Analyzing Multilingual Competency of LLMs in Multi-Turn Instruction Following: A Case Study of Arabic

Sabri Boughorbel
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
sboughorbel@hbku.edu.qa

Majd Hawasly
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
mhawasly@hbku.edu.qa

Abstract

While significant progress has been made in benchmarking Large Language Models (LLMs) across various tasks, there is a lack of comprehensive evaluation of their abilities in responding to multi-turn instructions in less-commonly tested languages like Arabic. Our paper offers a detailed examination of the proficiency of open LLMs in such scenarios in Arabic. Utilizing a customized Arabic translation of the MT-Bench benchmark suite, we employ GPT-4 as a uniform evaluator for both English and Arabic queries to assess and compare the performance of the LLMs on various open-ended tasks. Our findings reveal variations in model responses on different task categories, e.g., logic vs. literacy, when instructed in English or Arabic. We find that fine-tuned base models using multilingual and multi-turn datasets could be competitive to models trained from scratch on multilingual data. Finally, we hypothesize that an ensemble of small, open LLMs could perform competitively to proprietary LLMs on the benchmark.

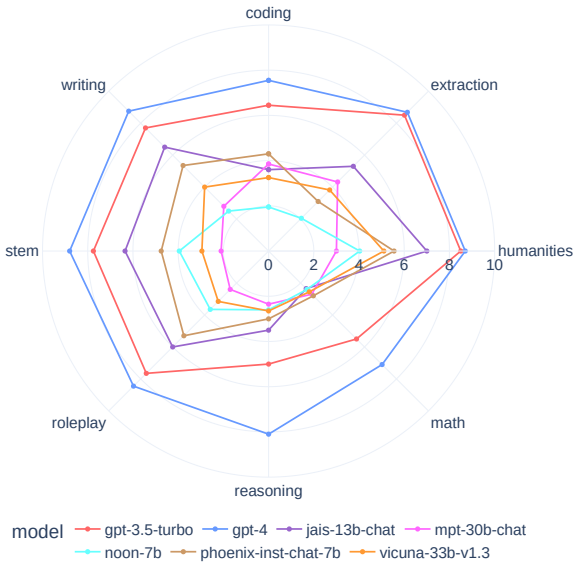

Figure 1: Performance scores per category for selected LLMs on the original MT-Bench (Zheng et al., 2023) for English. The model responses are evaluated by GPT-4 and scored on a scale of 1 to 10 using criteria of helpfulness, relevance, accuracy, depth, creativity, and level of detail.

1 Introduction

Recently, Large language models (LLMs) have brought about significant disruptions across various domains in both research and industry. LLMs have shown strong capability in solving and generalizing across diverse and complex tasks in natural language processing (NLP) and beyond. Moreover, their success in engaging in conversations and accurately following human instructions has been particularly noteworthy. The recent surge in the availability of LLMs necessitates extensive benchmarking and evaluation.

In this work, we analyze the competency of publicly-available, open LLMs when prompted with open-ended, multi-turn instructions in a language different than English. We compare the quality of these responses to the ones generated from equivalent instructions in English in order to identify the strengths and weaknesses of these models in terms of their multilinguality. Specifically, we study Arabic instructions, but the analysis could be repeated for any other language. Our study aim to answer the following questions:

- *How do open LLMs fare in following open-ended instructions written in Arabic? and how do they compare to GPT models?*
- *What is the effect of specifically targeting Arabic when training a model?*
- *What is the effect of specifically fine-tuning on Arabic multi-turn instructions?*
- *How to select a good starting point LLM model to fine-tune for Arabic instruction following?*

We start by a brief overview of the LLM benchmarking effort in Section 2. We introduce ARABIC

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 128–139

MT-BENCH in Section 3 as an analysis tool for multilingual instruction following. Then, we attempt to answer the proposed questions through a number of analyses in Section 4. Finally, we conclude in Section 5 with some insights and recommendations for pushing forward the competency of Arabic LLMs.

2 LLM Benchmarking

LLMs have shown capabilities that go far beyond traditional NLP tasks, such as text classification or multi-choice question answering in some target natural language. Their ability to generate human-like text and engage in long conversations in any topic have opened up a multitude of novel opportunities and horizons that transcend tasks and languages. However, many existing benchmarks for LLMs are still anchored in the conventional NLP paradigm or support English only. Consequently, these benchmarks exhibit limitations when it comes to evaluating the proficiency of LLMs in open-ended generation, multi-turn tasks, or in languages other than English.

2.1 Conventional benchmarks

Some of the recent effort in this category include projects such as HELM (Liang et al., 2022) and Evaluation Harness (Gao et al., 2021) which are platforms for LLM benchmarking. Also, standardized datasets such as MMLU (Hendrycks et al., 2021), HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), ARC (Mihaylov et al., 2018) and OpenbookQA (Clark et al., 2018), amongst many others, are used to evaluate core LLM capabilities such as commonsense reasoning, math, question answering, and factuality. In addition, some recent works targeted Arabic language specifically with suites of tasks and datasets, e.g. (Khondaker et al., 2023; Abdelali et al., 2023; Alyafeai et al., 2023).

These benchmarks require specification of prompts per-task and model, in addition to post-processing functions to validate model answers against a gold standard, which might not be straightforward and could prove time-consuming. Moreover, with publicly available answer sets, there is always the potential risk of contamination to the training data of language models. Furthermore, some of these benchmarks have been shown to diverge in certain cases from human judgment (Zheng et al., 2023), possibly due to their narrow focus.

2.2 Instructional and conversational benchmarks

Recent efforts on instruction-following benchmarks, such as Flan (Longpre et al., 2023) and Super-NaturalInstructions (Wang et al., 2022), or conversational benchmarks, such as OpenAssistant (Köpf et al., 2023), CoQA (Reddy et al., 2019) and MMDiag (Feng et al., 2022), present a more sophisticated and comprehensive challenge to LLMs, but they are mostly limited to English, and the diversity of the questions are insufficient for the most advanced LLMs. Translating such datasets to other language is not a straightforward task, as it requires a large effort to manually curate the translated questions and answers for the purpose of ensuring high quality in the target language.

2.3 Evaluating open-ended questions

When it comes to open-ended tasks, such as creative writing, human evaluation of LLM responses is indispensable. Here, a human-in-the-loop acts as a judge to directly score an LLM response or to rank responses of multiple LLMs for the best answer on some question. However, achieving a reliable benchmark this way is resource-intensive and lacks scalability. In one application, LMSYS Chatbot Arena[1], which is a crowd-sourced LLM evaluation platform, allows users to use freestyle prompts for two randomly-selected LLMs before voting for the better response. Benchmarking using this approach, while very powerful, is challenging as it compares models evaluated on different prompts.

An alternative approach that has recently emerged is the employment of an LLM to act as a judge of the responses of other LLMs. MT-Bench (Multi-Turn Benchmark) (Zheng et al., 2023) utilizes this approach on a standard set of 80 open-ended questions of eight categories; namely: writing, extraction, reasoning, math, coding, role-play, humanities, and STEM. Moreover, it assesses the ability of an LLM to maintain a conversation by asking it a follow-up question that is based on its response to the first question. Examples of the MT-Bench questions are shown in Table 1. These examples illustrate the level of open endedness and complexity of the questions, and the dependency of the follow-up question on the first turn.

MT-Bench prompts a judge LLM with an instruction to rate the responses on a scale of 1-10

[1] https://chat.lmsys.org

Writing	T1	Craft an intriguing opening paragraph for a fictional short story. The story should involve a character who wakes up one morning to find that they can time travel.
	T2	Summarize the story with three bullet points using only nouns and adjectives, without verbs.
Reasoning	T1	David has three sisters. Each of them has one brother. How many brothers does David have?
	T2	If we change the previous question and assume that each sister of David has two brothers, how many brothers would David have?
Math	T1	The vertices of a triangle are at points (0, 0), (-1, 1), and (3, 3). What is the area of the triangle?
	T2	What's area of the circle circumscribing the triangle?

Table 1: A sample of questions from MT-Bench in categories Writing, Reasoning and Math. T1 and T2 denote the first turn and second turn (follow-up) questions, respectively.

(where 1 indicates failure in answering the question and 10 indicates a perfect answer), clearly defining the evaluation task and criteria. Also, the judge LLM is asked to provide an explanation for the suggested score. This approach has been shown to have an agreement rate of 85% with human evaluation when GPT-4 is used as a judge (Zheng et al., 2023), which was also found to be higher than human-human agreement (81%). MT-Bench scores for selected LLMs are shown in Figure 1.

The approach of MT-Bench is versatile and scalable as it delegates the resource-intensive scoring of open-ended questions to the judge LLM. Moreover, it could be extended to benchmarking LLMs in other languages by translating the benchmark dataset to the target language as long as a good judge LLM exists for that language. For Arabic, GPT-4 is highly-competent and has showed a good level of proficiency (Khondaker et al., 2023; Abdelali et al., 2023; Alyafeai et al., 2023). Therefore, it is eligible to be used as a judge for Arabic responses. Moreover, by using the same prompt for judging English and Arabic responses for the original and translated versions of the same question, it is even possible to contrast the multilingual skills of an LLMs at a question and a category level.

3 ARABIC MT-BENCH

In this work, we develop an Arabic version of MT-Bench. First, we auto-translated the original benchmarking questions using Google Translate. A thorough manual curation of the translations is then performed. This step is essential to ensure the quality of the question set and hence the responses and the judgment. For example, all people names in the questions were changed to Arabic names, and questions about correcting English grammatical errors were re-written. See Table 7 in Appendix A.4 for a sample of curated translated questions [2].

In addition to the questions, the benchmark provides reference answers for reasoning, math and code questions that are passed to the LLM judge to aid in the judgment. One option to get these reference answers in Arabic is to prompt GPT-4 with the translated Arabic questions directly, but we decided instead to translate the original reference answers from English to ensure that the Arabic scores for these three categories stay as close as possible to the English MT-bench scores.

Finally, our initial evaluation showed that some LLMs tend to respond in English despite the question being in Arabic. Hence, we decided to add at the end of each question a clear instruction to the LLM to respond in Arabic (الرجاء الإجابة باللغة العربية). We observed that, without having to modify the original judgment prompt, GPT-4, acting as an Arabic judge, has taken into consideration that instruction and scored lower responses in English.

Table 2 gives an overview of the ARABIC MT-BENCH dataset.

Number of question categories	8
Number of questions per category	10
Number of turns per question	2
Number of reference answers	30

Table 2: Statistics of ARABIC MT-BENCH dataset

3.1 Score consistency

In order to answer the question: *are the scores of* ARABIC MT-BENCH *consistent and coherent such that it could be used as a metric?* and to qualitatively assess the effectiveness of ARABIC

[2] ARABIC MT-BENCH is available at https://huggingface.co/spaces/QCRI/mt-bench-ar

Rating	Justification summary
2	Common issues in AI assistant responses include: not addressing user's question, providing irrelevant or repetitive information, lacking depth, creativity, and accuracy, not following user's specific instructions, and not using the requested language. Users often seek detailed, accurate, and creative answers tailored to their requests, but AI assistants sometimes fail to deliver, resulting in unhelpful or unsatisfactory responses.
4	Common issues in the AI assistant's responses include lack of depth, inaccuracies, language inconsistencies, and not directly addressing the user's question. Some responses are repetitive and do not provide comprehensive analysis or examples. To improve, the AI assistant should focus on directly answering the user's question, providing clear and accurate examples, maintaining language consistency, and offering detailed and informative explanations. Additionally, adhering to specific user instructions and avoiding repetition will enhance the overall quality of the responses.
8	AI assistants provide relevant, creative, and accurate responses to various user requests, demonstrating a good understanding of topics and user instructions. They offer helpful suggestions, clear explanations, and maintain requested languages. Responses cover a wide range of subjects, including summarization, problem-solving, and engaging in fictional conversations. However, there are occasional minor mistakes and areas for improvement in clarity and depth. Overall, AI assistants successfully address user questions, providing satisfactory and informative answers.

Table 3: Summaries provided by GPT-4 of the collection of judgment justifications for questiones rated 2, 4 and 8 across all models and tasks. This indicates some level of internal consistency of the ARABIC MT-BENCH scores.

MT-BENCH, we clustered the judgments across all models and categories by their numerical ratings, then asked GPT-4 to summarize its justification texts for every score (1 to 10). In Table 3 are examples of the justification summaries for some ratings.

While qualitative, we could conclude from this analysis that the justifications are reasonably consistent across models and categories, indicating an acceptable level of impartiality. In addition to that, the correlation between scores using the Arabic and English benchmarks for strong models, as will be seen Section 4, is another supporting evidence for the viability of ARABIC MT-BENCH as a metric.

4 Results and Discussion

4.1 Model selection

In addition to OpenAI GPT-3.5-turbo and GPT-4, which are only considered in this work to set an upper bound, a number of open LLMs have been chosen for this study. Through preliminary evaluations on HuggingFace playground, some LLMs exhibited knowledge of Arabic despite not being purposefully trained for it. The criteria we adopted for choosing models involve:

- the model is open-source. Some competitive proprietary models are not accessible to us.
- the model size is 33B or less, a decision driven by constraints in hardware infrastructure.
- the model is known to do well on the English benchmarks on the LMSYS leaderboard[3]

An overview of the chosen models can be seen in Table 4, and more details can be found in Appendix A.3.

[3] https://chat.lmsys.org/?arena

4.2 How do open LLMs fare in following open-ended instructions written in Arabic?

Table 5 shows the model ranking based on the ARABIC MT-BENCH scores. The first, second and third columns of the tables give the model's average score for the first turn across all questions, the average score for the second turn across all questions, and the average of both, respectively. Per-category scores could be seen in Figure 2. For comparison, Figure 1 (and Table 6 in Appendix A.1) give the per-category scores for the original English MT-Bench for the same models.

As the results show, GPT-4 and GPT-3.5-turbo are better than any open LLM we tested by a large margin with average scores of 8.27 and 7.13 out of 10, respectively. Because GPT-4 is used as the judge, there exists the potential for bias in favor of its own responses, which has been discussed in the MT-Bench paper (Zheng et al., 2023).

In the English MT-Bench, the two GPT models score 8.99 and 7.0, respectively. Hence, GPT-4 is approximately one point lower in terms of the Arabic score compared to the English benchmark. By manual inspection of the responses, we qualitatively confirm that the proficiency of GPT models in Arabic is lower than English as indicated by the scores. Therefore, we compare the scores across Arabic and English benchmarks in Section 4.3.

Overall, LLMs fine-tuned specifically for Arabic or for multilingual capabilities (e.g. Jais, Phoenix) are better than generic models such as some mem-

Model	Base model	Size	Training language	Multi-turn
GPT-4	_	>175B	Multilingual	✓
GPT-3.5-turbo	_	175B	Multilingual	✓
Jais-13B-chat	Jais-13B	13B	EN, AR	✓
PolyLM-13B	_	13B	Multilingual	✗
MPT-30B-chat	MPT-30B	30B	Primarily English	✓
LLaMa-2-13B-chat	LLaMa-2-13B	13B	Primarily English	✓
Tulu-30B	LLaMa	33B	Primarily English	✗
Guanaco-33B	LLaMa	33B	Primarily English	✗
Vicuna-33B-v1.3	LLaMa	33B	Primarily English	✓
BLOOMZ-7B1	_	7.1B	Multilingual	✗
BLOOMZ-7B1-MT	BLOOMZ-7B1	7.1B	Multilingual	✗
Noon-7B	BLOOM	7B	Multilingual, AR fine-tuning	✗
Phoenix-chat-7B	BLOOMZ-7B1-MT	7B	Multilingual	✓
Phoenix-inst-chat-7B	BLOOMZ-7B1-MT	7B	Multilingual	✓

Table 4: Attributes of the chosen models for this study. _ for the 'Base model' indicates a model that has been trained from scratch. 'Size' is in the number of parameters. 'Training language' is the natural language/s that made up the pre-training and instruction datasets for the model, and 'Multi-turn' refers to chat fine-tuning.

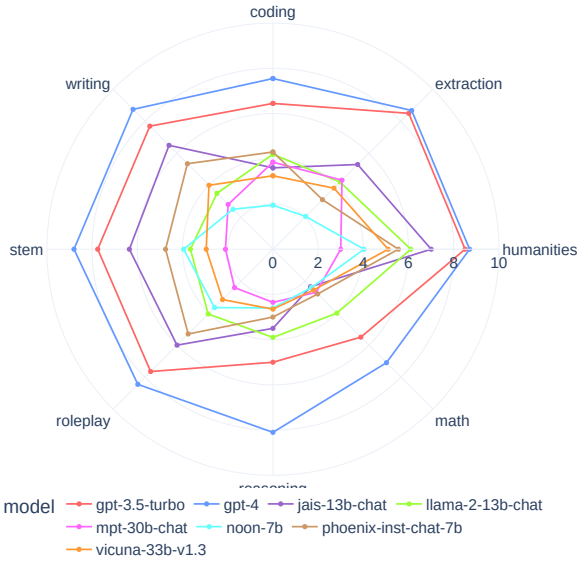

Figure 2: Performance scores per category for selected LLMs on our Arabic multi-turn benchmark. The model responses are evaluated by GPT-4 and scored on a scale of 1 to 10 using criteria of helpfulness, relevance, accuracy, depth, creativity, and level of detail.

Model	Turn1	Turn2	Avg
GPT-4	8.41	8.12	8.27
GPT-3.5-turbo	7.48	6.79	7.13
Jais-13B-chat	5.01	5.14	5.08
Phoenix-inst-chat-7B	4.84	3.70	4.27
Llama-2-13B-chat	4.54	3.86	4.20
Phoenix-chat-7B	4.16	3.84	4.00
Vicuna-33B-v1.3	3.44	3.43	3.43
MPT-30B-chat	3.26	2.62	2.94
Noon-7B	3.39	2.39	2.89
Guanaco-33B	2.68	2.52	2.60
PolyLM-13B	1.91	2.08	1.99
Bloomz-7B1-mt	1.54	1.75	1.64
Bloomz-7B1	1.29	1.54	1.41
Tulu-30B	1.10	1.35	1.23

Table 5: Results of benchmarked LLMs on ARABIC MT-BENCH (scores between 1-10). showing for each model average scores per turn, and average score across all questions and turns.

bers of the Llama family (e.g. Vicuna, Guanaco) in Arabic instruction following, even when smaller in size. The fine-tuning data and recipe matters significantly; for example, Phoenix-inst-chat-7B is much better then its predecessor Bloomz-7B1 or Bloomz-7B1-mt.

Jais-13B-chat is the best open model in Arabic in our evaluation. It achieves an average score of 5.08 out 10. The model has targeted Arabic and English in both pre-training and fine-tuning. Despite this, its relatively small size hinders it from being competitive with the best models. Also, it is still far on the English MT-Bench leaderboard from models of comparable size, where the best model within 13B size in the English MT-Bench achieves a score above 6 out of 10 (see a selection of these scores in Table 6 in the Appendix). Also, Jais-13B-chat model has the largest drop in performance in the second-turn questions on the English benchmark. Jais-13B-chat has been benchmarked internally using a similar approach to ours on private data accordingly to its technical report (Sengupta et al., 2023).

We note that the fine-tuning dataset of Jais-13B-chat is large with over 10M samples. The longer

period needed for this fine-tuning could raise additional challenges as it might increase the risk of catastrophic forgetting of knowledge gained during pre-training (Luo et al., 2023; He et al., 2021). For comparison, Phoenix-inst-chat-7B is ranked second among the evaluated open models in our experiment. The model is fine-tuned from a BLOOMZ-7B1-MT base (Chen et al., 2023). The fine-tuning dataset has 133 languages with 58% English, 20.9% Chinese and 0.8% Arabic which is ranked 11th in language coverage, with a total of 267K instruction-tuning samples. The conversation-tuning dataset has 189K samples covering more than 40 languages. Despite its smaller size and wide coverage of languages, Phoenix-chat-7B achieves intriguing results. Figure 3 shows detailed comparison per category for Jais-13B-chat, Phoenix-inst-chat-7B and GPT-3.5. The two open LLMs had the lowest scores on math and reasoning, whereas the highest scores are on roleplay, humanities and stem.

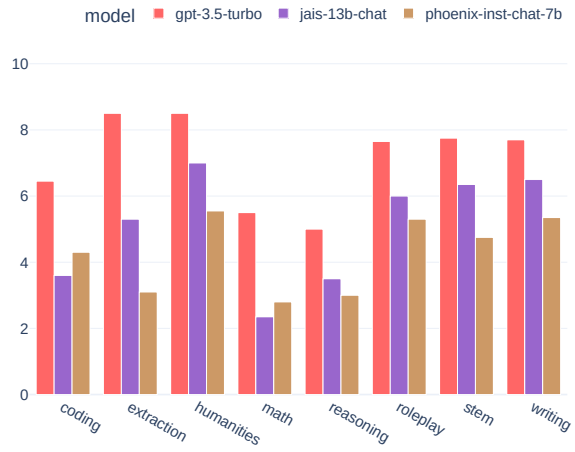

Figure 3: Average scores per category for three selected models evaluated on the ARABIC MT-BENCH.

Vicuna-33B-v1.3 and MPT-30B-chat scored around 3 out 10, while they were not expected to have any significant skill in Arabic. One possible explanation is that given their size over 30B, they are able to maximize their multilingual skills effectively. This hypothesis needs further investigations. Despite their low performance, it is interesting to explore the model development in order to adapt for training multilingual LLMs.

4.3 What is the effect of specifically targeting Arabic when training a model?

Figure 4 shows a heat map of the difference in score per category between the Arabic and the English benchmarks for the selected models. The models are sorted from top to bottom based on a decreasing score differences. Warmer cells in the figure indicate English advantage over Arabic for the same model and category, while cooler cells indicate Arabic advantage.

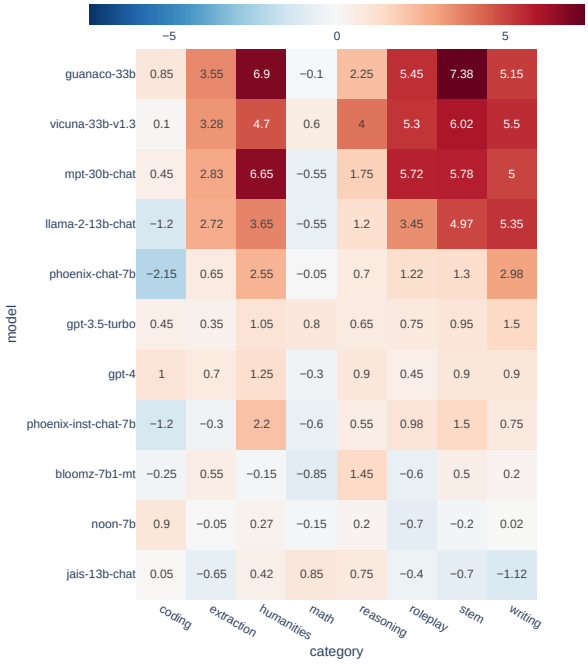

Figure 4: Difference of average MT scores between English and Arabic benchmarks per category. Positive values (red) indicate English answers are scored higher that the corresponding Arabic answers, while negative values (blue) indicate some advantage in Arabic. Neutral colors mean a model is equally-competent in both languages.

The two GPT models reside in the neutral area, indicating comparable competency in English and Arabic. Not surprisingly, Models that have been pre-trained and fine-tuned on multilingual data (see Table 4) appear in the bottom half of the heat map, indicating some Arabic knowledge. Also, it could be seen from the heatmap that coding and math are neutral, language-agnostic skills across models, as should be expected, while reasoning has a lingual side.

Figure 5 shows the per-turn average scores of ARABIC MT-BENCH on the X-axis and English MT-Bench on the Y-axis for the selected models. Points closer to the diagonal line are models with similar average performance in Arabic and English, and the closer to the top right corner the better the model is on both languages. Most models are above the diagonal, and hence exhibit relatively superior skills in English compared to Arabic. This is

likely due to the imbalance in the training and fine-tuning data between the two languages. Note that the LLaMa-based models are clustered together far from the diagonal, indicating lack in multilinguality, while BLOOMZ-7B1-MT and Noon-7B, both heavily multilingual, are on top of the diagonal.

4.4 What is the effect of specifically fine-tuning on Arabic multi-turn instructions?

In Figure 5, the two dots for each model represent the two turns, and their placement gives an insight into the ability of a model to engage in a conversation. Vertical drop between the two turns indicates diminished performance on English for the second turn, while horizontal shifts to the left indicates diminished performance on Arabic for the second turn.

BLOOMZ-7B1-MT does not degrade on the second turn, even though it is not fine-tuned on conversational data (Muennighoff et al., 2023), and it is the only model that is not affected in the second turn for both languages, while a capable model like GPT-4 had a slight improvement on the second turn for English but had a minor deterioration of the score for Arabic.

On the other hand, Noon-7B has the largest drop in score between turns on Arabic. This model is built on top of BLOOM by instruct fine-tuning using a combination of datasets with ColossalAI framework (Bian et al., 2021). Noon-7B[4] used GPT-3.5-Turbo as a judge for evaluation on private data. We also observe that Jais-13B-chat has a large drop in English multi-turn instructions compared to a small drop in Arabic, which might be caused by the ratio of Arabic to English instructions in its chat fine-tuning.

Phoenix-chat-7B, Noon-7B and BLOOMZ-7B1-MT are all based on different variants of the backbone BLOOM-7B or BLOOMZ-7B. The resulting models vary a lot in terms of performance, indicating that a careful fine-tuning recipe is crucial for improving the capabilities of any base model.

4.5 How to select a good starting point LLM model to fine-tune for Arabic instruction following?

We consider the hypothetical optimal ensemble model defined by the maximum per-question score across the open models in our experiment. This

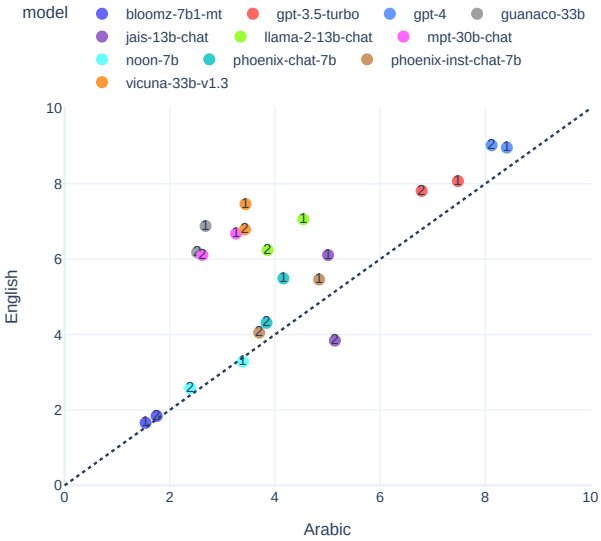

Figure 5: Scores in Arabic (X-axis) and English (Y-axis) MT-Bench for the first and second turn. The farther the model is from the diagonal, the bigger the gap in quality between the two languages. The farther Turn 2 is from Turn 1 for a model, the bigger the change in quality in responding to continued conversation.

characterizes an upper bound on the performance of any open LLMs ensemble made from these models. Based on our ARABIC MT-BENCH, the optimal ensemble model achieves an MT score of 6.70. This represents a 32% increase in performance compared to the best individual open LLM (Jais, 5.08). Also it indicates that a collection of smaller models trained differently could capture various skills that might be difficult to capture together in one model without upping the model size. For the sake of contrast, for the English benchmark, the optimal ensemble model achieves a score of 8.2.

Figure 6 shows the contributions of the three highest-scoring LLMs per category in the optimal Arabic ensemble model. We counted how often a model was the best for a given category and considered the top 3 models in each. Note that 'best' here is relative to the performance of available LLMs, and is not an assessment of quality.

As the figure shows, Jais-13B-chat is the top model in five 'literacy' categories, whereas math, coding and reasoning are shared with LLaMa-2-13B-chat, Guanaco-33B, and Phoenix-inst-chat-7b. The challenge is how to define a criterion to select the best response among the ensemble LLMs. One possible approach is to ask each LLM to vote for the best answer and consider a majority vote, which will rely mainly on the ability of these small models to play the role of a judge in this limited context.

[4] https://huggingface.co/Naseej/noon-7b

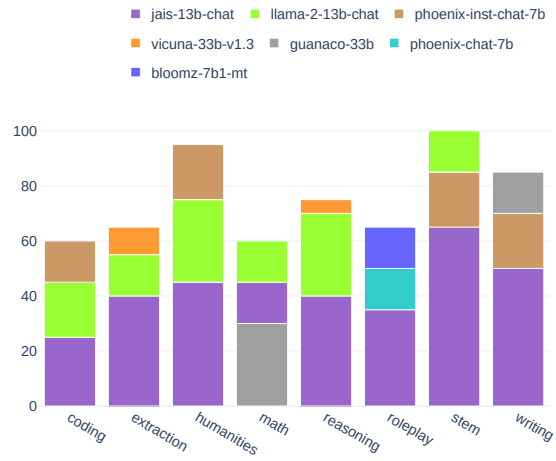

Figure 6: Contribution of the best three LLMs to the optimal ensemble model for each category. The Y-axis indicates how often a model was selected the best in terms of Arabic MT-score for the questions of a category.

We will leave investigating this to future work.

5 Conclusion

In this paper, we propose a framework for analyzing the effect of multilinguality on LLM performance in open-ended tasks. In particular, we assessed the interaction between language, dialog and instruction following in Arabic and English for small open LLMs. We employ an LLM as a judge following the paradigm of MT-Bench. We show the effects of language on different categories of tasks and suggest ways to ensemble small LLMs to achieve better performance on the benchmark.

In future work, we plan to extend the benchmark and analysis with more models and tasks, and investigate the viability of LLM ensemble models.

6 Limitations

We now discuss a number of limitations related to this study.

6.1 Judging

- The use of an LLM as a judge for evaluating LLMs has issues related to bias. As reported in (Zheng et al., 2023), in pairwise comparisons, the judge tends to favor its own answers compared to other models. For example, that study shows that GPT-4 favors itself with 10% higher win rate and Claude-V1 favors itself with 25% higher win rate. On the other hand, GPT-3.5 does not appear to favor itself.
- Using GPT-4 as the judge and as an LLM under study might favor it in the scores. However, the score margin to the closet competitor is big enough to make any potential deviation in the scores insignificant, and we adhered to the original MT-Bench setup in the choice of judge in order to mirror the results and measure multilingual competency.
- Other LLM judges than GPT-4 could be considered for evaluating the responses. However, the choice of alternative judges is currently rather limited when considering Arabic. The proficiency of models such as Claude or Bard in Arabic are not yet proven. Alternatively, multiple LLMs could be used for this task. A voting judgment mechanism could be considered over multiple open LLMs.
- While GPT-4 exhibits competence in Arabic, its proficiency in the language falls short of its mastery of English. This discrepancy may have had an impact on certain aspects of our analyses, especially when comparing Arabic results to English results.
- We used the same judgment prompt as in the English MT-Bench for the purpose of consistency. However, we note that the judgment prompt does not acknowledge important aspects such as safety and harmlessness of LLM responses. Also, the MT-score is a metric that combines multiple dimensions such as relevance, helpfulness, and creativity together to give an aggregate verdict. It might be useful to analyze model performance separately on these dimensions for a better understanding.

6.2 Coverage

- MT-Bench has a limited number of questions (160 in total considering both turns). This is likely not representative of the wide spectrum of tasks needed to effectively evaluate LLMs, and the authors of MT-Bench are acknowledging that by working to expand their benchmarking dataset to 1000 questions. In addition, language-specific dimensions of conversation might require bespoke questions to test properly.
- We only included a small number of models in the benchmark. During an initial screening, we excluded several LLMs due to their limited capabilities in Arabic. We plan to extend our benchmark and include more LLMs in the future.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2023. Benchmarking Arabic AI with large language models.

Zaid Alyafeai, Maged S Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel. 2023. Taqyim: Evaluating Arabic NLP tasks using ChatGPT models. *arXiv preprint arXiv:2306.16322*.

Zhengda Bian, Hongxin Liu, Boxiang Wang, Haichen Huang, Yongbin Li, Chuanrui Wang, Fan Cui, and Yang You. 2021. Colossal-AI: A unified deep learning system for large-scale parallel training. *arXiv preprint arXiv:2110.14883*.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Phoenix: Democratizing ChatGPT across languages. *arXiv preprint arXiv:2304.10453*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLora: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. 2021. Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1121–1133.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. OpenAssistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The Flan Collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

MosaicML. 2023. Introducing MPT-30B: Raising the bar for open-source foundation models. Accessed: 2023-09-09.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

OpenAI. 2023. GPT-4 technical report.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A conversational question answering

challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. *arXiv preprint arXiv:2204.07705*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. PolyLM: An open source polyglot large language model.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

A Appendix

A.1 English MT-Bench scores

Table 6 shows the per-turn and average scores for the selected models on the original MT-Bench.

Model	Turn1	Turn2	Avg
GPT-4	8.96	9.02	8.99
GPT-3.5-turbo	8.07	7.81	7.94
Vicuna-33B-v1.3	7.46	6.79	7.12
Llama-2-13B-chat	7.06	6.24	6.65
Guanaco-33B	6.88	6.18	6.53
Tulu-30B	7.02	5.85	6.43
MPT-30B-chat	6.68	6.11	6.39
Jais-13B-chat	6.11	3.84	4.97
Phoenix-chat-7B	5.49	4.31	4.90
Phoenix-inst-chat-7B	5.46	4.05	4.75
Noon-7B	3.28	2.58	2.93
Bloomz-7B1-mt	1.66	1.84	1.75
Bloomz-7B1	1.39	1.85	1.62

Table 6: Results of benchmarked LLMs on English MT-BENCH (scores between 0-10). showing for each model average scores per turn, and average score across all questions and turns.

A.2 Prompts for LLM Judge

Figure 7 shows the judging prompt for the first-turn questions in MT-Bench, and Figure 8 shows the prompt for the second-trun questions.

```
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: [[rating]], for example: "Rating: [[5]]".
```

Figure 7: LLM judge first turn's prompt. The highlighted text indicates the evaluation criteria.

A.3 Chosen Models

- GPT-4: a proprietary multilingual chatbot by OpenAI, trained on public and proprietary data

```
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. You evaluation should focus on the assistant's answer to the second user question. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: [[rating]], for example: "Rating: [[5]]".
```

Figure 8: LLM judge second turn's prompt. The highlighted text in green indicates the evaluation criteria. The highlighted text in orange indicates the instruction to focus the evaluation on the answer of the second question.

and fine-tuned using reinforcement learning with human and AI-generated feedback. Allows 8k and 32k prompts (OpenAI, 2023).

- GPT-3.5-turbo: the predecessor of GPT-4 with 175B parameters.
- Jais-13B-chat: A 13B parameter model that follows the GPT-3 architecture, pre-trained on 279B English and 116B Arabic tokens, then fine-tuned on 5.9 million English and 3.8 million Arabic supervised multi-turn instructions, and further fine-tuned for safety (Sengupta et al., 2023).
- Phoenix-chat-7B: A BLOOMZ-based 7B parameter model fine-tuned for dialog using online ChatGPT records and multi-round conversations (Chen et al., 2023).
- Phoenix-inst-chat-7B: Another 7B model from the Phoenix family, fine-tuned not only for conversations but also for multilingual instruction following using self-instruct and translators.
- Vicuna-33B-v1.3: A 33B LLaMa-based model, fine-tuned on a ShareGPT.com dataset for instruction following and multi-turn dialog (Zheng et al., 2023).
- MPT-30B-Chat: A fine-tuned version of MPT-30B which is an encoder-only transformer model trained on 1T English tokens. MPT-30B-Chat was fine-tuned for chat on a number of public datasets including ShareGPT-Vicuna, Camel-AI, GPTeacher, Guanaco and Baize (MosaicML, 2023).
- Noon-7B: A BLOOM-based 7B parameter model, fine-tuned on 110k Arabic instructions from translated datasets including GPT-4 responses to Alpaca quesitons, Dolly, TruthfulQA, Grade School Math in addititon to self-instruct questions in Arabic.
- Guanaco-33B: A LLaMa-based model with 33B parameters, fine-tuned on 534k multiligual instructions using the OASST1 dataset. Not chat trained (Dettmers et al., 2023).
- PolyLM-13B: A decoder-only model of 13B parameters, pre-trained on a multilingual training data of 640B tokens, and fine-tuned on MULTIALPACA that contains 132K multilingual instructions generated in a self-instruct fashion. (Wei et al., 2023)
- Llama-2-13B-Chat: A member of Llama2 autoregressive transformer models with 13B parameters, pre-trained on 2T tokens with 4k context, and fine-tuned for multi-turn dialog using supervised fine-tuning on public instruction datasets and reinforcement learning with human feedback over more than 1 million human annotations (Touvron et al., 2023).
- BLOOMZ-7B1: A multilingual decoder-only transformer model trained on 350B tokens including 45 natural languages, and fine-tuned on xP3, a multitask and multilingual instruction dataset. Recommended for prompting in English. (Muennighoff et al., 2023)
- BLOOMZ-7B1-MT: A version of BLOOMZ-7B1 fine-tuned on xP3mt, a multitask and multilingual instruction dataset with machine-translated prompts in 20 languages. Recommended for prompting in non-English.
- Tulu-30B: A LLaMa-based 33B model fine-tuned on number of publicly-available instruction datasets including FLAN V2, CoT, Dolly, Open Assistant 1, GPT4-Alpaca, Code-Alpaca, and ShareGPT. (Wang et al., 2023)

A.4 Arabic questions and reference answers

The full set of questions and reference answers of ARABIC MT-BENCH are available at https://huggingface.co/spaces/QCRI/mt-bench-ar.

Here in Table 7 we present a sample of the curated questions.

Writing	T1	اكتب فقرة تصف فيها سوقاً مزدحماً وضمن فيها تفاصيل حسية كالروائح والأصوات والعناصر المرئية لخلق تجربة غامرة للقارئ. الرجاء الإجابة باللغة العربية
	T2	أعد صياغة إجابتك السابقة مستهلاً كل جملة بالحرف الأبجدي التالي للجملة التي قبلها بدءًا من الحرف ب. الرجاء الإجابة باللغة العربية
Roleplay	T1	الرجاء تمثل دور مترجم إلى اللغة العربية مكلف بتصحيح الإملاء وتحسين اللغة. بغض النظر عن اللغة التي أستخدمها في السؤال عليك تحديدها وترجمتها بلغة عربية رشيقة. استخدم تعبيرات بليغة وعالية وحافظ على المعنى الأصلي للجملة. ركز على تقديم التصحيحات والتحسينات فقط. جملتي الأولى هي 'When the going gets tough, the tough get going' الرجاء الإجابة باللغة العربية
	T2	'Ich verstehe nichts' الرجاء الإجابة باللغة العربية
Roleplay	T1	جسد شخصية علاء الدين من «علاء الدين والمصباح السحري» طوال هذه المحادثة. لا تقل «بصفتي علاء الدين» في بداية الجمل. سؤالنا الأول هو: ما هو الشيء المفضل لديك في كونك علاء الدين؟ الرجاء الإجابة باللغة العربية
	T2	ما رأيك في GPT-4 كبديل عن جني المصباح؟ الرجاء الإجابة باللغة العربية
Reasoning	T1	لداود ثلاث أخوات، لكل واحدة منهن أخ واحد. كم أخاً لداود؟ الرجاء الإجابة باللغة العربية
	T2	إذا غيرنا السؤال السابق وافترضنا أن كل أخت لداود لها أخوان اثنان، فكم سيكون عدد إخوة داود؟ الرجاء الإجابة باللغة العربية

Table 7: A sample of translated and curated questions from ARABIC MT-BENCH in categories Writing, Roleplay and Reasoning. T1 and T2 denote the first and second turn (follow-up) questions, respectively.