TounsiBench: Benchmarking Large Language Models for Tunisian Arabic

Souha Ben Hassine[1], **Asma Arrak**[1], **Marouene Addhoum**[2], **Steven R. Wilson**[1]
[1]University of Michigan-Flint, [2]Oakland University
souhabh@umich.edu, asmaar@umich.edu,
marroueneaddhoum@oakland.edu, steverw@umich.edu

Abstract

In this work, we introduce the first benchmark for evaluating the capabilities of large language models (LLMs) in understanding and generating responses in Tunisian Arabic. To achieve this, we construct a dataset of Tunisian Arabic instructions and prompt ten widely-used LLMs that claim to support Arabic. We then assess the LLM responses through both human and LLM-based evaluations across four criteria: quality, correctness, relevance, and dialectal adherence. We analyze the agreement and correlation between these judgments and identify GPT-4o as our automated judge model based on its high correlation with human ratings, and generate a final leaderboard using this model. Our error analysis reveals that most of the LLMs that were evaluated struggle with recognizing and properly responding in Tunisian Arabic.

To facilitate further research, we release our dataset, along with gold-standard human-written responses for all 744 instructions, and our evaluation framework, allowing others to benchmark their own models.

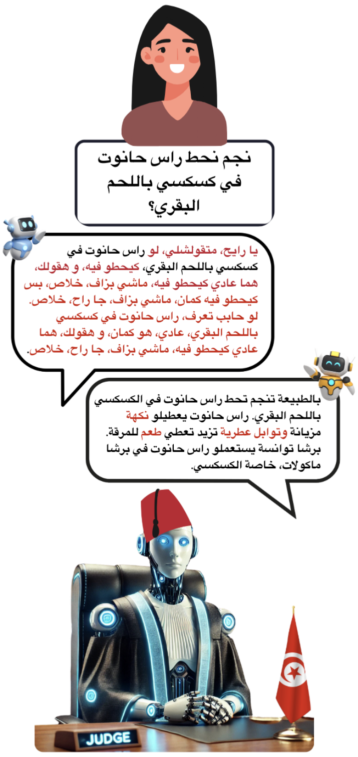
نجم نحط راس حانوت
في كسكسي باللحم
البقري؟
يا رايح، متقولشلي، لو راس حانوت في
كسكسي باللحم البقري، كيحطو فيه، و هقولك،
هما عادي كيحطو فيه، ماشي بزاف، خلاص، بس
كيحطو فيه كمان، ماشي بزاف، جا راح، خلاص.
لو حابب تعرف، راس حانوت في كسكسي
باللحم البقري، عادي، هو كمان، و هقولك، هما
عادي كيحطو فيه، ماشي بزاف، جا راح، خلاص.
بالطبيعة تنجم تحط راس حانوت في الكسكسي
باللحم البقري. راس حانوت يعطيلو نكهة
مزيانة وتوابل عطرية تزيد تعطي طعم للمرقة.
برشا توانسة يستعملو راس حانوت في برشا
ماكولات، خاصة الكسكسي.
JUDGE

Figure 1: **Evaluation of LLM Responses to a Tunisian Arabic Prompt:** The user is asking if they can put "Ras el Hanout"—a spice mix known in Tunisia and neighboring countries—into couscous with beef. The responses shown are from Llama-3.1-8b-instruct (first response) and GPT-4o (second response). We evaluate the extent to which an LLM can judge the quality, relevance, correctness, and dialectal adherence of these responses. The judge model assesses whether the responses stay true to Tunisian Arabic, both lexically and culturally. We used red font to highlight phrases or wordings that are not in Tunisian Arabic (e.g., borrowed from MSA, Maghrebi, or Levantine dialects).

1 Introduction

Tunisian Arabic, known colloquially to its speakers as *Derja*, is a unique and dynamic dialect spoken by over 12 million people in Tunisia (Wik, 2025; Eberhard et al., 2024). Unlike Modern Standard Arabic (MSA), which is primarily used in formal settings such as media, literature, and official communications, Tunisian Arabic is an evolving spoken language with significant phonetic, lexical, and syntactic differences from MSA. It exhibits substantial influence from languages such as French, Italian, Turkish, and Berber, making it distinct from other Arabic dialects. Moreover, Tunisian Arabic lacks a standardized writing system, leading to variability in orthographic representation, particularly in digital communication where Latin-based scripts and Arabic scripts coexist. These linguistic characteristics present unique challenges for natural language processing (NLP) models tasked with understanding and generating text in Tunisian Arabic. Despite the increasing interest in Arabic NLP, most research has focused on MSA or high-resource dialects such as Egyptian and Levantine Arabic (Guellil et al., 2021).

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34627–34642
November 4-9, 2025 ©2025 Association for Computational Linguistics

Tunisian Arabic remains underrepresented in existing language models and datasets, limiting the development of effective NLP applications for the dialect. Given its linguistic uniqueness and practical significance in daily communication, evaluating large language models (LLMs) on their ability to comprehend and generate Tunisian Arabic is crucial for advancing Arabic dialect processing.

In this paper, we introduce the first benchmark for Tunisian Arabic by systematically evaluating LLMs on their ability to generate responses in this dialect. To achieve this, we construct a dataset of Tunisian Arabic instructions through two complementary approaches: (1) crowdsourcing naturally occurring chatbot queries from native speakers and (2) translating a subset of the HuggingFace Helpful Instructions dataset into Tunisian Arabic. The resulting dataset comprises 744 unique instructions covering a wide range of topics. We then prompt ten widely used LLMs that claim Arabic language support, instructing them to generate responses strictly in Tunisian Arabic. To assess the quality of these responses, we conduct both human- and LLM-based evaluations using four criteria: quality, correctness, relevance, and adherence to Tunisian Arabic. We analyze the agreement and correlation between human and LLM-based rankings and define GPT-4o as our judge model, as explained in Figure 1, to generate the final leaderboard. Additionally, our analysis reveals that most LLMs struggle with Tunisian Arabic, particularly in recognizing and adhering to the dialect. Our findings highlight the challenges of processing underrepresented Arabic dialects and emphasize the need for improved dialect-specific NLP solutions. We release our evaluation code and dataset along with gold-standard human-written responses for all 744 instructions, allowing anyone to easily benchmark their own models.

2 Related work

The evaluation of natural language generation systems has traditionally relied on human annotators or reference-based automatic metrics like BLEU, ROUGE, and METEOR. However, human evaluation is costly, time-consuming, and prone to inconsistencies among evaluators. More recently, there has been a shift toward reference-free evaluation techniques that leverage high-performing (LLMs) to assess generated text.

The foundation for LLM-based evaluation was laid by Chiang and Lee (2023), who explored replacing human assessment with LLM-driven methods. Studies such as AlpacaEval (Li et al., 2023), Arena-Hard (Li et al., 2024), and WildBench (Lin et al., 2024) demonstrated that LLM-based evaluation can serve as an efficient and scalable alternative to traditional human judgment. Expanding on these efforts, Fu et al. (2023) introduced GPTScore, showcasing how LLMs can provide flexible, task-specific evaluation criteria. Further investigations (Wang et al., 2023; Zheng et al., 2023) revealed biases and inconsistencies in LLM-based evaluation, particularly in non-English contexts.

While LLM-based evaluation has become increasingly popular, research has shown that LLMs struggle in multilingual settings, particularly for low-resource languages. Zhao et al. (2024) found that LLMs often internally translate non-English inputs into English before reasoning and translating responses back, raising concerns about fidelity and bias. Kew et al. (2024) and Shaham et al. (2024) demonstrated that minimal multilingual fine-tuning can enhance cross-lingual generalization, yet performance disparities persist. These findings highlight the need for comprehensive multilingual benchmarks that better capture linguistic diversity.

Recent benchmarking efforts have revealed significant LLM performance gaps in low-resource languages. Khondaker et al. (2024) found that LLaMA-3.3-70B underperforms closed models in both Modern Standard Arabic (MSA) and dialectal Arabic. Bhatia et al. (2024) introduced ArabicMTEB, a benchmark designed to evaluate Arabic embeddings across dialects and domains. Similarly, Adelani et al. (2025) proposed IrokoBench for 17 African languages, exposing stark disparities between high- and low-resource languages. These efforts emphasize the pressing need for better evaluation frameworks tailored to underrepresented linguistic communities.

Despite increasing attention to Arabic NLP, research on Tunisian Arabic remains scarce. Previous efforts have tackled specific tasks, such as (Sadat et al., 2014), which developed a framework for translating Tunisian Arabic social media text into MSA. Mulki et al. (2018) investigated sentiment analysis techniques, while TunBERT (Haddad et al., 2023) adapted BERT for Tunisian dialect understanding. In the domain of hate speech detection, Haddad et al. (2019) introduced the first publicly available Tunisian Hate and Abusive Speech (T-HSAB) dataset, aiming to establish a benchmark for automatic detection of online Tunisian toxic content. Additionally, Mdhaffar et al. (2024) evaluated speech encoders for spoken Tunisian Arabic in SLU and ASR tasks, and Gugliotta et al. (2020) proposed a multi-task sequence prediction system for Arabizi annotation.

However, to the best of our knowledge, no prior work has systematically benchmarked LLMs for Tunisian Arabic in chat settings, which represent the ways in which everyday Tunisian speakers may hope to use these models. Our work fills this gap by providing the first comprehensive evaluation of LLMs on Tunisian Arabic, highlighting their strengths and limitations while paving the way for more inclusive multilingual NLP research.

3 Data Collection

To develop our benchmark, we first aimed to collect examples of instructions that Tunisian speakers would normally use. We collected Tunisian Arabic instructions through two methods: (1) crowdsourcing from native speakers and (2) translating an existing dataset.

3.1 Crowdsourced Instructions

To gather naturally occurring chatbot queries, we asked native Tunisian Arabic speakers to provide instructions they would typically use. For inclusivity, we recruited participants of different genders (32 female, 16 male), ages (15–48), and regions across Tunisia. While dialectal variation within Tunisia has minimal impact on written communication, this ensured a broad range of perspectives. This effort resulted in **417 unique instructions**.

3.2 Augmentation via Translation

To expand the dataset, one of the authors (a native speaker of Tunisian Arabic) translated **327 instructions** from the HuggingFace Helpful Instructions dataset [1] to Tunisian Arabic, bringing the total to **744 instructions**. Examples of these adaptations can be found in Appendix B. Each instruction was manually labeled with one or more topics, covering diverse areas (e.g., Culture, Society, Education). Full topic details are in Appendix A.

3.3 Gold-Standard Responses

To enhance the utility of our dataset, three of the authors, all native speakers of Tunisian Arabic, collaborated to write high-quality gold-standard responses for each of the 744 instructions. These responses are designed to be clear, natural, and culturally appropriate, providing a valuable reference point for researchers interested in fine-tuning models or conducting reference-based evaluations. While our primary analysis focuses on **reference-free evaluation**, we believe that making these gold responses available will facilitate a broader range of future research on Tunisian Arabic. Example entries are shown in Table 13, with English translations provided for readability.

4 Model Selection and Prompting

After constructing our dataset, we selected a range of LLMs and prompted them to follow the instructions.

4.1 Model Selection

To ensure a comprehensive evaluation of LLMs for generating responses in Tunisian Arabic, we selected models based on their widespread use and strong performance at the time of writing. The chosen models included GPT-4o, GPT-4o-mini, GPT-4, GPT-3.5 Turbo, LLaMA-3.3-70B (Llama Team, 2024), Aya-23-8B (Aryabumi et al., 2024), jais-13B-chat (Sengupta et al., 2023), SILMA-9B-Instruct-v1.0 (SilmaTeam, 2024), LLaMA-3.1-8B-Instruct (Llama Team, 2024), and Mixtral-8x7B-Instruct-v0.1 (MistralAI, 2024). Further details can be found in Appendix C.

4.2 Prompting Strategy

Each of the selected LLMs was prompted using our instructions dataset. The **temperature parameter was set to its default value**. To prescribe the use of Tunisian Arabic, we appended the following directive to each instruction:

أجب باللهجة التونسية في جميع ردودك،
مع استخدام المفردات والتعابير المحليةْ

This phrase explicitly instructed the models to generate responses in Tunisian Arabic, ensuring that their ability to follow dialectal constraints could be evaluated. Following the collection of responses for all instructions from all models, we proceeded to the evaluation phase, wherein responses were assessed using a pairwise comparison methodology. The details of this evaluation process are described in the subsequent sections.

[1] https://huggingface.co/datasets/HuggingFaceH4/helpful_instructions

5 Evaluation Methodology

Next, we explored the extent to which an LLM can evaluate these responses. To do this, we collected ratings from both human annotators and LLM models to systematically compare their judgments.

5.1 Evaluation Setup

Given our selection of 10 LLMs, we adopted a **pairwise evaluation strategy**, resulting in a total of 45 unique model comparisons (the number of ways to choose 2 models out of 10, C(10,2) = 45). For our initial evaluation model outputs, we used a subset of our responses such that each of the 744 instructions was assigned to one or two model pairs, ensuring that every LLM participated in an equal number of comparisons (20 times). This setup yielded a total of **900 evaluation instances**. To mitigate position bias (Zheng et al., 2023), we randomly shuffled the order of response pairs before presenting them for evaluation. This ensured that neither model had a systematic advantage based on response positioning. Each evaluation involved assessing two responses to the same prompt and determining the quality of the responses based on the evaluation criteria that will be introduced in the next subsection. We employed two types of judges: **Human annotators** and **LLMs-as-judges**.

5.2 Evaluation Criteria

We assessed the quality of model responses according to the following evaluation criteria:

- **Quality:** Which response is preferred overall, considering clarity, coherence, and engagement?
- **Correctness:** Which response provides more factually accurate information?
- **Relevance:** Which response better addresses the given prompt and remains on-topic?

The judge selected whether A or B is better, if they are tied, or if both responses are inadequate. Options: `[A is better / B is better / Tie / Both are bad]`

Additionally, each response was individually assessed for **Tunisian Arabic Usage** on a scale from 0 to 2. A score of 0 indicates that no Tunisian Arabic was used (e.g., MSA). A score of 1 signifies that some Tunisian Arabic was present, such as mixed dialects or partial Tunisian expressions. Finally, a score of 2 means that the response was fully in Tunisian Arabic. These evaluation criteria were chosen to address key issues observed during preliminary analysis. **Relevance** ensures that model responses stay on-topic, as we frequently encountered hallucinations where outputs were unrelated to the prompt. **Correctness** distinguishes between responses that are factually accurate and those that contain misinformation, helping us assess truthfulness separately from relevance. **Quality** captures fluency, coherence, and engagement, which influence user-perceived response quality and aligns closely with commonly used human judgments of LLM outputs (Chiang et al., 2024). More details on failure cases can be found in the **Error Analysis** section 6.3.

5.3 Human as a Judge

To evaluate model responses, three authors of this paper served as annotators, collectively labeling the 900 evaluation instances. Annotation was conducted in three rounds. In **Round 1**, each annotator independently labeled 66 instances, with 33 instances overlapping between each pair of annotators. We computed $Krippendorff's\ alpha$ to measure inter-annotator agreement, which was found to be low. This was largely due to borderline cases for the correctness evaluation criteria, where both annotators often recognized flaws in the responses, but one selected "Tie" or "Both are bad" while the other slightly favored one LLM. Direct disagreements, where one annotator preferred one LLM and the other preferred the other, were extremely rare, only one instance, compared to 20 instances involving a tie or "Both are bad" versus a preferred LLM. These borderline disagreements explain the initially low correctness agreement. To improve consistency, we conducted an **adjudication phase**, where disagreements were discussed and resolved. Following adjudication, in **Round 2**, each judge independently labeled an additional set of 33 instances. We then computed Krippendorff's alpha again and found that agreement had improved to a sufficient level (Table 1).

Therefore, in **Round 3**, the annotators continued to annotate all of the remaining evaluation instances where each instance was labeled by one annotator.

5.4 LLM-as-a-Judge

We leveraged LLMs themselves to assess the responses in a self-evaluation framework. Following the LLM-as-a-Judge approach (Zheng et al., 2023), we prompted the same LLMs to evaluate model re-

Evaluation Criterion	Round 1	Round 2
Quality	0.335	0.509
Correctness	0.215	0.414
Relevance	0.240	0.507
Tunisian Usage (Avg)	0.309	0.493

Table 1: Krippendorff's α for each evaluation criterion at different annotation stages.

sponses using the same instruction-response pairs and evaluation criteria. **The temperature was set to zero** to ensure deterministic outputs. The evaluation prompt was adapted from the AlpacaEval benchmark (Li et al., 2023) and tailored to our criteria. The full prompt can be found in Appendix D. All LLMs successfully adhered to the evaluation format, except for **jais-13B-chat** (Sengupta et al., 2023) and **SILMA-9B-Instruct-v1.0** (SilmaTeam, 2024). jais-13B-chat failed to follow the structured response guidelines entirely, leading to its exclusion as a judge. SILMA-9B-Instruct-v1.0, on the other hand, correctly evaluated **Quality**, **Correctness**, and **Relevance** but ignored **Tunisian Usage**. As a result, we retained SILMA-9B-Instruct-v1.0's evaluations for the first three criteria while excluding it from Tunisian Usage assessment. We parsed the evaluations for each criterion from the respective LLM responses, ensuring consistency in the extracted judgments. This approach enabled us to compare human and LLM-based evaluations systematically.

6 Results and Analysis

Finally, we used the LLM with highest alignment with human ratings to automatically evaluate our full dataset, and analyzed the results.

6.1 Evaluating LLMs as Judges

To assess the reliability of LLMs as evaluators, we compared their annotations against human judges across all evaluation criteria in terms of both (1) Cohen's Kappa (Vieira et al., 2010) and (2) Spearman correlation (Zar, 2005) between the leaderboard rankings produced using each approach. For the first three comparison-based criteria, the leaderboard is based on head-to-head Elo ratings (Elo and Sloan, 1978) (details below), and for Tunisian usage, the average scores were used. The agreement levels vary significantly across models, with **GPT-4o** achieving the highest agreement in most categories as shown in Table 2.

While there was some disagreement between the models and human ratings at the individual instance level, our main goal was to automatically produce a reliable *ranking* of models in terms of our four criteria. Therefore, we used an Elo rating system (Elo and Sloan, 1978; Boubdir et al., 2023) with penalties for poor-quality evaluations (`Both are bad`). The Elo score update functions are defined as:

$$\begin{aligned} R'_A &= R_A + K(S_A - E_A) \\ R'_B &= R_B + K(S_B - E_B) \end{aligned} \qquad (1)$$

where:

- R_A and R_B are the current Elo ratings of models A and B.
- $E_A = \frac{1}{1+10^{(R_B - R_A)/400}}$ is the expected probability of model A winning.
- S_A is the actual outcome: 1 if A wins, 0.5 if a tie, and 0 if B wins.
- K is the scaling factor (set to 32).

To penalize unreliable evaluations, if both models were judged as poor-quality (`Both are bad` rating), a fixed penalty of P points was subtracted from their scores:

$$R'_A = R_A - P, \quad R'_B = R_B - P \qquad (2)$$

where the penalty factor, P was set to $\frac{K}{2} = 16$ to avoid updates that are significantly more dramatic than would be received for other outputs. The resulting Elo-based leaderboards for each judge, including human annotators, are reported in Table 3a for Quality, Table 3b for Tunisian Usage, Table 12a for Correctness, and Table 12b for Relevance. For the Tunisian usage criterion, instead of Elo rankings, we computed **the average score** for each model, as this evaluation was ordinal rather than comparative.

6.1.1 Ranking Correlation Across Judges

To assess the consistency of rankings produced by different judges, we computed the **Spearman correlation** between Elo rankings derived from each judge's evaluations. This correlation measures the monotonic relationship between rankings, indicating how similarly different judges ranked the models. We report the heatmap of Spearman correlation values in Figure 2.

LLM	Quality	Correctness	Relevance	Tunisian Usage	*Average*
GPT-4o	**0.382**	**0.323**	0.238	**0.311**	***0.313***
GPT-4	0.340	0.301	**0.263**	0.184	*0.272*
GPT-4o-mini	0.342	0.292	0.207	0.205	*0.262*
LLaMA3.3_70b	0.329	0.291	0.205	0.182	*0.251*
SILMA-9B	0.229	0.17	0.147	-	*0.182*
LLaMA3.1_8b	0.214	0.173	0.129	0.113	*0.157*
GPT-3.5	0.208	0.17	0.145	0.108	*0.158*
mixtral-8x7B	0.213	0.144	0.144	0.104	*0.151*
aya-23-8B	0.018	0.047	0.012	0.057	*0.034*

Table 2: Cohen's Kappa agreement scores between each LLM and human annotators across different evaluation criteria.

Rank	Human	GPT-4o	GPT-4o-mini	GPT-4	GPT-3.5	LLaMA3.370B	aya-23-8B	SILMA-9B	Llama-3.1-8B	mixtral-8x7B
1	GPT-4o	GPT-4o	GPT-4o-mini	GPT-4o-mini	aya-23-8B	GPT-4o	GPT-4o-mini	GPT-4o	aya-23-8B	aya-23-8B
2	GPT-4o-mini	GPT-4o-mini	GPT-4o	GPT-4o	GPT-4o-mini	GPT-4o-mini	Llama-3.1-8B	GPT-4o-mini	LLaMA3.370B	LLaMA3.370B
3	LLaMA3.370B	LLaMA3.370B	aya-23-8B	LLaMA3.370B	GPT-4o	LLaMA3.370B	GPT-4	aya-23-8B	GPT-4o	GPT-4o-mini
4	GPT-4	GPT-4	LLaMA3.370B	aya-23-8B	LLaMA3.370B	aya-23-8B	LLaMA3.370B	LLaMA3.370B	GPT-4o-mini	GPT-4o
5	GPT-3.5	aya-23-8B	GPT-4	GPT-4	mixtral-8x7B	GPT-4	GPT-3.5	GPT-4	Llama-3.1-8B	GPT-4
6	aya-23-8B	GPT-3.5	GPT-3.5	GPT-3.5	Llama-3.1-8B	GPT-3.5	jais-13B	Llama-3.1-8B	GPT-4	Llama-3.1-8B
7	SILMA-9B	Llama-3.1-8B	Llama-3.1-8B	mixtral-8x7B	GPT-3.5	Llama-3.1-8B	aya-23-8B	SILMA-9B	GPT-3.5	mixtral-8x7B
8	Llama-3.1-8B	jais-13B	SILMA-9B	Llama-3.1-8B	GPT-4	mixtral-8x7B	mixtral-8x7B	GPT-3.5	mixtral-8x7B	jais-13Bt
9	jais-13B	mixtral-8x7B	mixtral-8x7B	jais-13B	jais-13B	jais-13B	GPT-4o	mixtral-8x7B	jais-13B	GPT-3.5
10	mixtral-8x7B	SILMA-9B	jais-13B	SILMA-9B	SILMA-9B	SILMA-9B	SILMA-9B	jais-13B	SILMA-9B	SILMA-9B

(a) Leaderboards of LLM Quality Rankings Across Different Judges.

Rank	Human	GPT-4o	GPT-4o-mini	GPT-4	GPT-3.5	LLaMA3.370B	aya-23-8B	Llama-3.1-8B	mixtral-8x7B
1	gpt-4o	GPT-4o	GPT-4o	GPT-4o	GPT-4o	GPT-4o	GPT-4o-mini	GPT-4o	GPT-4o
2	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	aya-23-8B	GPT-4o-mini	GPT-4o	GPT-4o-mini	GPT-4o-mini
3	GPT-4	GPT-4	GPT-4	GPT-4	GPT-4o-mini	GPT-4	LLaMA3.370B	Llama-3.1-8B	Llama-3.1-8B
4	LLaMA3.370B	GPT-3.5	LLaMA3.370B	GPT-3.5	GPT-4	LLaMA3.370B	aya-23-8B	GPT-4	Mixtral-8x7b
5	GPT-3.5	LLaMA3.370B	aya-23-8B	LLaMA3.370B	LLaMA3.370B	GPT-3.5	Llama-3.1-8B	LLaMA3.370B	LLaMA3.370B
6	Silma9B	aya-23-8B	GPT-3.5	Silma9B	GPT-3.5	Llama-3.1-8B	GPT-4	aya-23-8B	aya-23-8B
7	aya-23-8B	Silma9B	Llama-3.1-8B	aya-23-8B	Mixtral-8x7b	aya-23-8B	Mixtral-8x7b	GPT-3.5	GPT-4
8	Llama-3.1-8B	Llama-3.1-8B	Silma9B	Llama-3.1-8B	Llama-3.1-8B	Silma9B	GPT-3.5	Mixtral8	GPT-3.5
9	Mixtral-8x7b	Mixtral-8x7b	Mixtral-8x7b	Mixtral-8x7b	Silma9B	Mixtral-8x7b	Silma9B	Silma9B	Silma9B
10	Jais-13B	Jais-13B	Jais-13B	Jais-13B	Jais-13B	Jais-13B	Jais-13B	Jais-13B	Jais-13B

(b) Leaderboards of LLM Tunisian Usage Rankings Across Different Judges.

Table 3: The tables below present Elo-based leaderboards showing LLM rankings for each judge, including human annotators. Subtable (a) reflects rankings based on overall quality, while subtable (b) reflects rankings based on Tunisian Arabic usage

The results reveal notable patterns across different evaluation criteria. For **quality**, rankings exhibit strong correlations among high-performing models such as **GPT-4**, **GPT-4o**, and **LLaMA-3-70B**, suggesting a consistent evaluation of their outputs, while **aya-23-8B** shows much weaker correlations with human rankings, indicating a divergence in perceived quality. A similar pattern emerges for **correctness**, where **GPT-4o** and **GPT-4** display the highest agreement, reinforcing their close performance in generating accurate responses, whereas **aya-23-8B** and **Mixtral-8x7B** show lower correlations, reflecting greater inconsistency in their rankings. For **relevance**, the correlation structure aligns closely with that of quality, as top-tier models maintain strong agreement, while **aya-23-8B** and **Mixtral-8x7B** exhibit weaker correlations. The most variability appears in Tunisian usage, where **GPT-4** and **GPT-4o** continue to demonstrate high agreement, yet lower-tier models show weaker correlations, likely due to differing judgments on their ability to produce responses that align with Tunisian dialect expectations. Overall, these results highlight strong agreement among top models while weaker-performing models show inconsistent rankings across judges, reinforcing the importance of evaluating multiple linguistic and cultural dimensions to capture variations in model performance.

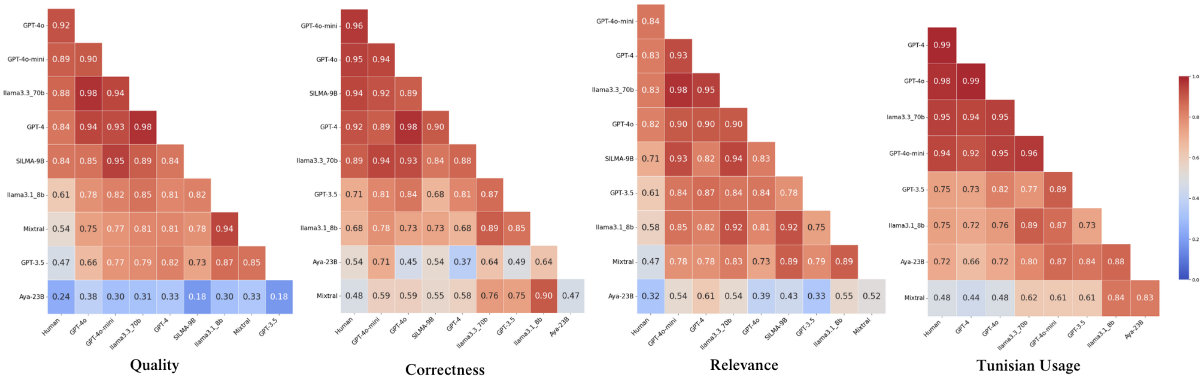

Figure 2: Spearman Correlation Heatmaps: The heatmaps illustrate the Spearman correlation between Elo rankings assigned by different judges across four evaluation criteria: Quality, Correctness, Relevance, and Tunisian Usage. Higher correlation values (darker red) indicate stronger agreement between rankings, while lower values (blue) reflect greater divergence.

Rank	Quality	Quality ELO	Correctness	Correctness ELO	Relevance	Relevance ELO	Tunisian Usage	Avg Tunisian Usage Score
1	GPT-4o	-483.37	GPT-4o	-4177.30	GPT-4o	-1686.74	GPT-4o	1.9476
2	GPT-4o-mini	-581.95	GPT-4o-mini	-4279.94	GPT-4o-mini	-1855.58	GPT-4o-mini	1.8734
3	aya-23-8B	-945.94	GPT-4	-4368.32	GPT-4	-2102.43	GPT-4	1.5974
4	LLaMA3.370B	-957.74	LLaMA3.370B	-4375.74	LLaMA3.370B	-2191.56	GPT-3.5	1.4101
5	GPT-4	-966.34	jais-13B	-4423.47	aya-23-8B	-2199.53	LLaMA3.370B	1.0028
6	GPT-3.5	-1154.47	aya-23-8B	-4427.32	GPT-3.5	-2297.40	aya-23-8B	0.8715
7	jais-13B	-1169.86	GPT-3.5	-4511.12	jais-13B	-2305.24	SILMA-9B	0.7648
8	Llama-3.1-8B	-1262.61	Llama-3.1-8B	-4595.23	mixtral-8x7B	-2406.23	mixtral-8x7B	0.5722
9	mixtral-8x7B	-1280.76	mixtral-8x7B	-4606.19	Llama-3.1-8B	-2427.64	Llama-3.1-8B	0.5593
10	SILMA-9B	-1284.96	SILMA-9B	-4691.35	SILMA-9B	-2455.65	jais-13B	0.0582

Table 4: Final leaderboard of LLM evaluations using GPT-4o as the primary judge. The table ranks the 10 evaluated models based on Quality, Correctness, and Relevance using Elo scores derived from 33,480 pairwise comparisons. Additionally, Tunisian Usage ranking and the average score are reported, showing how well each model incorporates Tunisian Arabic in responses.

6.2 Selection of GPT-4o as Primary Judge

Among all evaluated LLMs, **GPT-4o** exhibited the highest average agreement with human annotations across all evaluation criteria. It also demonstrated the strongest average ranking correlation with human judges across all evaluation criteria, achieving an **average Spearman correlation of 0.9175**, surpassing both **GPT-4-mini (0.9075)** and **GPT-4 (0.895)**. These results established GPT-4o as the most reliable automatic evaluator, leading to its selection as the **primary judge** for large-scale evaluation. To employ GPT-4o as a judge, we followed the same prompting strategy used for all previous judges. However, to produce the final leaderboard, we evaluated *all* possible **pairwise comparisons** among LLM responses, covering $744 * C(10, 2) = 33480$ unique comparisons. The evaluations were parsed from GPT-4o's responses using the same approach as before, ensuring consistency in the judgment process. Using these evaluations, we generated **leaderboards** for each criterion by computing **Elo scores**, maintaining alignment with prior ranking methods. The final leaderboards are in Table 4.

6.3 Discussion and Error Analysis

6.3.1 Human and LLM Agreement on Tunisian Usage

The human rankings of Tunisian usage scores exhibit a very high Spearman correlation with judge LLMs (Table 5), with values of 0.99 for GPT-4, 0.98 for GPT-4o, and 0.95 for LLaMA-3.3-70B. However, the Cohen's kappa agreement on these scores remains low, with values of 0.184, 0.3105, and 0.1815 for the same models, respectively. This suggests that while the ranking order of Tunisian proficiency is aligned between humans and LLMs, the absolute judgments diverge significantly. In other words, LLMs tend to rank models in the same order as humans but systematically overestimate their proficiency in the Tunisian dialect. Table 6

	Human	GPT-4o Judge	*Diff.*
GPT-4o	1.80	1.95	*+0.15*
GPT-4o-mini	1.68	1.87	*+0.20*
GPT-4	1.24	1.60	*+0.36*
GPT-3.5	1.13	1.41	*+0.28*
LLaMA-3.3-70B	1.15	1.00	*-0.14*
aya-23-8B	0.79	0.87	*+0.08*
SILMA-9B	0.96	0.77	*-0.20*
mixtral-8x7B	0.41	0.57	*+0.16*
LLaMA-3.1-8B	0.77	0.56	*-0.21*
Jais-13B Chat	0.36	0.06	*-0.31*

Table 5: Comparison of average Tunisian usage scores as rated by human annotators and GPT-4o as the final judge.

LLM Response	Human Tunisian Usage Score	GPT-4o Tunisian Usage Score
أوه، يا صاحبي! كلاه والله الحال متعك لا أعلن. إيا شنوا الهموم التي تكون فيها؟ كيفاش بش نقدر نساعدك؟	1	2
عندي شغل بالبايثون؟ وربى ما يحويكش! لي تو أقصر واحد، قولي نتوما نيهو! ياخي البهجة!	0	2
لعمل دوّارة ونقلبلها قبل الولد من اليسرا لليمين وبعد الولد من اليمين لليسرا.	0	2
العيد الصغير هو يوم عاشوراء، اللي هو ١٠ من شهر محرم.	2	0

Table 6: Examples of disagreement cases between human annotators and GPT-4o. In the first three rows, human annotators assigned low Tunisian usage scores (0 or 1), while GPT-4o assigned a score of 2. In the first row, the response contains a mix of MSA and other dialects, while the second and third responses fail to correctly interpret the prompt and are not in Tunisian Arabic. The last row illustrates the opposite situation, where human annotators assigned a score of 2, but GPT-4o judged the response as not being in Tunisian Arabic (score of 0).

presents some examples where human annotators gave low Tunisian usage scores (0 or 1), while GPT-4o consistently assigned 2.

6.3.2 Influence of Output Length on Model Preferences

To determine whether output length influenced model preference, we analyzed the percentage of instances in which judges preferred the longer response over the shorter one. Our hypothesis regarding length bias is informed by prior work which demonstrated that LLM-based evaluators tend to favor longer responses in English (Dubois et al., 2024). Given this documented bias, we wanted to investigate whether similar preferences exist when evaluating responses in Tunisian Arabic. The results, summarized in Table 7, indicate that human judges were the least biased towards longer responses (45%), followed by Aya23B (47%) and GPT-4o (63%). To further investigate this phenomenon, we examined the correlation between the final rankings of GPT-4o (our primary judge) and the rankings based on average response length for each LLM. The resulting Spearman correlation was very low (0.42 for quality, 0.15 for correctness, 0.31 for relevance, and 0.115 for Tunisian Usage), suggesting that response length alone cannot account for the final rankings.

6.3.3 Self-preference bias

We analyzed whether models preferred their own responses (i.e., self-preference bias (Koo et al., 2023; Deutsch et al., 2022)) by comparing the number of times each model selected its own output as the best against the number of times human judges preferred responses from that model. Table 8 illustrates the discrepancies between these numbers, revealing a consistent trend where models favored their own responses more frequently than humans did. This trend was particularly pronounced for mixtral-8x7B and LLaMA-3.1-8B, which exhibited the highest self-preference bias.

6.3.4 Content Issues and Annotation Flags

During dataset annotation, annotators were tasked with flagging content for various issues, including Hate Speech, Non-Arabic Responses, Inappropriate Content, Untruthful Information, and Personal Information Disclosure. A notable finding was that the mixtral-8x7B model was flagged 30 out of 180 times for producing responses that were not in

Judge	Preference for Longer Response (%)
Human	45.11
aya-23-8B	47.00
GPT-4o	63.78
SILMA-9B	67.55
GPT-4	68.55
GPT-4o-mini	71.00
LLaMA-3.3-70B	71.44
LLaMA-3.1-8B	74.22
GPT-3.5	74.55
mixtral-8x7B	75.33

Table 7: Percentage of times each judge preferred the longer response in pairwise comparisons.

LLM	Selection Count		
	Self	Human	Difference
mixtral-8x7B	78	29	49
LLaMA-3.1-8B	94	46	48
GPT-4o-mini	154	116	38
LLaMA-3.3-70B	119	84	35
GPT-4	108	84	24
GPT-4o	161	134	18
aya-23-8B	78	65	13
SILMA-9B	53	40	13
GPT-3.5	62	51	11

Table 8: Comparison of how often each LLM preferred its own response versus how often it was chosen by human evaluators.

Arabic—neither Modern Standard Arabic (MSA) nor Tunisian dialect. This highlights a significant limitation in mixtral-8x7B's language generation capabilities. The other types of flags were either used only once or not at all for each model.

7 Conclusion

In this paper, we have introduced a novel benchmark for evaluating the capabilities of large language models (LLMs) in understanding and responding in Tunisian Arabic, a dialect that is underrepresented in current language processing technologies. We developed a diverse dataset of Tunisian Arabic instructions and assessed ten known LLMs, revealing significant gaps in their ability to recognize and accurately respond in this dialect. Through comprehensive evaluations, GPT-4o emerged as the most reliable automated judge, aligning most closely with human assessments. Using an LLM-as-a-judge approach, we then produce a final ranking of ten models across four metrics. By releasing this benchmark and dataset, we aim to spur further research in developing dialect-specific NLP applications, enhancing language models for more effective communication in Tunisian Arabic.

8 Limitations

While our study provides valuable insights into the performance of LLMs on Tunisian Arabic, several limitations should be acknowledged. First, human evaluations may introduce bias, as annotators might have different views on response quality—some may prefer short and concise answers, while others favor detailed and structured responses. Second, our study is restricted to Tunisian Arabic written in Arabic script, whereas many speakers also use Latin script or a mix of both in digital communication. This limitation may impact the generalizability of our findings to other forms of Tunisian Arabic writing. Third, the dataset used for evaluation consists of 744 instructions, which, while diverse, does not capture the full range of possible queries and conversational contexts. Future work could expand on this by incorporating a larger and more representative set of instructions. Finally, our study evaluates existing LLMs without adaptation; future research could explore fine-tuning models on Tunisian Arabic data to improve their ability to generate more accurate and natural dialectal responses.

9 Ethical Considerations

Large language models (LLMs) trained on diverse datasets may reflect biases, stereotypes, or inaccuracies, particularly given the underrepresentation of Tunisian Arabic in major NLP datasets. This can lead to unfair or misleading outputs, favoring certain dialectal forms while marginalizing others. Additionally, models may generate misinformation, which is especially concerning when users rely on them for cultural, medical, or legal advice. Ensuring factual accuracy remains a challenge that future research should address. To uphold ethical standards in data collection, we obtained Institutional Review Board (IRB) exemption for gathering instructions from native Tunisian Arabic speakers, ensuring compliance with ethical guidelines and informed consent protocols. While our dataset does not include personally identifiable information, responsible data handling and transparency remain priorities for future work.

References

2025. Demographics of tunisia.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Salomey Osei, Shamsuddeen Hassan Muhammad, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stenetorp. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2024. Swan and arabicmteb: Dialect-aware, arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. *arXiv preprint arXiv:2411.01192*.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 339–352, Singapore. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. Ethnologue: Languages of the world - tunisian arabic (aeb). Accessed: 2025-02-14.

Arpad E Elo and Sam Sloan. 1978. The rating of chessplayers: Past and present.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

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

Elisa Gugliotta, Marco Dinarelli, and Olivier Kraif. 2020. Multi-task sequence prediction for tunisian arabizi multi-level annotation. *arXiv preprint arXiv:2011.05152*.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-hsab: A tunisian hate speech and abusive dataset. In *International conference on Arabic language processing*, pages 251–263. Springer.

Hatem Haddad, Ahmed Cheikh Rouhou, Abir Messaoudi, Abir Korched, Chayma Fourati, Amel Sellami, Moez Ben HajHmida, and Faten Ghriss. 2023. Tunbert: pretraining bert for tunisian dialect understanding. *SN Computer Science*, 4(2):194.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2024. Turning English-centric LLMs into polyglots: How much multilinguality is needed? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13097–13124, Miami, Florida, USA. Association for Computational Linguistics.

Md Tawkat Islam Khondaker, Numaan Naeem, Fatimah Khan, Abdelrahim Elmadany, and Muhammad Abdul-Mageed. 2024. Benchmarking llama-3 on arabic language generation tasks. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 283–297.

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

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

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

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*.

AI @ Meta Llama Team. 2024. The llama 3 herd of models.

Salima Mdhaffar, Haroun Elleuch, Fethi Bougares, and Yannick Estève. 2024. Performance analysis of speech encoders for low-resource slu and asr in tunisian dialect. *arXiv preprint arXiv:2407.04533*.

MistralAI. 2024. Mixtral-8x7b-instruct-v0.1.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Ismail Babaoğlu. 2018. Tunisian dialect sentiment analysis: a natural language processing-based approach. *Computación y Sistemas*, 22(4):1223–1232.

Fatiha Sadat, Fatma Mallek, Mohamed Mahdi Boudabous, Rahma Sellami, and Atefeh Farzindar. 2014. Collaboratively constructed linguistic resources for language variants and their exploitation in nlp application–the case of tunisian arabic and the social media. In *Proceedings of workshop on Lexical and grammatical resources for language processing*, pages 102–110.

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. *Preprint*, arXiv:2308.16149.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *arXiv preprint arXiv:2401.01854*.

SilmaTeam. 2024. Silma.

Susana M Vieira, Uzay Kaymak, and João MC Sousa. 2010. Cohen's kappa coefficient as a performance measure for feature selection. In *International conference on fuzzy systems*, pages 1–8. IEEE.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

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

A Instruction Topic Categories

Table 9 lists all topic categories used to annotate the Tunisian Arabic instructions, including the number of instructions per topic, an example instruction, and its English translation. Instructions may belong to multiple topics.

B Augmentation via Translation: Example translations and cultural adaptations.

To complement Section 3.2, Table 10 provides examples of how English instructions were translated and culturally adapted into Tunisian Arabic.

C Full List of LLMs Used

Table 11 provides a complete list of the LLMs used in our study, including their release dates and access methods to ensure reproducibility.

Model	Version / Release Date	Access Method
GPT-4o	Aug 2024	OpenAI API
GPT-4o-mini	Jul 2024	OpenAI API
GPT-4	Jun 2023	OpenAI API
GPT-3.5 Turbo	Jan 2024	OpenAI API
LLaMA-3.3-70B-Instruct	Apr 2024	Hugging Face
aya-23-8B	May 2024	Hugging Face
jais-13B-chat	Aug 2024	Hugging Face
SILMA-9B-Instruct-v1.0	Jan 2025	Hugging Face
LLaMA-3.1-8B-Instruct	Jul 2024	Hugging Face
Mixtral-8x7B-Instruct-v0.1	Dec 2023	Groq

Table 11: List of LLMs used in our study, including their specific versions and access methods for reproducibility. *OpenAI API* indicates that the model was accessed directly via the official API at https://platform.openai.com/. *Hugging Face* indicates that the model was downloaded from https://huggingface.co/models and run locally. *Groq* indicates that a hosted version of the model from the Groq API (https://console.groq.com/) was used.

D Full prompt used for LLM-as-a-Judge

As part of our self-evaluation framework, we used the **LLM-as-a-Judge** approach (Zheng et al., 2023) to assess model responses. To ensure consistency and fairness in evaluation, we prompted the same LLMs to compare responses using a standardized evaluation prompt adapted from the AlpacaEval benchmark (Li et al., 2023). This prompt guides the model in ranking responses based on predefined criteria while ensuring deterministic outputs by setting the temperature to zero.

Below, we present the full prompt structure, which consists of a **system message** that defines the model's role and a **user message** that provides detailed evaluation instructions.

System Message:

> "You are a helpful assistant that ranks models by the quality of their answers."

User Message:

> "Evaluate the following responses to the given prompt based on the criteria below. Provide your choices directly in the specified format without any additional explanations.

Prompt: {prompt}

Response A: {response_a}

Response B: {response_b}

### Pairwise Comparison

Evaluate the responses according to these criteria and choose one of the following for each:

- A is better
- B is better
- Tie
- Both are bad

1. **Overall Quality:** Which response would a human reader prefer overall? (Consider clarity, coherence, and engagement.)
2. **Correctness:** Which response is more factually accurate and provides correct information?
3. **Relevance:** Which response better addresses the prompt and stays focused on the topic?

### Tunisian Arabic Usage Evaluation

Assign a score to each response based on the following scale:

- 0: No Tunisian Arabic used.
- 1: Some Tunisian Arabic used (e.g., mixed dialects, partial Tunisian expressions).
- 2: Fully in Tunisian Arabic (100%).

### Response Format:

Pairwise Comparison:

- Overall Quality: [A is better / B is better / Tie / Both are bad]
- Correctness: [A is better / B is better / Tie / Both are bad]
- Relevance: [A is better / B is better / Tie / Both are bad]

Tunisian Arabic Usage: - Response A Score: [0 / 1 / 2]
- Response B Score: [0 / 1 / 2] "

E Leaderboards of LLM Correctness and Relevance Rankings Across Different Judges

We present the Elo-based leaderboards for each judge, including human annotators, across the two evaluation criteria: correctness in Table 12a and relevance in Table 12b. The results for quality and Tunisian usage are reported in Section 6.1.

F Sample Gold Reference Entries

Table 13 shows some example entries illustrating the addition of gold-standard responses to our dataset (English translations are manually added for readability):

Topic	Count	Example Instruction (Tunisian Arabic)	English Translation
Food & Cooking	64	نجم نحط راس حانوت في كسكسي باللحم البقري؟	Can I put "Ras el-hanout" in couscous with beef?
Religion & Beliefs	15	شنو اتجاه القبلة	What is the direction of the Qibla?
Health & Body	35	كيفاش انحي حب الشباب في اقرب وقت؟	How do I remove acne as quickly as possible?
Education & Studying	85	كيفاش نتميز في قرايتي؟	How can I excel in my studies?
Personal Development & Life Advice	84	كيفاش نظم وقتي؟	How can I organize my time?
Social & Romantic Relationships	65	كيفاش نكتب فقرة على العلاقات بين الصحاب؟	How do I write a paragraph about friendships?
Society & Current Affairs	92	علاش الناس تبدلت وماعادش تضحك؟	Why have people changed and stopped smiling?
Legal & Administrative	15	اذا كان شدوني و في جيبي ١غرام زطلة، نجم ندخل للحبس؟	If I get caught with 1 gram of cannabis, can I go to jail?
Money & Finance	17	١٠٠ دينار تونسي قداش تجي بالدولار	How much is 100 Tunisian dinars in dollars?
Technology & Digital Life	66	كيفاش نشري عملة رقمية BITCOIN ؟	How do I buy digital currency like Bitcoin?
Languages & Translation	65	تنجم تعرف الفرق بين اللغة التونسية والجزايرية؟	Can you tell the difference between Tunisian and Algerian Arabic?
Culture & Entertainment	105	قداش فما من حلقة في شوفلي حل؟	How many episodes are there in "Shoufli Hal"?
Travel & Immigration	14	شنو الفرق بين الدكتوراه في تونس والدكتوراه في امريكا؟	What is the difference between a PhD in Tunisia and a PhD in America?
Greetings & Small Talk	23	شعامل لباس عليك	How's everything with you?
Other	54	اللون البرتقالي اسمو برتقالي خطر اشبه للون البرتقال والا البرتقال اسمو برتقالي خطر لونو برتقالي؟	Is orange called orange because of its color or is it the other way around?

Table 9: Tunisian Arabic instruction topic categories, number of instructions per topic, and example instructions with English translations.

Translation Approach	Original English Instruction	Tunisian Arabic Translation
Standard Instruction Translation	Create a 3-turn conversation between a customer and a grocery store clerk - that is, 3 per person. Then tell me what they talked about.	اعملي حوار ب 3 ادوار بينات واحد يخدم في كاسا و حريف. كل واحد عندو 3 ادوار. قلي كل واحد شنوا قال

Code-Related Translation	Can you find and correct any logical errors in the following code snippet and output the corrected code? function countVowels(str) let vowels = [’a’, ’e’, ’i’, ’o’, ’u’]; let count = 0; for (let i = 0; i < str.length; i++) if (vowels.includes(str[i])) count++; return count;	تنجم تلوجلي الحاجات الغالطة في الكود هذا؟ اعطيني الكود الصحيح في الاخر
		function countVowels(str) {
		let vowels = [’a’, ’e’, ’i’, ’o’, ’u’];
		let count = 0;
		for (let i = 0; i < str.length; i++) {
		if (vowels.includes(str[i])) {
		count++; } }
		return count;}

Proper Names Adaptation	I need you to write a resignation letter to my boss. My name: Anthony Company Name: AirTek Position: Data Analyst Boss Name: Albert Last Day: 2 weeks from today (today is 02/10/2023).	نحبك تكتبلي رسالة تقاعد للعرف متاعي. اسمي أنور الشركة متاعي: تونيسار. رتبة:محلل معلومات اسم العرف : محمد. اخر نهار : بعد جمعتين من اليوم (02/10/2023)
Geographical Adaptation	List some interesting things to do in Idaho.	شنوا حاجات باهيا نجمو نعملوها في توزر.

Table 10: Augmentation via Translation: Examples of English instructions translated and culturally adapted into Tunisian Arabic.

Rank	Human	GPT-4o	GPT-4o-mini	GPT-4	GPT-3.5	LLaMA3.370B	aya-23-8B	SILMA-9B	Llama-3.1-8B	mixtral-8x7B
1	GPT-4o	GPT-4o	GPT-4o-mini	GPT-4o	aya-23-8B	GPT-4o	GPT-4o-mini	GPT-4o	aya-23-8B	aya-23-8B
2	GPT-4o-mini	GPT-4o-mini	GPT-4o	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	Llama-3.1-8B	GPT-4	GPT-4o	GPT-4o
3	LLaMA3.370B	LLaMA3.370B	LLaMA3.370B	GPT-4	GPT-4o	LLaMA3.370B	LLaMA3.370B	GPT-4o-mini	LLaMA3.370B	LLaMA3.370B
4	GPT-4	GPT-4	GPT-4	LLaMA3.370B	LLaMA3.370B	aya-23-8B	GPT-4	LLaMA3.370B	GPT-4o-mini	GPT-4o-mini
5	GPT-3.5	aya-23-8B	aya-23-8B	aya-23-8B	GPT-4	GPT-4	aya-23-8B	aya-23-8B	GPT-4	mixtral-8x7B
6	aya-23-8B	GPT-3.5	GPT-3.5	GPT-3.5	jais-13B	GPT-3.5	GPT-4o	GPT-3.5	Llama-3.1-8B	GPT-4
7	SILMA-9B	jais-13B	Llama-3.1-8B	jais-13B	GPT-3.5	Llama-3.1-8B	GPT-3.5	SILMA-9B	GPT-3.5	Llama-3.1-8B
8	Llama-3.1-8B	SILMA-9B	SILMA-9B	SILMA-9B	Llama-3.1-8B	jais-13B	mixtral-8x7B	Llama-3.1-8B	mixtral-8x7B	jais-13B
9	jais-13B	Llama-3.1-8B	jais-13B	mixtral-8x7B	SILMA-9B	mixtral-8x7B	SILMA-9B	mixtral-8x7B	jais-13B	GPT-3.5
10	mixtral-8x7B	mixtral-8x7B	mixtral-8x7B	Llama-3.1-8B	mixtral-8x7B	SILMA-9B	jais-13B	jais-13B	SILMA-9B	SILMA-9B

(a) Leaderboards of LLM Correctness Rankings Across Different Judges.

Rank	Human	GPT-4o	GPT-4o-mini	GPT-4	GPT-3.5	LLaMA3.370B	aya-23-8B	SILMA-9B	Llama-3.1-8B	mixtral-8x7B
1	GPT-4o	GPT-4o	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini	GPT-4o	GPT-4o-mini	GPT-4o	GPT-4o	aya-23-8B
2	GPT-4o-mini	GPT-4o-mini	GPT-4o	GPT-4o	aya-23-8B	GPT-4o-mini	Llama-3.1-8B	aya-23-8B	LLaMA3.370B	GPT-4o
3	GPT-4	LLaMA3.370B	LLaMA3.370B	LLaMA3.370B	GPT-4o	LLaMA3.370B	LLaMA3.370B	GPT-4o-mini	aya-23-8B	GPT-4o-mini
4	LLaMA3.370B	GPT-4	aya-23-8B	GPT-4	LLaMA3.370B	aya-23-8B	GPT-4	LLaMA3.370B	GPT-4o-mini	LLaMA3.370B
5	GPT-3.5	aya-23-8B	GPT-4	aya-23-8B	GPT-3.5	GPT-4	mixtral-8x7B	GPT-4	Llama-3.1-8B	GPT-4
6	SILMA-9B	GPT-3.5	GPT-3.5	GPT-3.5	GPT-4	GPT-3.5	GPT-4o	Llama-3.1-8B	GPT-4	Llama-3.1-8B
7	aya-23-8B	jais-13B	Llama-3.1-8B	mixtral-8x7B	jais-13B	Llama-3.1-8B	aya-23-8B	GPT-3.5	GPT-3.5	mixtral-8x7B
8	mixtral-8x7B	Llama-3.1-8B	SILMA-9B	Llama-3.1-8B	mixtral-8x7B	mixtral-8x7B	GPT-3.5	SILMA-9B	mixtral-8x7B	jais-13B
9	Llama-3.1-8B	SILMA-9B	mixtral-8x7B	jais-13B	Llama-3.1-8B	SILMA-9B	jais-13B	mixtral-8x7B	jais-13B	GPT-3.5
10	jais-13B	mixtral-8x7B	jais-13B	SILMA-9B	SILMA-9B	jais-13B	SILMA-9B	jais-13B	SILMA-9B	SILMA-9B

(b) Leaderboards of LLM Relevance Rankings Across Different Judges.

Table 12: The tables below present Elo-based leaderboards showing LLM rankings for each judge, including human annotators. Subtable (a) reflects rankings based on correctness, while subtable (b) reflects rankings based on relevance.

Prompt (Tunisian Arabic)	Gold Standard Response (Tunisian Arabic)	English Translation of Prompt	English Translation of Response
كم من ساعة في اليوم؟	اليوم فيه 24 ساعة	How many hours are there in a day?	There are 24 hours in a day
وين موجودة تونس في خريطة العالم ؟	تونس موجودة في شمال إفريقيا، على الساحل متع البحر الأبيض المتوسط، تحديداً في المغرب العربي.	Where is Tunisia located on the world map?	Tunisia is in North Africa, on the Mediterranean coast, specifically in the Maghreb region.
قداش فما من حلقة في شوفلي حل؟	مسلسل شوفلي حل فيه 135 حلقة.	How many episodes are there in Shoufli Hal?	The TV series Shoufli Hal has 135 episodes.

Table 13: Gold-standard examples of Tunisian Arabic prompts and responses, with English translations for readability.