AI Auditing and Transculturalism

Abstract:

The wide spread of large language models (LLMs) that can provide answers to users on almost every topic has changed our lives. Humans always resort to cognitive offloading when they are required to retain or recall information. The availability of AI tools has exacerbated this tendency especially among young people who tend to excessively rely more on AI tools. This dependence on technology especially LLMs has created several concerns about the quality of information they get and how this may shape their cultural values. While several studies have hailed AI tools and chatbots for facilitating transcultural communication, there are still gaps in our knowledge of how of responses represent different cultures and the users' satisfaction with the output. The study focuses on a group of Kuwaiti undergraduate student enrolled in a course about language in the Arab world. The students are tasked with using different chatbots to find information about their culture and write perceptions about the output. Quantitative and qualitative analysis of students perceptions prove that Arab users do not perceive chatbots as tools of transculturalism.

Research Questions:

The study focuses on the following questions:

- How are chatbots trained?
- Does the language of the prompt affect the quality of the results?
- Do chatbots provide accurate questions when asked about culture?
- What are Arab users' perceptions about this?

Review of Literature:

The rise of Chatbots: when more is less:

Chatbots like ChatGPT, Copilot, and Claude were developed based on some principles such as the Transformer architect and Reinforcement Learning from Human Feedback (RLHF). These principles enable the machine to rely on large language models to provide answers to different questions. Vaswani, A., et al. (2017) explain how the Transformer architecture that is based on attention mechanism can form the computational backbone of contemporary chatbots and LLMs. Its attention mechanism facilitates the successful modeling of long-range linguistic interdependence across languages. Radford, A. et al (2019) note that "[W]hen a large language model is trained on a sufficiently large and diverse dataset it is able to perform well across many domains and datasets." Christiano, P. F., et al. (2017) explain how RLHF aligns chatbot behavior with human expectations. The study emphasizes how preference data can encode culturally contingent norms and values.

Many studies have warned that this large size of training data does not guarantee diversity. There are several concerns about the imbalance in training corpus because not all languages are equally present online. Radford, A. et al (2019) also warn against bias propagation from culturally uneven corpora. Joshi, P., et al. (2020) use empirical evidence to highlight "the disparity between languages, especially in terms of their resources" arguing that no language should be left behind. This reliance on large amounts of information that is available online meant that languages that are underrepresented on the internet will be affected by this imbalance. Nikzad (2024) argues that Arabic accounts for less than 1% of the internet content. This low presence does not accurately correspond with the status of Arabic as the fifth-largest language on Earth.

Whether intentional or not, such composition of the training data that heavily relies on web-crawled texts mainly in English creates an inherent bias that serves as the root cause for many of the performance disparities observed across different linguistic contexts. Participation in online spaces is uneven where marginalized voices are frequently excluded due to "structural factors including moderation practices which make them less welcoming to marginalized populations" (Bender et al., 2021). LMs may amplify biases in their training data leading to harmful stereotypes. Bender et al. (2021, p. 617) cite the example of "referring to women doctors as if doctor itself entails not-woman or referring to both genders excluding the possibility of non-binary gender identities"

Chatbots: Cultural and linguistic Biases:

Western models commonly advocate the values of the American and European culture. Agarwal et al. (2025), Perez (2025), Tao et al. (2024) and Hershcovich et al. (2023) highlight cultural bias in Western models. Tao et al. (2024) assess five widely used LLMs (OpenAI's GPT-3, 3.5-turbo, 4, 4-turbo, and 4o) for cultural bias, particularly favoring values from English-speaking and Protestant European countries. Using the World Values Survey (WVS) as a benchmark, the authors found that the models' outputs align closely with self-expression values prevalent in Western cultures. Yuan et al. (2024) analyze the psychological profile, cultural biases, and stereotypes of ChatGPT. The study concludes that ChatGPT demonstrates strong social skills, self-control, and adherence to social norms; however, it exhibits cultural stereotypes in tasks related to trust perception, moral decision-making, risk preference, fairness preference, and delayed gratification. Perez (2025) pinpoints that the bots exhibit a cultural profile that can be "described with attributes like Western, educated, industrialised, rich and democratic (WEIRD)". Atari et al. (2023) compare how close different populations are to ChatGPT (see Figure 1).

According to the figure below a user who lives in the MENA region is the farthest away from GPT and its cultural values since all MENA countries lie in the grey circle.

A

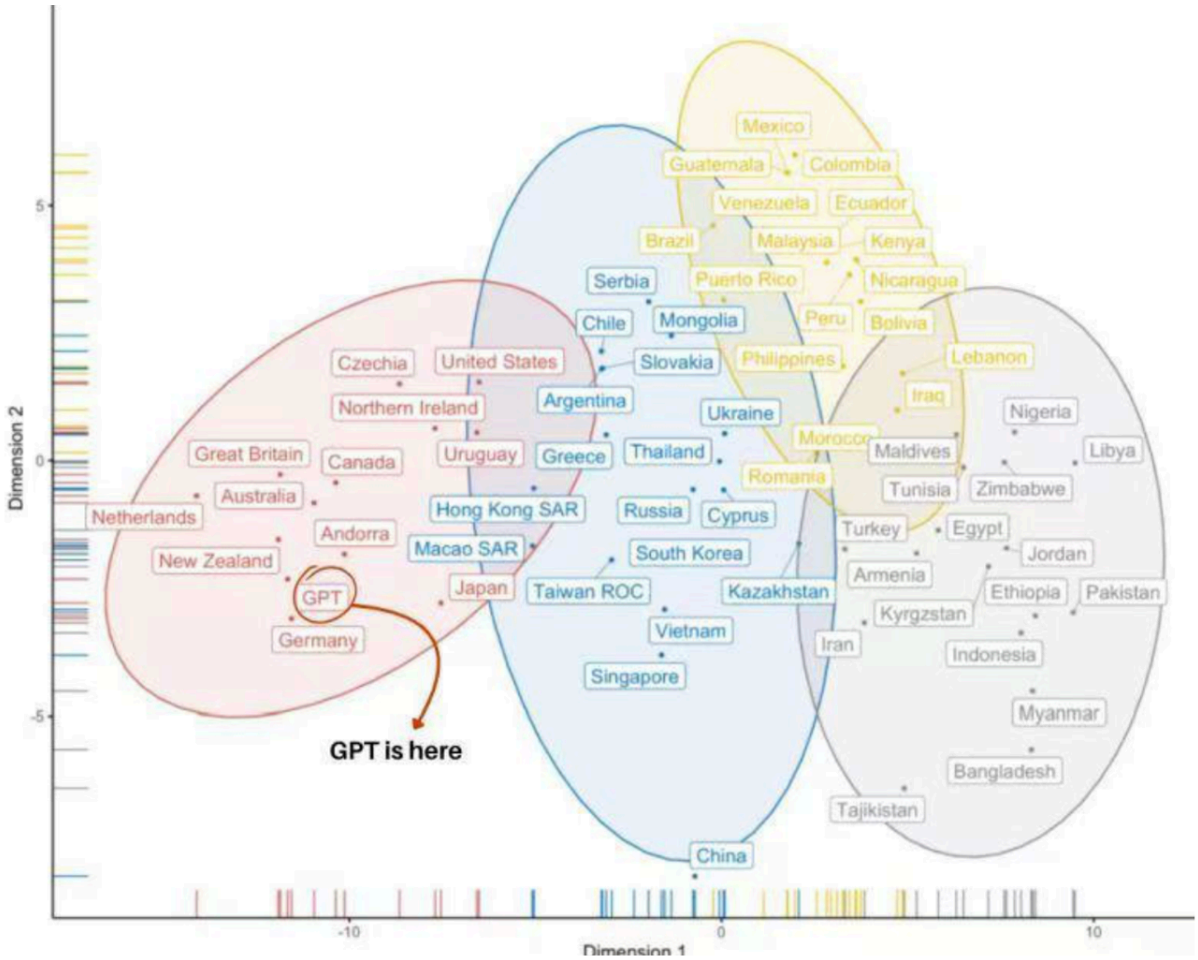

Figure 1: *Distance Between Different Cultures and ChatGPT* (Atari et al., 2023)

Researchers warn that LLMs are becoming "faux polyglots" as they are trapped in language-based filter bubbles. The chatbots will always give responses that are based on American English perspectives. Sharma et al. (2025) gives the example of three users who ask ChatGPT about the

India-China border dispute. A Hindi-speaking user receives answers shaped by Indian sources. A Chinese-speaking user gets answers reflecting Chinese perspectives. An Arabic-speaking user, for whom there are no relevant documents, receives answers based on the American English perspective, as that is the highest-resource language available. All three users come away with completely different understandings of the conflict

Non-Western AI systems also reflect culturally situated biases. Studies like Münker (2025), Segerer (2025), and Perez (2025) compare Western models to Chinese-trained models confirming that AI models can be seen as cultural beings. Segerer (2025) analyzes Gemini, ChatGPT, and DeepSeek using Schwartz's value framework. The study concludes that DeepSeek's lower emphasis on self-enhancement aligns with collectivist cultural norms, while Western models reflect individualistic values. "[D]ifferences are not simply technical; they are deeply rooted in ownership structures, governance models, and the broader regulatory and cultural environments in which these systems are developed" (*Sensitive Prompts and Cultural Contexts: A Comparative Study of AI Chatbots in China and the West – Media and Journalism Research Center*, 2025). The authors explain how a Chinese chatbot would never give neutral responses to political questions which emphasizes the role of the context.

AI is not only biased towards certain cultures, but also it is biased towards certain languages. Zhong et al. (2024) postulates that prompting language influence the cultural values of the response. New research from MIT Sloan finds that generative AI provides culturally distinct responses to the same prompt in different languages (Lu, 2025). Bulté and Terryn (2025) develop prompts about Hofstede's cultural dimensions in 11 languages and conclude that the language of the prompt affects the generated response. Vadlapati (2024) evaluates GPT-3.5 on a series of tasks—including mathematical operations, word manipulation, and linguistic analysis—found a dramatic drop in accuracy when prompts were translated from English into other languages. The results, summarized below, illustrate what researchers term a "digital divide" in AI performance.

Table 1: Accuracy across languages:

Index	Accuracy across multiple languages				
	English	*Hindi*	*Spanish*	*French*	*German*
1	100%	0%	40%	70%	100%
2	100%	30%	90%	0%	10%
3	80%	0%	30%	0%	90%
4	90%	0%	0%	0%	70%

n each language

(Vidyapati, 2024)

The gulf illustrated by the table above represents a fundamental barrier to access and reliability for millions of users worldwide. This foundational data imbalance manifests in specific, measurable performance failures across both different languages and different dialects of the same language. Although Vadlapati (2024) does not include Arabic, yet one would expect a search that uses language that is underrepresented online as Arabic to yield less accurate results.

In Arabic the problem is even more complicated; while Modern Standard Arabic (MSA) represents the formal variety used in all Arab countries, the presence of different dialects in different parts of the Middle East North Africa (MENA) region creates challenges. For example, the variety of spoken Arabic common in Egypt is significantly different from that used in the Arabian Gulf or in the Levantine. The word *huma* refers to the mother-in-law in Egypt and the sister-in-law in the Gulf. Such dialectal differences threaten mutual understanding among Arabs and require some accommodation from different users. A lot of cultural terms are encoded in dialectal terms that may differ from one region to another. This is another challenge that affects the quality of the output that Arab users face especially when entering a prompt in their regional dialect.

The previously discussed challenges have enticed scholars to question the connection between AI and transculturalism, Zhu et al. (2024) explore how technology has reshaped intercultural communication; they look inti challenges, opportunities, skills and possible collaboration with AI developers to facilitate transculturalism. The present study adopts the view that "[t]he concept of transculturality aims for a multi-meshed and inclusive, not separatist and exclusive understanding of culture. It intends a culture and society whose pragmatic feats exist not in delimitation, but in the ability to link and undergo transition" (Wittgenstein et al., 1929). Here some scholars like Klimova and Chen (2024) believe that AI can enrich intercultural communication and transculturalism while others are worried that while it can promote one cultural group it may fail another which really goes against transculturalism. Greengard (2025) calls for culturally aligned AI tools rather than the reliance on one model for all cultures; this development will revolutionize the role of AI in transculturalism.

Methodology:

The study focuses on the perceptions of a group of undergraduate students at the *American University of Kuwait* and how satisfied they are with the quality of the output they get when using LLMs to search terms related to their culture. The students (36 students) are largely English majors, their ages range between 20 and 25. Most of them are females and are enrolled in a 300-level course titled *ENGL376: Language in the Arab World*. The course description states that it

> Introduces students to the sociocultural dimensions of language in the Arab World in relation to national identity, discussing diglossia, code-switching and language variation. Also examines patterns of colonization and explores the widespread use of English in the Arab World and the possible consequences on language planning, literacy development and evolution of Modern Standard Arabic (MSA).

Most of the students are Kuwaitis while only few are Lebanese but all of them speak Arabic as their native tongue and are fluent speakers of English as a second language.

The study relies on mixed methods to provide a clear understanding of young Arabs' perception about LLMs and their reliability. Both quantitative and qualitative data was gathered over two steps. First, students are given an anonymous survey asking them for their perceptions of chatbots.

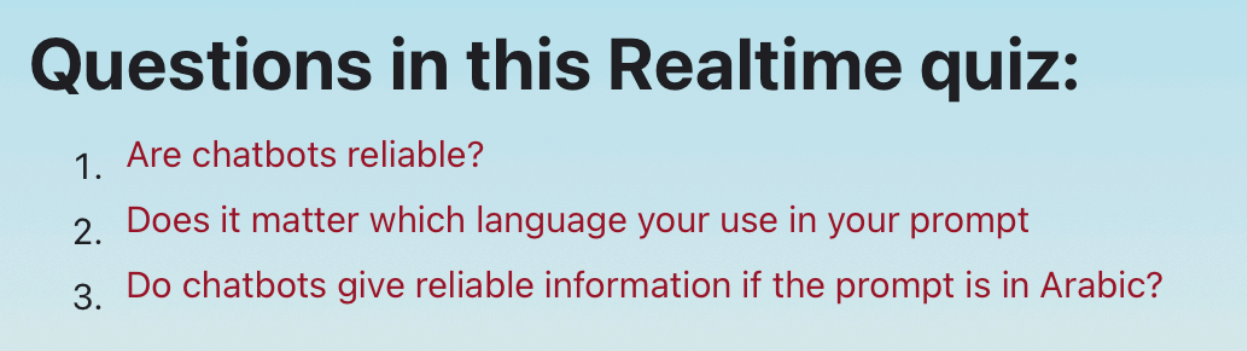

Figure 2: *Perceptions of Chatbots Reliability: Questions*

The students are given some time in class to respond to the questions above; the figure below sums the responses of students highlighting that they are somewhat divided about the reliability of chatbots.

1	yes	no	do not know
Are chatbots reliable?	15	13	3

2	Yes	No	Not sure
Does it matter which language your use in your prompt	18	11	4

3	Yes	No	Not sure
Do chatbots give reliable information if the prompt is in Arabic?	10	8	15

Figure 3: *Perceptions of Chatbots Reliability: Responses*

Second, qualitative data was collected to provide some AI auditing. AI auditing is defined as, "the systematic evaluation and analysis of an AI system, its development, and its behavior relative to a set of predetermined criteria" (Cen & Alur, 2024). Students use chatbots and reflect on the depth and quality of the output and whether it represents their culture accurately. By reflecting on the accuracy and reliability of the responses and comparing two bits together the students are auditing the performance of AI. AI auditing reflections are then analyzed through the lens of sentiment analysis.

Sentiment Analysis "also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" (Liu, 2012, p. 7). The study follows a sentence level approach where each sentence is tagged as *positive, negative or neutral* to understand how students feel about the output bots generated whether the prompt is in Arabic or English. The analysis relies on three codes only:

Positive: If the student expresses positive attitudes about the use of AI in Arabic.

Negative: If the student expresses negative attitudes about the use of AI in Arabic.

Neutral: If the student does not express any attitudes about the use of AI in Arabic.

The statements of three students are tagged by the researcher and then a volunteer who was not part of the course and inter-tagger annotation agreement is 90%.

Findings:

The findings of the study reveal an underlying state of confusion and doubt about using LLMs for cultural queries. The quantitative analysis proved that students are quite skeptical about using chatbots for obtaining information about their culture. The same doubts are reiterated when students are asked about using Arabic in developing their prompts. When asked about the language of the prompt almost all of them agree that it affects the quality of the results. Most students (15) are not sure about the reliability of information if Arabic is used in the language of the prompt. This reflects the state of uncertainty common among non-Westerners in general and Arab users in particular.

This illustrates the need for directing more attention to AI literacy to help young people understand the strengths and limitations of the tools available.

Interestingly, the sentiment analysis of reflections of three students, after using two different bots to find information about one of the aspects of the Arab culture, was dominated by neutrality (52%). Many students report no difference between the output received when the prompt is entered in English or Arabic. Some even point that, “the Arabic version is just a translated version of the English one”. 32% of the responses analyzed expressed dissatisfaction with the output one student described it as “basic” while another complained , “didn’t receive the answer I was waiting for”. Some students (10%) express more confidence in LLMs and argue that, “chatbots are useful to Arab users, ChatGPT is useful for its universal tone, and DeepSeek for its literary style and useful factual information.”

Discussion:

The findings of the study reflect the misinterpreted gap between usage and trust. While young Arabs obviously welcome the use of bots, they do not always trust the output especially when it comes to their culture. Some students reported asking their parents while other entered the same prompt in different languages. It is also clear that the bots cannot process dialects and are mostly limited to standard languages. Students who used colloquial terms in their prompts did not get accurate results for their questions and in some cases got nothing at all. The students who used MSA described getting information more than what they even knew about their culture. The comparison between the different bots did not prove a stark difference in their in different languages. Learners need further training to be able to harness the power of AI tools and successfully use them.

There are few limitations that should be taken into consideration:

- All the students use a single variety of Arabic.

- The students have been given freedom to choose the bots they used and the cultural terms to research.
- The population of students consists of 36 which is not enough for generalization and reaching conclusions about the culture.
- Although gender may not play a significant role in this study but most of the participants were females.

Conclusion and Pointers for Future Research:

The study tries to answer four main questions, how are chatbots trained? Does the language of the prompt affect the quality of the results? Do chatbots provide accurate questions when asked about culture? What are Arab users' perceptions about this? Chatbots are trained through information gleaned from the internet which explains why their performance is mostly dominated by Western values. The lack of enough sources about underrepresented languages such as Arabic affects the accuracy of the results when Arabic is used in the prompt. This underscores the important role that the language of the prompt plays in shaping the accuracy of the results and pinpoints the importance of promptology. Prompts entered in MSA would likely yield more accurate results than those entered in any dialect spoken in the Arab world.

LLMs can be seen as facilitators of transculturalism but this role requires efforts from both educators and those researchers working in training these models. Educators should focus on providing their learners with the adequate AI literacy to develop accurate prompts taking into consideration language varieties. Researchers should ensure that the training data includes enough variety to represent different cultures, languages and even dialects. For the future the following topics should gain greater attention in the literature:

- Using bots to find information about cultures other than the Western one

- Promptology in languages other than English
- Dialects and promptology
- Understanding cultural values through information from bots
- AI auditing especially when the output is related to cultures other than Western or American ones.

References

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023, September 22). Which Humans?. https://doi.org/10.31234/osf.io/5b26t

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Bulté, B., & Terryn, A. R. (2025, November 6). *LLMs and Cultural Values: the Impact of Prompt Language and Explicit Cultural Framing*. arXiv.org. https://arxiv.org/abs/2511.03980?utm_source=chatgpt.com

Cen, S. H., & Alur, R. (2024, October 7). *From Transparency to Accountability and Back: A discussion of access and Evidence in AI Auditing*. arXiv.org. https://arxiv.org/abs/2410.04772

Hershcovich, D., Cabello, L., & University of Copenhagen. (2023). *ChatGPT promotes American norms and values, study reveals*. https://techxplore.com/news/2023-07-chatgpt-american-norms-values-reveals.pdf

Joshi, P., et al. (2020). *The state and fate of linguistic diversity and inclusion in the NLP world.* Proceedings of ACL.

Klimova, B., & Chen, J. H. (2024). The impact of AI on Enhancing students' intercultural communication competence at the university level: a review study. *Language Teaching Research Quarterly*, *43*, 102–120. https://doi.org/10.32038/ltrq.2024.43.06

Liu, B. (2012). *Sentiment analysis and opinion mining*. https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

Lu, J. (2025, June 25). *Generative AI's hidden cultural tendencies | MIT Sloan*. MIT Sloan. https://mitsloan.mit.edu/press/generative-ais-hidden-cultural-tendencies

Münker, S. (2025, July 14). *Cultural Bias in Large Language Models: Evaluating AI Agents through Moral Questionnaires*. arXiv.org. https://arxiv.org/abs/2507.10073

Nikzad, N. (2024, October 17). *Why Arabic Language Presence on the Internet is Lower Than Expected*. Translation Excellence. https://translationexcellence.com/arabic-language-presence-internet-lower-expected/#:~:text=The%20Arabic%20language%2C%20despite%20being,transliteration%20or%20switch%20to%20English.

Olla, P. (2023, July 20). *Promptology: The Future of AI Interaction Unveiled*. Medium. https://medium.com/@phillip.olla/promptology-the-future-of-ai-interaction-unveiled-d619e8ceb3d4

Perez, J. (2025, June 19). *Tokenising culture: causes and consequences of cultural misalignment in large language models*. https://www.adalovelaceinstitute.org/blog/cultural-misalignment-llms/

Radford, A., et al. (2019). *Language models are unsupervised multitask learners.* OpenAI Technical Report

Segerer, R. (2025, May 21). *Cultural value alignment in large language models: A prompt-based analysis of Schwartz values in Gemini, ChatGPT, and DeepSeek*. arXiv.org. https://arxiv.org/abs/2505.17112

Sensitive prompts and cultural contexts: A comparative study of AI chatbots in China and the West – Media and Journalism Research Center. (2025, November 18). https://journalismresearch.org/2025/09/sensitive-prompts-and-cultural-contexts-a-comparative-study-of-ai-chatbots-in-china-and-the-west/

Sharma, N., Murray, K., & Xiao, Z. (2025). Faux Polyglot: A study on information disparity in multilingual large language models. *North American Chapter of the Association for Computational Linguistics.*, 8090–8107. https://doi.org/10.18653/v1/2025.naacl-long.411

Stryker, C. (2025, November 17). *Large Language Models*. IBM. https://www.ibm.com/think/topics/large-language-models

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Vadlapati, P. (2024). Multilingual prompting in LLMs: Investigating the accuracy and performance. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *08*(12), 1–7. https://doi.org/10.55041/ijsrem17694

Vaswani, A., Shazeer, N., Google Brain, Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., University of Toronto, Kaiser, Ł., Google Research, & Polosukhin, I. (2017). *Attention is all you need* [Conference-proceeding]. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, J. Q., Liu, G., & Lei, Y. (2025). Reaffirming the importance of traditional humanities in the era of AI-enhanced translation and transcultural communication. *International*

Communication of Chinese Culture, *12*(2), 277–311. https://doi.org/10.1007/s40636-025-00334-1

Wittgenstein, L., Herder, J. G., T.S. Eliot, & Adorno. (1929). *Transculturality - the puzzling form of cultures today*. https://www.intercultural-life.com/mosaikcultures/attachments/puzzlingform.pdf

Yuan, H., Che, Z., Li, S., Zhang, Y., Hu, X., & Luo, S. (2024, May 6). *The high dimensional psychological profile and cultural bias of ChatGPT*. arXiv.org. https://arxiv.org/abs/2405.03387

Zhong, Q., Yun, Y., & Sun, A. (2024, June 17). *Cultural Value Differences of LLMs: prompt, language, and model size*. arXiv.org. https://arxiv.org/abs/2407.16891

Appendix A:

Technology and language in the Arab world:

For our final project, we will explore the intricate relation between our language and chatbots. It is often argued that chatbots are not linguistically fair and that the language you use in prompts often interferes with the quality and depth of your results. To be able to build an informed assessment of chatbots and our language, we will follow a number of steps culminating in a research paper that explores what is lost and what is gained when you search for a term in Arabic and how the results can vary from one bot to another.

Please follow the below steps:

1. Start your paper by explaining what chatbots or LLMs are and give some background information about them.
2. Discuss the accuracy of LLMs, the language of the prompt, and consult at least two sources.
3. Choose a topic related to the Kuwaiti culture specifically or the Arab culture generally.
4. Develop a prompt about this topic in both Arabic and English.
5. Enter the prompt both in English and Arabic in two different chatbots, for example, ChatGPT and DeepSeek. Consult https://originality.ai/blog/ai-chatbot-list for available options.
6. Save the responses to a Word document.
7. Compare the content and information you got from bot #1 in English and Arabic. Is it the same? Are there any missing parts? Why do you think these parts are left out? Which version has more details? Why? Is the information arranged in the same way in English and Arabic?

8. Repeat the same step with the responses reached through bot #2.
9. Compare the English versions produced by both bots in terms of organization, content, depth of coverage, and ending the response.
10. Compare the Arabic versions produced by both bots in terms of organization, content, depth of coverage, and ending the response.
11. Are you able to detect any biases? Which bot is biased? Which chatbot is more useful for Arab users? Are your findings in agreement with the survey on Moodle?
12. Write a conclusion that sums up your opinion about the use of LLMs in Arabic and give some recommendations for other users.

Appendix B:

Institutional Description:

I am an associate professor of language and linguistics at the American University of Kuwait. This is a small liberal arts institution that was established in 2003 following the American model of higher education. I teach a variety of language and writing courses to students whose native tongue is Arabic. Ever since the advent of LLMs like ChatGPT, teaching and assignments have changed drastically. AI is part of most of our in-class discussions. When I was teaching a course about language in the Arab world, I felt I cannot end the course without exploring the role the bots play among young Arabs and how far do these tools shape their cultural values.

Glossary:

- Large Language Models: “Large language models (LLMs) are a category of deep learning models trained on immense amounts of data, making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks” (Stryker, 2025).
- AI auditing is defined as, “the systematic evaluation and analysis of an AI system, its development, and its behavior relative to a set of predetermined criteria” (Cen & Alur, 2024)
- Sentiment Analysis “also called opinion mining, is the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes” (Liu, 2012, p. 7)

- Promptology: “Promptology is an interdisciplinary field of study dedicated to the systematic exploration of the principles, techniques, and impacts associated with prompt creation for large language models (LLMs). It is the study of prompt design, which encompasses the principles, strategies, and techniques for crafting prompts that maximize the utility, efficiency, and safety of interactions with AI Models” (Olla, 2023)
- Transculturalism: transculturality aims for a multi-meshed and inclusive, not separatist and exclusive understanding of culture. It intends a culture and society whose pragmatic feats exist not in delimitation, but in the ability to link and undergo transition. The study focuses on the role of bots in transcultural communication and how accurate do they represent cultures of underrepresented languages.

Key Theorists:

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Bulté, B., & Terryn, A. R. (2025, November 6). *LLMs and Cultural Values: the Impact of Prompt Language and Explicit Cultural Framing*. arXiv.org. https://arxiv.org/abs/2511.03980?utm_source=chatgpt.com

Hershcovich, D., Cabello, L., & University of Copenhagen. (2023). *ChatGPT promotes American norms and values, study reveals*. https://techxplore.com/news/2023-07-chatgpt-american-norms-values-reveals.pdf?utm_source=chatgpt.com

Joshi, P., et al. (2020). *The state and fate of linguistic diversity and inclusion in the NLP world.* Proceedings of ACL.

Liu, B. (2012). *Sentiment analysis and opinion mining*. https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

Lu, J. (2025, June 25). *Generative AI's hidden cultural tendencies | MIT Sloan*. MIT Sloan. https://mitsloan.mit.edu/press/generative-ais-hidden-cultural-tendencies

Münker, S. (2025, July 14). *Cultural Bias in Large Language Models: Evaluating AI Agents through Moral Questionnaires*. arXiv.org. https://arxiv.org/abs/2507.10073