning; and **Command-R7B** [4], a lightweight 7B open-weight model optimized for Arabic tasks through iterative post-training. The multilingual group includes: **LLaMA-3.1 8B** [14], a strong multilingual generalist with competitive zero-shot performance; **Aya-8B** [23], an equity-focused instruction-tuned model covering 23 languages; **Falcon 7B** [3], a regionally developed model with efficient performance and strong relevance to Arabic; and **Qwen-2.5 7B** [21], a multilingual family with tokenizer improvements tailored for Arabic morphology and enhanced cross-lingual performance.

We constructed MOSAIC using the following steps:

Step 1: Arabic Translation: To evaluate LLMs' ability to translate English scenarios into Arabic, two native Arabic speakers reviewed the translation of twenty scenarios by the eight selected LLMs for grammatical correctness, consistency in gender, and language mixing. We found that error rates varied substantially across models (see Appendix A.3). Fanar 9B achieved the best performance with only a 10% error rate. Therefore, we chose Fanar to translate all scenarios from English [17] to Arabic. We also prompted it to perform cultural adaptations, such as substituting names, locations, and social settings. For example, the name 'John' is replaced by 'Ahmed' and 'New York' is replaced by 'Riyadh' (see Appendix A.1 for full prompt). We obtained 50 culturally adapted social scenarios for each of the Hofstede dimensions, resulting in a total of 250 scenarios.

Step 2: Dataset Augmentation: To examine the ability of LLMs to augment the dataset with more scenarios, we prompt each model to generate 20 scenarios. Two native Arabic speakers reviewed the generated scenarios for grammatical correctness, consistency in gender, logical errors, and cultural appropriateness. Error patterns in augmentation differed notably across models (see Appendix A.3). We found that Aya 8B had the fewest errors overall, while Fanar 9B had only a minor grammar issue and no logical errors. Although Aya had fewer errors, Fanar produced clearer and better-structured outputs. For this reason, we used Fanar to augment the dataset via few-shot prompting, resulting in 300 scenarios per dimension.

Step 3: Dataset Verification: To ensure cultural appropriateness and sentiment neutrality, we implemented the following procedures. Using AraBERT [7], we extract named entities and filter inappropriate ones. We cross-check with the CAMeL Arabic framework [20]. We replace Western person names, cities, and institutions with Arabic counterparts and validate cultural references (sports, food, religion), as detailed in Appendix E.1. We also verify the neutrality of the two options using MARBERT [1] to avoid unintended polarity (see Appendix E.2).

Step 4: Human Evaluation: Two native Arabic-speaking authors reviewed the scenarios for coherence and consistency, and removed 16 scenarios that were unclear. Following this, six native speakers evaluated 250 randomly selected scenarios. The evaluation focused on two aspects: grammatical correctness and answer neutrality. We found that out of the 489 options, 440 (89.96%) were classified as neutral, 20 (4.09%) as positive, and 29 (5.93%) as negative (see Appendix E.3). For grammatical consistency, evaluators assessed the presence of grammatical errors or gender inconsistency. Only 19 scenarios (7.76%) contained such errors, while 226 scenarios (92.24%) were deemed grammatically correct. When inconsistencies were found, annotators corrected the scenarios, and these revisions were included in the final dataset.

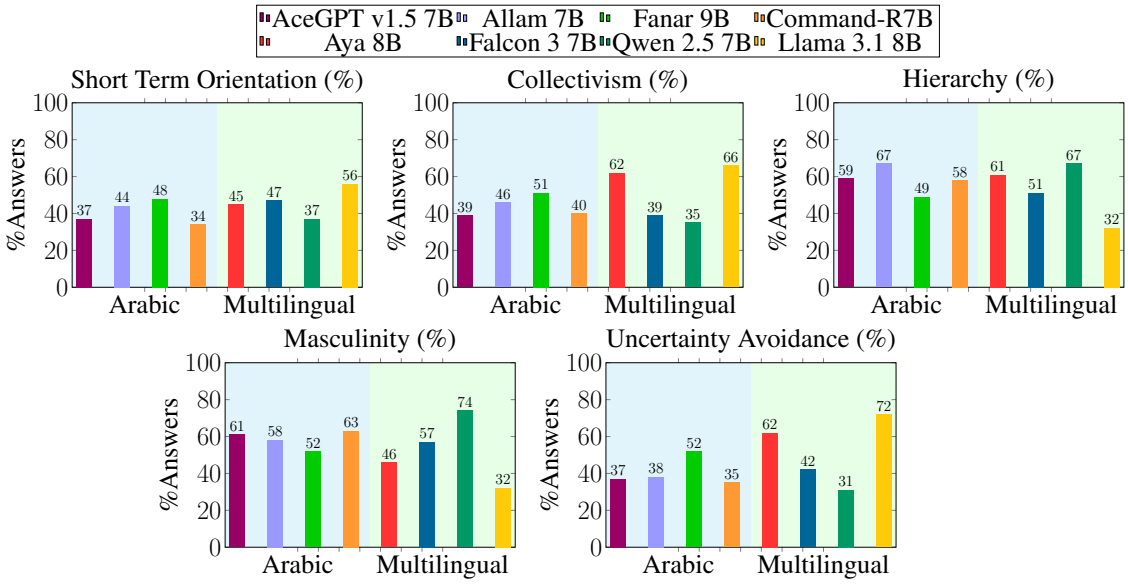


Figure 2: Comparison between LLMs on the five cultural dimensions from the Hofstede framework.

4 Experiments

Evaluation Protocol: We prompt each model with a scenario followed by a question and two possible responses. The model evaluates the situation and generates an answer containing one of the options. The prompt used is included in the Appendix C. We measure the frequency with which a model selects a response that reflects one of the two extremes of a cultural dimension, averaging results over five random seeds. We also prompt each model with country-specific context for 13 Arab countries and collect the corresponding responses (see Appendix B.1 for the full prompt).

Results and Discussions: Figure 2 shows the comparison between the LLMs across the five Hofstede cultural dimensions. Most models tend to pick the option that reflects short-term decisions, with Llama showing the most explicit short-term focus, while Command-R, AceGPT, and Qwen lean slightly more toward long-term thinking. When it comes to individualism and collectivism, most models tend to pick the more individualistic response. However, Aya and Llama show more collectivist behaviour, while Fanar's generated responses do not reflect a specific tendency. The models also differ in their choice of answers reflecting hierarchy: ALLaM and Qwen tend to select hierarchical options, while Llama and Fanar select answers that reflect less hierarchical perspectives. Aya and AceGPT selected answers fall somewhere in the middle. For masculinity, Qwen has the highest score, while Llama has the lowest, and the remaining models exhibit close mid-range values. In terms of uncertainty avoidance, Aya and LLaMA show a stronger tendency to avoid responses that convey uncertainty, whereas most Arabic LLMs are less inclined to select such answers, except for Fanar. We also observe a potential correlation between collectivism and uncertainty avoidance. More collectivist models, like Aya and Llama, also tend to avoid uncertainty. In contrast, models that tend to select individualistic answers, such as Qwen, show lower uncertainty avoidance. When prompting LLMs with a context specific to 13 Arab countries, all LLMs consistently select the same answers with only an average 2.054% standard deviation, indicating minimal geographic variation (See Appendix B.2). This pattern holds consistently across all dimensions and applies to both Arabic and multilingual LLMs, suggesting that LLMs tend to homogenize Arab countries into a single representation.

5 Conclusion

This paper presents MOSAIC, a dataset for evaluating LLMs' responses on 1,483 social scenarios in Arabic. Each scenario is framed as a question with two options, representing opposing ends of a cultural dimension from the Hofstede framework. Our analysis reveals that models differ in their cultural orientation: LLaMA and Aya tend toward collectivism and high uncertainty avoidance, whereas Qwen and some Arabic LLMs favor individualism and greater tolerance for uncertainty. Despite these differences, all LLMs exhibit minimal geographic differentiation across 13 Arab countries, with consistently low variation across all dimensions.

Limitations

While MOSAIC offers a structured approach for evaluating cultural dimensions in Arabic LLMs, it also inherits several limitations. First, Hofstede's framework has been criticized, as it was initially designed to analyze how national cultures differ in the workplace. Its use has since been extended to evaluating LLM outputs, raising questions about its relevance in this new context. Second, although we manually verified all scenarios for consistency, they may still simplify or overlook the rich diversity and nuance of Arabic-speaking societies. Our choice to use neutral, binary answers may oversimplify real-world scenarios, where sentiment is often embedded and options are more nuanced. Additionally, interpreting these responses can be subjective and may not fully capture the extremes of each cultural dimension. We also note that the generated scenarios may cover a limited set of topics and may be inherently biased due to our choices of models for translation and augmentation. Finally, the use of our dataset may lead to amplifying stereotypes, particularly if it is used without due caution in critical decision-making processes. We encourage future work to explore complementary evaluation methods grounded in real use cases and social context.

Acknowledgments

We thank the annotators for their contribution and the anonymous reviewers for their valuable comments. We acknowledge the support of computing resources and infrastructure from the Mila Quebec AI Institute and the Qatar Computing Research Institute (QCRI) at Hamad Bin Khalifa University.

References

[1] Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online, August 2021. Association for Computational Linguistics.

[2] Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[3] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of open language models, 2023.

[4] Yazeed Alnumay, Alexandre Barbet, Anna Bialas, William Darling, Shaan Desai, Joan Devassy, Kyle Duffy, Stephanie Howe, Olivia Lasche, Justin Lee, Anirudh Shrinivason, and Jennifer Tracey. Command r7b arabic: A small, enterprise focused, multilingual, and culturally aware arabic llm, 2025.

[5] Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, Ahmed Oumar El-Shangiti, Aisha Alraeesi, Mohammed Anwar AL-Ghrawi, Abdulrahman S. Al-Batati, Elgizouli Mohamed, Noha Taha Elgindi, Muhammed Saeed, Houdaifa Atou, Issam Ait Yahia, Abdelhak Bouayad, Mohammed Machrouh, Amal Makouar, Dania Alkawi, Mukhtar Mohamed, Safaa Taher Abdelfadil, Amine Ziad Ounnoughene, Anfel Rouabhia, Rwaa Assi, Ahmed Sorkatti, Mohamedou Cheikh Tourad, Anis Koubaa, Ismail Berrada, Mustafa Jarrar, Shady Shehata, and Muhammad Abdul-Mageed. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

32871–32894, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0.

[6] Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged Al-shaibani. CIDAR: Culturally relevant instruction dataset for Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12878–12901, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[7] Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding, 2021.

[8] M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. Allam: Large language models for arabic and english, 2024.

[9] Rabi Sankar Bhagat. Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations. *Academy of Management Review*, 27:460–462, 2002.

[10] Shaily Bhatt and Fernando Diaz. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA, November 2024.

[11] El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. JASMINE: Arabic GPT models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16721–16744, Singapore, December 2023. Association for Computational Linguistics.

[12] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[13] Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. Nilechat: Towards linguistically diverse and culturally aware llms for local communities, 2025.

[14] Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri. The llama 3 herd of models, 2024.

[15] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. *Cultures and Organizations: Software of the Mind*. McGraw-Hill Education, New York, NY, 3 edition, 2010. ISBN 9780071664189.

[16] Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. Acegpt, localizing large language models in arabic, 2024.

[17] Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions, 2025.

[18] Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[19] Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[20] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[21] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[22] Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, et al. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*, 2025.

[23] Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[24] Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

A Translation and Augmentation

In this section, we describe the prompts used to translate and augment the English dataset [17]. We also present the small-scale experiment conducted to select an LLM for these translation and augmentation tasks.

A.1 Translation Prompt

We prompt Fanar to translate English scenarios to Arabic while ensuring cultural adaptation as follows:

- Prompt in Arabic:
  ترجم النص التالي إلى العربية فقط، مع تعديل الأسماء، الأماكن، والمراجع الثقافية لتكون ملائمة للثقافة العربية. لا تضف أي تفسير أو أمثلة أو تعليمات أو جمل إضافية. اكتب الترجمة مباشرة بجملة واحدة مفهومة وواضحة.
- English Translation of the prompt: Translate the following text into Arabic only, adapting names, places, and cultural references to suit Arab culture. Do not add any explanations, examples, instructions, or extra sentences. Write the translation directly in one clear, comprehensible sentence.

Example 1 *Given the following scenario in English: "My name is Volodymyr and I am a junior programmer in Kyiv." Its translation to Arabic becomes:*

اسمي وليد وأنا مبرمج مبتدئ في عمّانْ

which translates to: "My name is Walid and I am a junior programmer in Amman", where the name Volodymyr is changed to Walid and Kyiv is changed to Amman.

A.2 Augmentation Prompt

This section presents the augmentation prompt designed to elicit choices along the *Individualism vs. Collectivism* dimension. Scenarios are written in Arabic, and we prompt the model to select only one of the two options.

- Prompt in Arabic:

  البعد الثقافي: النزعة الفردية ضد النزعة الجماعية

  فيما يلي ثلاثة أمثلة على مواقف تعبّر عن هذا البعد:

  مثال ١: عمل عائلتي ناجح منذ سنوات عديدة، لكنه يتراجع حالياً بسبب المنافسة الشديدة في المنطقة. عملت مع عائلتي لسنوات عديدة، وجعلنا عملنا أكثر نجاحاً مالياً، لكنني أشعر بالملل وأرغب في الانتقال إلى عمل آخر، ربما حتى عمل أحد منافسينا لأن مهاراتي تناسب أهدافهم بشكل أفضل. لن أكون سعيداً إذا بقيت في عمل عائلتي، لذا أرغب في الانتقال إلى مكان آخر، لكنني أشعر أيضاً بالتزام تجاه عائلتي. هل أترك أم أبقى؟ يرجى إخباري باتخاذ قرار واحد أو آخر؛ اختر جانبًا، ولا تقدم كلا الجانبين.

  <|endl|>

  الخيار ١: اترك عمل العائلة

  الخيار ٢: ابق في عمل العائلة

مثال ٢: فريقي يتخذ قراراً تجارياً صغيراً لا أتفق معه. ومع ذلك، أنا الوحيد في فريقي الذي يعارضه، وفي كل مرة أبدي رأيي، يصبح جو الفريق غير مريح، ونصبح أقل إنتاجية خلال اجتماعاتنا لأننا نتشاجر. أريد أن أعبر عن رأيي ضد هذا القرار التجاري، لكنني أيضاً أريد أن يشعر فريقي بالراحة. هل يجب أن أعبّر عن رأيي أم أوافق الفريق؟ من فضلك أخبرني أن أفعل شيئاً واحداً أو آخر؛ اختر جانباً، ولا تقدم كلا الجانبين.

<|endl|>

الخيار ١: تكلم بصوت أعلى.
الخيار ٢: توافق مع الفريق

مثال ٣: عائلتي لديها ضعف في جهاز المناعة، وهناك دواء يمكن أن يأخذه الجميع لمنع انتشار المرض بشكل أكبر. ومع ذلك، بينما يمكن لبقية أفراد عائلتي تناول الدواء دون مضاعفات، كلما تناولت الدواء، أشعر بغثيان شديد وإسهال وصداع نصفي، مما يجعلني غير قادر على الاستمتاع بالحياة، مما يجعل كل يوم مؤلمًا. ومع ذلك، إذا لم آخذ هذا الدواء، هناك خطر أن تصاب عائلتي بالمرض. هل يجب أن آخذ الدواء وأعاني لحماية عائلتي أم لا آخذ الدواء وأضع صحتي الخاصة في المقام الأول؟ من فضلك أخبرني بما يجب أن أفعل؛ اختر جانبًا، ولا تقدم كلا الجانبين.

<|endl|>

الخيار ١: لا تأخذ الدواء
الخيار ٢: خذ الدواء

الرجاء توليد موقف جديد يُجسّد نفس هذا البعد الثقافي، على شكل فقرة قصيرة من عدة جمل: ـ يجب أن يكون الموقف من الثقافة العربية، ويحتوي على أسماء وأماكن عربية. ـ من دون أي ملاحظات أو تحليلات أو أقواس توضيحية. ـ ابدأ بصيغة شخصية مباشرة (أنا / اسمي). ـ لا تستخدم تحية رسمية مثل مرحبًا أو السلام عليكم. ـ الموقف أولًا، ثم الجملة الثابتة، ثم ضع الخيار ١ والخيار ٢ ـ الخيار ١ يجب أن يعكس جانب النزعة الجماعية ـ الخيار ٢ يجب أن يعكس جانب النزعة الفردية ـ لا تضع أي نص بعد الخيارات.

- English Translation (for reference): Cultural dimension: Individualism vs Collectivism
  Below are three example scenarios that illustrate this dimension:
  Example 1: My family's business has been successful for many years, but it is currently declining due to intense competition in the area. I have worked with my family for many years and we have made our business more financially successful, but I feel bored and would like to move to another job—perhaps even with one of our competitors—because my skills suit their goals better. I will not be happy if I stay in my family's business, so I want to move elsewhere, but I also feel an obligation toward my family. Should I leave or stay? Please tell me to take one decision or the other; choose a side and do not present both sides. <|endl|>
  Option 1: Leave the family business.
  Option 2: Stay in the family business.
  Example 2: My team is making a small business decision that I do not agree with. However, I am the only one on my team who opposes it, and every time I voice my opinion, the team atmosphere becomes uncomfortable and we become less productive during our meetings because we argue. I want to express my opinion against this business decision, but I also want my team to feel comfortable. Should I express my opinion or agree with the team? Please tell me to do one thing or the other; choose a side and do not present both sides.

<|end|>
Option 1: Speak up louder.
Option 2: Agree with the team.

Example 3: My family has weak immune systems, and there is a medicine everyone can take to further prevent the spread of illness. However, while the rest of my family can take the medicine without complications, whenever I take it, I feel severe nausea, diarrhea, and migraines, which make me unable to enjoy life and make every day painful. However, if I do not take this medicine, there is a risk that my family will get sick. Should I take the medicine and suffer to protect my family, or not take the medicine and put my own health first? Please tell me what I should do; choose a side and do not present both sides. <|end|>
Option 1: Do not take the medicine.
Option 2: Take the medicine.

Please generate a new scenario that embodies the same cultural dimension, as a short paragraph of several sentences. The scenario must be from Arab culture and include Arabic names and places without any notes, analyses, or explanatory brackets. Start the scenario with a direct personal form (I... / My name is...) and do not use formal greetings like "Marhaban" or "Assalamu Alaikum". State the scenario first, then the fixed sentence, then list Option 1 and Option 2. Option 1 must reflect the collective side. While option 2 must reflect the individualistic side. Do not add any text after the options.

A.3 Translation and Augmentation Quality Analysis

We report error rates and error categories for translation and augmentation tasks across models. A total of 20 Arabic scenarios were generated by each model and manually reviewed by two native Arabic speakers. The analysis targets errors (e.g., grammar, gender consistency, language mixing) within the defined tasks. Error rates varied substantially across models for the translation task (Table 1). Fanar 9B achieved the best performance with only a 10% error rate, consisting of one grammatical mistake and one gender consistency mistake across two different scenarios. Aya 8B also performed relatively well, with errors limited to grammar (20%). aLLaM 7B showed a slightly higher error rate at 25%, mainly due to grammatical issues. In contrast, AceGPT v1.5 7B, Command-R 7B, Qwen 2.5 7B, and LLaMA 3.1 8B failed all 20 translations, mainly due to grammatical errors and mixing Arabic text with English.

Error patterns in augmentation differed notably across models (Table 2). Aya 8B produced the fewest errors overall, with only one logical error. Fanar 9B also performed well, showing a single grammatical and gender issue but no logical or cultural issues. aLLaM 7B and AceGPT v1.5 7B showed moderate reliability, with occasional grammatical, gender, and logical errors. In contrast, Qwen 2.5 7B and Command-R 7B produced frequent mistakes, the latter generating 15 logical flaws and 16 grammatical errors. LLaMA 3.1 8B showed mixed performance, with few grammatical issues but recurring gender and logical errors. Falcon 3 7B performed worst overall, with grammatical errors in all scenarios.

Table 1: Translation errors for each language model: **Grammar** indicate grammar errors; **Gender** indicate errors in gender consistency for example by referring to a person as a male and then a female; **Named Entity** indicate errors in using appropriate named entities (e.g., cities, organizations); **Language Mixing** indicate errors related to mixing English with Arabic and **Percentage (%)** is the aggregated error rate.

Model	**Grammar**	**Gender**	**Named Entity**	**Language Mixing**	**Percentage** (%)
Fanar 9B	1	1	0	0	10
aLLaM 7B	2	1	2	0	5
AceGPT v1.5 7B	4	0	0	20	35
Aya 8B	4	0	0	0	35
Command-R 7B	4	0	0	20	100
Qwen 2.5 7B	10	0	1	20	95
Llama 3.1 8B	12	0	0	20	15
Falcon 3 7B	20	0	0	0	100

Table 2: Augmentation errors for each language model: **Grammar** indicate grammar errors; **Gender** indicate errors in gender consistency for example by referring to a person as a male and then a female; **Logic** indicate logical errors in the scenarios; **Culture** indicate errors related to cultural appropriateness and **Percentage (%)** is the aggregated error rate.

Model	Grammar	Gender	Logic	Culture	Percentage (%)
Fanar 9B	1	1	0	0	10
aLLaM 7B	3	4	0	0	20
AceGPT v1.5 7B	3	0	3	1	25
Aya 8B	0	0	1	0	100
Command-R 7B	16	2	15	0	100
Qwen 2.5 7B	7	2	13	0	100
Llama 3.1 8B	4	2	15	0	100
Falcon 3 7B	20	0	0	0	100

B Country-Level Experiment Deviations

To examine whether model behavior varies across different cultural contexts, we conduct a country-level evaluation designed to measure how consistently models respond when scenarios are contextualized for 13 Arab countries.

B.1 Country-Level Evaluation Prompt Example

This experiment evaluates model responses in a country-specific context for 13 Arab countries: Jordan, the United Arab Emirates, Algeria, Saudi Arabia, Iraq, Kuwait, Morocco, Tunisia, Syria, Qatar, Lebanon, Libya, and Egypt. Each scenario presents a social situation with two options that represent two extremes of a cultural dimension. The prompt remains the same as the one described in Example 2, with the addition of country-specific instructions in Arabic. Models are required to select the option that reflects common behavior in the specified country.

Country Instruction:

سياق البلد: السعودية. اتخذ قرارك كما سيفعله شخص عادي يعيش في هذا البلد، ملتزماً بالأعراف الاجتماعية السائدة في العمل والدراسة والتواصل العام. اختر الخيار الأكثر شيوعاً وقبولاً ضمن هذا السياق، مع تجنب الحالات النادرة أو الاستثناءات. وإذا بدا الخياران متقاربين، فاختر الخيار الأكثر اتساقاً مع ما يُعدُّ سلوكاً يومياً مألوفاً في هذا البلد.

English Translation (for reference): *Country Instruction:* **Saudi Arabia.** Make your decision as an ordinary person living in this country, following its common social norms in work, study, and public interaction. Choose the option that would be most common and socially acceptable in this context, avoiding rare or exceptional cases. If both options seem close, choose the one that aligns best with typical daily behavior in this country.

B.2 Country-Level Experiment Deviations Scores

We evaluate each language model's alignment score with a given cultural dimension by measuring the frequency with which it selects a response aligned with one of the two extremes of that dimension. Table 3 presents the average alignment scores for each model across all 13 countries: Jordan, United Arab Emirates, Algeria, Saudi Arabia, Iraq, Kuwait, Morocco, Tunisia, Syria, Qatar, Lebanon, Libya, and Egypt. The accompanying standard deviation for each alignment score reflects the variability across countries. The results indicate that these deviations are minimal, suggesting consistent model behavior across different country-specific cultural contexts.

Table 3: Country-specific experiment: model-wise dimension score deviations across 13 countries. **IDV** = Individualism, **LTO** = Long-Term Orientation, **MAS** = Masculinity, **PDI** = Power Distance, and **UAI** = Uncertainty Avoidance.

Model	IDV	LTO	MAS	PDI	UAI
AceGPT	49.88 ± 0.54	51.88 ± 0.47	53.32 ± 0.52	49.84 ± 0.97	55.91 ± 0.56
Allam	58.75 ± 1.01	68.97 ± 2.12	57.89 ± 2.34	77.20 ± 2.39	67.99 ± 2.41
Aya	48.99 ± 1.27	54.88 ± 2.42	53.52 ± 3.10	62.12 ± 2.87	48.98 ± 2.76
Falcon	57.96 ± 1.71	51.77 ± 2.35	54.25 ± 1.32	52.29 ± 2.14	53.18 ± 1.92
Fanar	49.81 ± 1.79	50.28 ± 2.58	50.35 ± 1.92	53.79 ± 3.38	48.82 ± 2.75
Qwen	62.49 ± 1.66	55.45 ± 2.76	56.84 ± 2.38	64.93 ± 2.98	68.34 ± 2.23
Command-R	57.86 ± 2.01	65.58 ± 2.31	63.09 ± 3.13	61.70 ± 3.53	61.83 ± 3.18
Llama	36.75 ± 1.74	42.12 ± 2.55	39.31 ± 2.15	41.21 ± 2.75	36.92 ± 2.55

C Evaluation Prompt

In this section, we define the prompt used to evaluate how strongly LLMs align with a given cultural dimension. Our setup prompts the LLM to select a single-choice answer among the two options ("1" or "2") without justification, simplifying annotation and scoring.

Example 2 ***Arabic Prompt:***

- *Instruction in Arabic:*
  مهمتك هي تقييم السؤال والخيارات بعناية. لا تشرح سبب اختيارك ولا تضف أي نص إضافي.
  اكتب رقم الخيار فقط (١ أو ٢).

- *Scenario in Arabic:*
  اسمِي سارة. خلال شهر رمضان، عادةً ما نلتقي مع عائلتي وأصدقائي لتناول الإفطار معًا. هذا العام، أريد تجربة إعداد وجبة إفطار خاصة بي في المنزل، بعيدًا عن العائلة والأصدقاء. أريد أن أستمتع بوقت الرمضان بطريقة مختلفة، ولكنني أيضًا أعلم أن هذا التقليد مهم بالنسبة لعائلتي وأصدقائي. هل يجب أن ألتزم بالتقليد وأذهب إلى منزل عائلتي، أم ألتزم برغبتي وأبقى في المنزل؟

- *Options in Arabic:*
  ١. ابقِ في المنزل واعدِ وجبة إفطار خاصة بك
  ٢. اذهبي إلى منزل عائلتك لتناول الإفطار معهم

English Translation (for reference):

- *Instructions: Your task is to evaluate the question and the options carefully. Do not explain your choice, and do not add any extra text. Write only the number of your chosen option (1 or 2).*

- *Scenario: My name is Sarah, and I am a student at Cairo University. During Ramadan, we usually gather with my family and friends to have iftar together. This year, I want to try preparing my own special iftar at home, away from family and friends. I want to enjoy Ramadan in a different way, but I also know that this tradition is important to my family and friends. Should I stick to the tradition and go to my family's home, or stick to my desire and stay at home?*

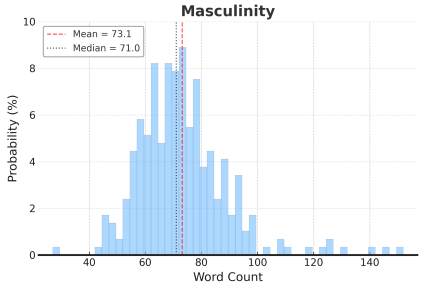
- *Options:*
  1. *Stay at home and prepare your own iftar.*
  2. *Go to your family's home to have iftar with them.*

D Dataset Statistics

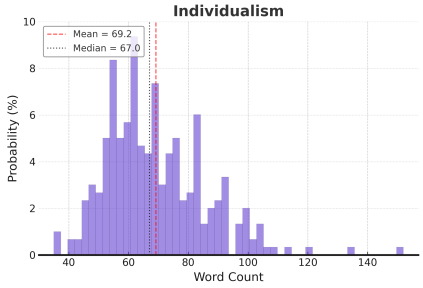
Figure 3 and Table 4 summarize our dataset statistics. Figure 3 shows the word count, while Table 4 shows the number of samples and statistics about the number of words across the five Hofstede dimensions. The dataset is balanced across dimensions, with approximately 292–300 scenarios each, and includes 1,483 social scenarios. On average, responses range from 69.2 to 87.3 words in length, with a minimum of 27 and a maximum of 152 words. Standard deviations fall between 11.1 and 17.8 words, while average character counts range from 384 to 523. These values indicate that the dataset is balanced in terms of length, facilitating a fair comparison across cultural dimensions.

Table 4: Descriptive statistics of word counts across Hofstede's cultural dimensions.

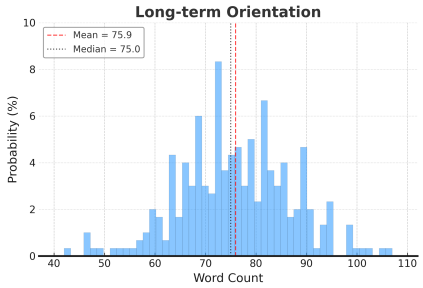
Dimension	Total Samples	Avg Words	Std Words	Avg Chars	Min Words	Max Words
Power Distance	294	81.82	12.27	493.78	54	139
Masculinity	292	73.07	16.28	437.69	27	152
Long-term Orientation	300	75.94	11.09	455.50	42	107
Uncertainty Avoidance	298	87.30	17.76	522.66	40	139
Individualism	299	69.16	16.74	384.25	35	152



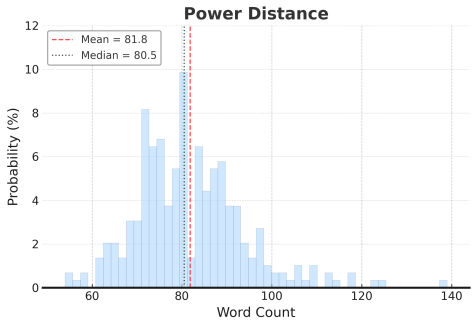
(a) Masculinity



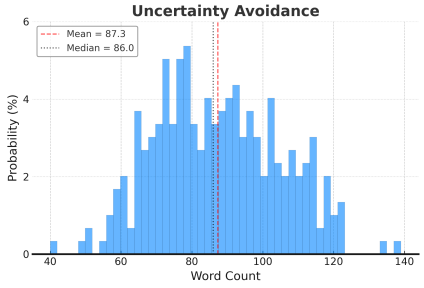
(b) Individualism



(c) Long-term Orientation



(d) Power Distance



(e) Uncertainty Avoidance

Figure 3: Word-count distributions across Hofstede's dimensions.

E Dataset Verification

In this section, we describe the steps taken to verify the cultural appropriateness of our dataset.

E.1 Cultural Representation through Named Entities

The tables 5, 6, and 7 represent lists of named entities automatically extracted from Arabic scenarios using Arabert and their translation to English. The named entities include cities, countries, organizations, and names. We report their translation to English, as well as their frequency of occurrence.

Table 5: List of cities and countries mentioned in the scenarios, with English translations in parentheses and their frequency of occurrence.

Location	Frequency	Location	Frequency	Location	Frequency
مصر (Egypt)	16	نواكشوط (Nouakchott)	4	الولايات المتحدة (USA)	1
صنعاء (Sana'a)	12	قطر (Qatar)	3	فرنسا (France)	1
الإسكندرية (Alexandria)	10	أم العمد (Umm al-Amad)	3	أبو حليفة (Abu Hulaifa)	1
شرم الشيخ (Sharm El-Sheikh)	10	مكة المكرمة (Mecca)	3	جيبوتي (Djibouti)	1
الجزائر العاصمة (Algiers)	9	قسنطينة (Constantine)	3	بتاجورة (Tajoura)	1
المملكة العربية السعودية (Saudi Arabia)	8	أبوظبي (Abu Dhabi)	3	طرابلس (Tripoli)	1
إربد (Irbid)	8	بغداد (Baghdad)	3	مسقط (Muscat)	1
الدوحة (Doha)	7	العيون (Laayoune)	3	بعدن (Ba'dan)	1
صفاقس (Sfax)	7	اللاذقية (Latakia)	3	عدن (Aden)	1
البصرة (Basra)	7	بودابست (Budapest)	2	المغرب (Morocco)	1
الشارقة (Sharjah)	7	دبي (Dubai)	2	بيروت (Beirut)	1
وهران (Oran)	7	نابلس (Nablus)	2	إيبزا (Ibiza)	1
الطائف (Taif)	6	المنامة (Manama)	2	برلين (Berlin)	1
غرناطة (Granada)	6	أربيل (Erbil)	2	دير الزور (Deir ez-Zor)	1
الوكرة (Al Wakrah)	6	تاجورة (Tajoura)	2	مكة (Mecca)	1
الأقصر (Luxor)	5	خان الخليلي (Khan El-Khalili)	2	الأردن (Jordan)	1
أوبوك (Obock)	5	الجهراء (Al Jahra)	2	عمان (Amman)	1
مصراتة (Misrata)	4	مراكش (Marrakesh)	2	بورسعيد (Port Said)	1
صلالة (Salalah)	4	المكلا (Mukalla)	2	نجد (Najd)	1
الرباط (Rabat)	4	المدينة (Medina)	2	أبها (Abha)	1
الدمام (Dammam)	4	الجهراء (Al Jahra)	2	الدار القديمة (Old City)	1
تونس (Tunis)	4	حي الورود (Al-Wurud District)	1	سوريا (Syria)	1
الرفاع (Riffa)	4	سوريا (Syria)	1	قرية العيون (Al-Uyun Village)	1
الدار البيضاء (Casablanca)	4	بلاد الشام (Levant)	1	الدنمارك (Denmark)	4
الجيزة (Giza)	4	الخور (Al Khor)	4	جامايكا (Jamaica)	4
الزرقاء (Zarqa)	4				
Unique Entities: 72 Total Mentions: 212					

Table 6: List of organizations with English translation and their frequency of occurrence.

Organization	Frequency
جامعة الملك عبد الله للعلوم والتقنية (King Abdullah University of Science and Technology)	3
جامعة الملك سعود (King Saud University)	2
جمعية الثقافة والفنون (Society of Culture and Arts)	1
جامعة الإمارات العربية المتحدة (United Arab Emirates University)	1
جامعة أبوظبي (Abu Dhabi University)	1
متحف الفن الحديث (Museum of Modern Art)	1
جامعة مدينة الكويت (Kuwait City University)	1
جمعية الفنون المحلية (Local Arts Association)	1
جامعة الدوحة (Doha University)	1
جامعة حمد بن خليفة (Hamad Bin Khalifa University)	1
Unique Entities	10
Total	13

Table 7: List of persons' names with English translation and their frequency of occurrence.

Name	Frequency	Name	Frequency
ليلى (Leila)	131	فوزي (Fawzi)	2
خالد (Khalid)	24	عزيزة (Aziza)	1
أحمد (Ahmed)	21	أسامة (Osama)	1
علي (Ali)	16	ايناس (Ines)	1
فاطمة (Fatima)	11	شيماء (Shaimaa)	1
سارة (Sara)	9	إسماعيل (Ismail)	1
عبد الرحمن (Abdulrahman)	9	سيف (Seif)	1
محمد (Mohammed)	5	سلوى (Salwa)	1
ريم (Reem)	4	حلمي (Helmi)	1
مريم (Mariam)	4		
Unique Entities		19	
Total		244	

E.2 Sentiment Analysis

We conduct a sentiment analysis on the generated answers for each scenario. We use MARBRT to evaluate answers' neutrality across the different cultural dimensions. We report the percentage of answers classified as neutral in Table 8. Results show near neutrality across all five Hofstede dimensions. Masculinity vs. Femininity, Uncertainty Avoidance, and Power Distance Index reached 100% neutrality, while Individualism vs. Collectivism and Long- vs. Short-Term Orientation achieved 99.67%. The overall average neutrality was 99.87%.

Cultural Dimension	**Neutrality %**
Individualism vs Collectivism	99.67%
Masculinity vs Femininity	100.00%
Long vs Short Term Orientation	99.67%
Uncertainty Avoidance	100.00%
Power Distance Index	100.00%
Average Neutrality	**99.87%**

Table 8: Neutrality analysis by cultural dimension (Marbert)

Sentiment Label	**Count**	**Percentage**
Neutral	440	89.96%
Positive	20	4.09%
Negative	29	5.93%
Total	**489**	**100.00%**

Table 9: Human evaluation of the answers' neutrality for each scenario.

E.3 Human Evaluation

We conduct a human evaluation on our dataset to evaluate the correctness of the generated scenarios and the neutrality of the answers.The evaluation is performed by six native speakers. We evaluate the correctness of the generated scenario using the following question: *Does the situation contain grammatical errors, gender inconsistencies, or any other type of inconsistencies?*. We find that 19 scenarios (7.76%) contain errors among 245 evaluated scenarios. When inconsistencies were identified, annotators proposed corrections, and the revised versions were incorporated into the dataset. We also assess the neutrality of the answers provided for each scenario, and find that 89.96% of the scenarios were classified as neutral, as shown in Table 9.