Are Dialects Better Prompters? A Case Study on Arabic Subjective Text Classification

Leila Moudjari[1] **and Farah Benamara**[1,2]
[1] IRIT, Université de Toulouse, ANITI, Toulouse, France
[2] IPAL, CNRS-NUS-A*STAR, Singapore

Abstract

This paper investigates the effect of dialectal prompting, variations in prompting script and model fine-tuning on subjective classification in Arabic dialects. To this end, we evaluate the performances of 12 widely used open source LLMs across four tasks and eight benchmark datasets. Our results reveal that specialized fine-tuned models with Arabic and Arabizi scripts dialectal prompts achieve the best results, which provides new findings on the fine-tuning of LLMs for low-resource languages.

1 Motivations

Large Language Models (LLMs) have gained significant attention for their ability to perform various NLP tasks, including text classification. Studies have demonstrated that these models achieve high accuracy on a wide range of classification tasks by leveraging prompt engineering and customized prompts (Brown et al., 2020; Sun et al., 2023; Bareiß et al., 2024). However, the results for subjective classification remain a challenge. For example, Zhang et al. (2024) have shown that LLMs perform poorly in classifying highly subjective content, such as irony and emotions, where contextual subtleties play a crucial role. Fine-tuning these models on large affective datasets has been shown to improve performance for tasks involving nuanced emotional understanding (Liu et al., 2024).

Additionally, studies indicate that LLMs tend to perform better when prompts are either in English or translated to English compared to when they are in the target language of the testset (Perak et al., 2024; Ondrejová and Šuppa, 2024; Nguyen et al., 2024; Cahyawijaya et al., 2024; Bareiß et al., 2024). This is particularly salient for low-resource languages such as dialectal Arabic and its regional variants, where similar declines were observed (Abdelali et al., 2024; Ahuja et al., 2023; Koto et al., 2024; Zhang et al., 2023; Abdelaziz et al., 2024). Improvements have, however, been observed with code-mixed prompts in few-shot settings, i.e., instructions in English and examples in MSA (Ahuja et al., 2023).

When it comes to subjective text classification in Arabic, three main challenges remain: (1) Task-specific transformer fine-tuned models consistently outperform LLMs when prompted in zero or few-shot settings (Zhang et al., 2023), (2) The results when using native language prompts in MSA vs. English were not conclusive (e.g., increase in subjectivity detection but decrease in sentiment analysis (Abdelali et al., 2024)), and (3) LLMs struggle with more nuanced subjective tasks like detecting sarcasm and emotions (Abdelali et al., 2024).

In this paper, we aim to address these challenges by investigating, for the first time as far as we know, the effectiveness of dialect-specific prompting in improving subjective text classification in Arabic. Our work makes contributions in both advancing prompt engineering methodologies and studying the effects of linguistic variation in prompt design, focusing on the following dimensions:

1. ***Dialectal prompting.*** We explore how LLMs respond to prompts written in Maghrebi (Algerian, Tunisian) and Levantine dialects, comparing performances with prompts in English and MSA.

2. ***Variations in prompting script.*** Code-switching may impact LLMs performances. We therefore vary prompts' scripts, exploring Latin, Arabic, and newly Arabizi (Yaghan, 2008), the informal Arabic chat alphabet in which words are written in their transliterated form using Latin characters and numerals (e.g., "*a3tini*" (*give me*)).

3. ***LLMs fine-tuning.*** We compare fine-tuned vs. non fine-tuned versions of 12 open source LLMs while varying their sizes and the training process

Findings of the Association for Computational Linguistics: ACL 2025, pages 17356–17371
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

in terms of training strategy (instruction vs. chat-tuned), language (Arabic vs. multilingual) and specialization (affective vs. general-purpose).

4. ***Robustness and generalization.*** We design experiments to assess how well fine-tuned models generalize across dialects, tasks, and scripts, shedding light on the impact of linguistic variation on model adaptability.

Our results provides new findings on fine-tuning LLMs for low-resource languages: (a) Fine-tuned models with dialect prompts written in Arabic script and Arabizi achieve the best, (b) Arabic-centric models demonstrate stronger efficiency, (c) LLMs achieve strong performances outperforming fine-tuned Bert-like smaller models, (d) Transfer learning of fine-tuned models across datasets/tasks as well as switching to a different dialect at inference time, significantly degrade performance. Our models are available to the research community.[1]

2 Methodology

2.1 Classification tasks

We explore 4 subjective tasks of various complexity, selecting publicly available datasets for each task where instances from various length are written either in MSA only, a specific dialect, or mixture of MSA and/or different dialects. We consider a wide range of Arabic dialects from different regions: North Africa (Algerian (Dz), Tunisian (Tn)), Levantine (Lv), and Gulf (Gl). Our datasets are as follows (see Appendix A for a detailed description, including the train/test split):

(1) *Sentiment analysis (SA):* OCA$_{MSA}$ (Rushdi-Saleh et al., 2011), Sem17$_{MSA+Mixed}$(Rosenthal et al., 2017), Twifil-sent$_{Dz}$ (Moudjari et al., 2020), and TSAC$_{Tn}$ (Medhaffar et al., 2017).

(2) *Emotion recognition (EM):* Twifil-emo$_{Dz}$ (Moudjari et al., 2020), Sem18$_{MSA+Mixed}$ (Mohammad et al., 2018).

(3) *Irony detection (ID):* IDAT$_{MSA+Mixed}$ (Ghanem et al., 2019).

(4) *Crisis management (CM):* Flood$_{Lv+Gl}$ (Alharbi and Lee, 2019).

In order to study the impact of dialect prompting in LLMs performances, we further analyzed the distribution of script in our datasets (cf. Appendix A.2). Arabic scripts are a majority on all datasets, except for OCA$_{MSA}$ where data is mixed with Latin scripts coming from English movie titles. Maghrebi dialect datasets contain instances exclusively written in Latin characters. In Table 2 from the same Appendix, we also analyzed the frequency of MSA samples relying on the ALDi score (Keleg et al., 2024) where $0 \leq ALDi < 0.1$ indicating that an instance is expected to be in MSA, otherwise it is likely in dialect. All ALDi scores are > 0.1 with Flood$_{Lv+Gl}$ having the smallest scores. After a manual check, we observed that these scores may not reflect the real percentage of dialect, in particular when the inputs are in Arabizi. We provide examples and discussion on this issue in the Limitation Section and Appendix A.2.

2.2 Experimental Settings

Models. See Appendix B.1:

- ***Transformers.*** We employ three models that have been pre-trained on MSA and social media content from various dialects: AraBERTv2 (Antoun et al., 2020), CAMeLBERT (Inoue et al., 2021), and DarijaBERT (Gaanoun et al., 2024).

- ***Non fine-tuned LLMs.*** We use 12 LLMs from three families: (a) *General-purpose multilingual LLMs*: SOLAR$_{10.7B}$ (Kim et al., 2023), Mistral$_{7b}$, Mistral-Instruct$_{7b}$(Jiang et al., 2023), LLaMA3$_{8b}$ (Dubey et al., 2024) and LLaMA3-Instruct$_{8b}$, (b) *Arabic-centric LLMs:* Jais-family$_{1p3b}$ (Sengupta et al., 2023), AceGPT$_{8b}$ (Huang et al., 2024) and AceGPT-chat$_{8b}$, and (c) *Specialized affective models*: EmoLlama$_{7b}$ (Liu et al., 2024), EmoLlama-chat$_{7b}$ and EmoLlama-chat$_{13b}$.

- ***Fine-tuned LLMs***. These 12 models have also been fine-tuned with QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2024).

Prompts. Given a task and to avoid bias, all LLMs share the same zero-shot prompts[2] and parameters (see Appendix C and B.2, resp.). We utilize role-playing by assigning roles to LLMs (Liu et al., 2023) and direct the models to generate a single response based on the provided labels only to limit hallucination. As LLMs may have different output formats, we consider a class correct if it exactly corresponds to task categories. When the model is unable to confidently assign a label,

[1] https://huggingface.co/DialectPrompting/Dialectal_LLMs

[2] Experiments with few-shot and in-context mixing (Shankar et al., 2024) were not conclusive.

we implemented two strategies: assigning a "neutral" label when it is included in the task annotation scheme, and resorting to a random false label, as suggested by Zhang et al. (2023), otherwise.

We also vary the languages and scripts:

- ***English Prompts:*** Instructions are written in English, with English labels, and examples are randomly sampled from the task dataset following (Abdelali et al., 2024; Koto et al., 2024).
- ***Arabic Prompts:*** English instructions are translated into MSA following (Ahuja et al., 2023; Zhang et al., 2023). We use ChatGPT then manually checked the translations by a native speaker, while examples are kept unchanged (i.e., same language as in the target task).
- ***Dialectal Prompts:*** Arabic instructions as well as labels are manually translated into a dialect $d = \{\text{Lv, Dz, Tn}\}$ written either in Arabic (*ar*) or Arabizi (*abz*) scripts. For Lv, we experiment with *ar* and *abz*, however and given that Dz and Tn are two Magherbi dialects, we explore Dz prompting with *abz* script to capture its distinct code-mixed nature while Tn with *ar*. This is also motivated by the very similar preliminary results obtained when prompting LLMs in Dz and Tn using Arabic script. In the following, the notation d_s indicates a prompt written in a dialect d using script $s \in \{ar,\ abz\}$.

Impact of fine-tuning and prompting variations. We design four experimental settings to evaluate the impact of dialectal and script variation on model generalization:

- `In-dataset`: All inputs are from a distribution present during training/testing. This is the standard within-dataset baseline.
- `In-dataset`$_{DA}$: A variant of the above where dialectal instances are isolated based on their ALDi scores, teasing apart MSA samples to better assess the effectiveness of our prompting strategies on dialect-specific data.
- `Cross-dataset transfer`: It focuses on evaluating task-level generalization across datasets with different dialects/scripts, without any exposure to the test distribution. Therefore, a model trained on one dataset is tested on another—within the same task but differing in dialect and/or script—without additional fine-tuning on the target data.
- `Fine-tuning transfer`: This setting examines the ability of LLMs to generalize when the prompting conditions at inference time differ from those seen during fine-tuning. It helps answer whether switching the prompting language (e.g., from Tunisian to MSA or Arabizi) degrades performance, and whether adaptation to one variant enhances or constrains generalization.

3 Results

To reduce randomness and enhance generalizability, all models are evaluated on the same test set and run three times with different seed of examples, if any; we report the averaged macro-F1 across runs.

`In-dataset` results. The main findings are as follows (cf. Table 1, together with Figure 3 as well as Tables 11 and 7 in Appendix D.1):

(a) *Dialect*: Prompting with the target dataset dialect is not always effective, likely due to varying sensitivities of LLMs to specific dialects, as acknowledged in a recent study (Mousi et al., 2025). For instance, LLaMA and its derivative as well as AceGPT excel in Gulf Arabic, while JAIS in Levantine. Additionally, mutlilingual models often confuse dialects (e.g., Gulf with Levantine or MSA) and tend to switch to MSA in uncertain cases.

(b) *Script*: We observe a correlation between the dominant script in the dataset and the prompt used (see Figure 1 in Appendix D.1). For instance, in Twifil-sent$_{Dz}$, where 50% of instances use a mix of scripts, the best results were achieved after fine-tuning using Algerian prompting. Similarly, in Sem18$_{MSA+Mixed}$, where 83% of instances are written in Latin script, the best results were achieved using Levantine with Arabizi. There are, however, two exceptions: (i) OCA$_{MSA}$, where 73% of the data is mixed between Arabic and Latin script. Best scores have been achieved with MSA which is consistent as sentiment tags are mostly derived from the review (and not from the movie titles). (ii) Twifil-emo$_{Dz}$, where English prompts LLaMA3$_{8b}$ were the best. This is likely due to these LLMs being trained on significantly larger English datasets compared to Arabic. After fine-tuning, Levantine with Arabizi achieved the best results (35.81), with MSA closely behind.

	Sem17$_{MSA+Mixed}$	**Twifil-sent**$_{Dz}$	**OCA**$_{MSA}$	**TSAC**$_{Tn}$	**Sem18**$_{MSA+Mixed}$	**Twifil-emo**$_{Dz}$	**IDAT**$_{MSA+Mixed}$	**Flood**$_{Lv+Gl}$
Best Baseline								
Reported SOTA$^{\Delta}$	61.00	79.52	90.60	63.35	49.53 (6 classes)	59.19 (6 classes)	84.40	–
Transformer	71.22	74.46	91.00	93.90	24.23	38.06	83.75	70.87
Non Fine-tuned LLMs								
Best Eng	42.18	65.25	87.92	**92.86**	**24.77**	**27.44**	**58.45**	**38.05**
Best MSA	**50.94**	68.28	**90.99**	67.30	**24.77**	21.97	39.54	31.63
Best Dialect	41.31	**75.92**	89.00	70.80	22.86	23.98	56.80	33.98
→ dialect$_{script}$	Dz$_{abz}$	Lv$_{ar}$	Lv$_{abz}$	Dz$_{abz}$	Dz$_{abz}$	Lv$_{ar}$	Lv$_{ar}$	Lv$_{abz}$
Fine-tuned LLMs								
Best Eng	80.36	**89.59**	98.00	97.98	30.66	27.78	84.54	67.04
Best MSA	81.39	86.77	**99.98**	98.89	30.20	33.38	87.34	**67.57**
Best Dialect	**85.23***	87.13	**99.98**	**99.34***	**32.41***	**35.81***	**94.89***	67.06*
→ dialect$_{script}$	Tn$_{ar}$	Dz$_{abz}$	Lv$_{ar}$	Dz$_{abz}$	Lv$_{abz}$	Lv$_{abz}$	Lv$_{ar}$	Lv$_{ar}$

Table 1: Best baseline and best LLMs results before and after fine-tuning in terms of macro F1-score. Best score per language/dialect are in bold while overall best performance for each dataset is highlighted in green. *: The difference between dialect vs. MSA prompting is statistically significant at $p < 0.01$. $^{\Delta}$: Reported state of the art results using different experimental settings (see Section 4 for a discussion).

(c) *Fine-tuning (FT) vs. non fine-tuning (N-FT)*: FT LLMs outperform transformers in all tasks, except CM and ED in Twifil-emo$_{Dz}$ with a small decrease of around 3%. N-FT results are in-line with state of the art findings where English prompts performed the best (Koto et al., 2024). FT results on the other hand, show consistent improvement, with dialectal prompting yielding the best performance. Also, dialects with shared linguistic features, such as Algerian and Tunisian, led to comparable outcomes across models, indicating the models' ability to leverage these similarities.

(d) *Models size, language and specialization.* We observe that specialized LLMs fine-tuned on affective data (i.e. EmoLlama family) do not exhibit specific improvement over general-purpose LLMs. When comparing pre-training strategies, instruction-tuned models, particularly Llama3-inst$_{8b}$ outperformed other non Arabic-centric models across nearly all datasets, with Emollama-chat$_{13b}$ closely following in two datasets. When it comes to Arabic vs. multilingual LLMs, language-specific AceGPT achieves best results in almost all datasets beating JAIS, the other Arabic LLM. Even without fine-tuning, prompting in MSA exhibits better performances than English, while dialects are the more productive prompts after fine-tuning.

`In-dataset`$_{DA}$ results. Table 8 in Appendix D.2 shows that when a dataset contains a high percentage of MSA, as seen in TSAC$_{Tn}$ (ALDi > 0.1), performance drops significantly with a lost 62.94% in F-score, while Flood$_{Lv+Gl}$ experienced a 19.75% decline. Conversely, when the MSA proportion is lower, the impact of ALDi-based filtering is minimal, and in some cases, even beneficial. For instance, Twifil-sent$_{Dz}$ saw a slight improvement of 0.2% after removing texts with ALDi=0. Additionally, the presence of dialectal texts can also enhance classification performance on MSA instances, as observed in most cases, whereas for OCA$_{MSA}$, the inclusion of dialect samples did not impact the results.

`Cross-dataset` results. We measured transfer learning of our best models across SA and ED tasks (the two others containing one dataset each). For SA (see Table 9 Appendix D.2), models perform best when prompted in the same script. For instance, in Sem17$_{MSA+Mixed}$, the top-performing model used Tunisian, while the best transfer learning results were achieved with Levantine. Levantine prompts generally yield stronger results, as Arabic-centric models tend to perform better on non-Maghrebi Arabic.

For ED (see Table 10 in Appendix D.2), we observe the same tendency. For example, in Sem18$_{MSA+Mixed}$, the best model was prompted in Levantine Arabizi, while the strongest transfer learning occurred with Algerian Arabizi. Algerian and Tunisian dialects exhibit strong mutual transferability. In TwiFil, the best-performing Algerian model showed optimal transfer when prompted in Tunisian Arabic, and vice versa.

`Fine-tuning transfer` results. Models fine-tuned on a specific dialect exhibit varying levels of robustness when tested on different dialects (see yellow cells in Tables 9 and 10 in Appendix D.2). In particular, models initially trained on larger data from MSA and Levantine/Gulf exhibit greater adaptability when fine-tuned on other regionally specific dialects (e.g., Tunisian, Algerian

Arabizi), see for e.g., the Sem17 (Tn$_{ar}$) model that achieved the highest performance in SA, followed by the Twifil-sent (Dz$_{abz}$) model —both built on AceGPT$_{8b}$.

A similar trend is observed in ED, demonstrating superior generalization, therefore maintaining a competitive score comparable to the best in-dataset fine-tuning.

4 Discussions

Comparison with state of the art. We acknowledge that while dialectal prompting improves performance over MSA prompting, the magnitude of these improvements varies across dialects. Overall, the improvements range from 1% (Twifil-sent$_{Dz}$) to 7% (IDAT$_{MSA+Mixed}$) as shown in the first line of Table 1. Our results beat SOTA models for Sem17$_{MSA+Mixed}$, Twifil-sent$_{Dz}$, OCA$_{MSA}$, TSAC$_{Tn}$ and IDAT$_{MSA+Mixed}$.

For the other datasets, it is important to mention that direct comparison with reported results is unfair as we are using different experimental settings due to differences in the main task (e.g., Sem18 emotion shared task is a multi-label classification and not a multiclass classification), number of classes (e.g., the latest reported results (Moudjari et al., 2021) on Twifil-emo$_{Dz}$ (resp. Sem18$_{MSA+Mixed}$) multiclass task use a set of 6 emotion classes and not the whole set composed of 10 (resp. 11) classes as we do here), or evaluation setting (e.g., in the Flood corpus, the train/test split is done by crisis events vs. random in our case). To allow a better comparison, we therefore designed similar baseline transformer models that we fine-tuned on each dataset following the same train/test split ratio across datasets.

Additionally, we conducted statistical significance testing to assess the impact of dialectal prompting compared to MSA prompting. Results reveal that dialect-specific prompting significantly outperforms MSA prompting (at $p < 0.05$) on six out of eight datasets, providing strong empirical support for incorporating dialectal variation in prompt design.

Mixed dialectal data. Our findings reinforce the advantage of using dialectal Arabic for prompting, especially when fine-tuned to align with the dataset's dominant script. We have however to consider cases where datasets are composed of mixed dialectal data. In this case, selecting a single dominant dialect for prompting can be challenging. However, most dialects share a substantial overlap with MSA, which serves as a linguistic backbone for many Arabic dialects.

Maghrebi dialects. They are among the least represented, as they incorporate more Amazigh and French words. Despite this, they are quite similar which benefits fine-tuning. For example, in TSAC$_{Tn}$, using Algerian prompts yielded the best results. Similarly, in Sem17$_{MSA+Mixed}$, MSA performed best before fine-tuning, but post-fine-tuning, Tunisian showed greater consistency, followed by Algerian, especially using AceGPT. This suggests that fine-tuning amplifies dialect-specific nuances, enabling the model to better adapt to unique characteristics of dialectal data. Notably, Maghrebi data within the dataset, which might have been misclassified when using MSA. This is illustrated in the following example from Twifil-emo$_{Dz}$:

ربي يجيب الخير اللهم أسقينا غيثا مغيثا نافعا غير ضار عاجلا غير اجلا
(*May God bring goodness. O Allah, grant us abundant, life-giving, beneficial, and non-harmful rain—soon, not delayed.*),
where the model prompted in ENG/MSA predicted "Sadness", possibly influenced by the expressions of supplication. However, the gold label is "Trust", reflecting a hopeful, faith-driven tone common in religious invocations. When prompted in dialect, the model correctly predicted "Trust", indicating that dialectal input facilitated better alignment with the culturally embedded emotional cues of the text.

5 Conclusion

This study examined the ability of 12 open source LLMs to handle Arabic dialects in both Arabic and Arabizi scripts to improve subjective text classification. We highlighted the impact of Arabic-centric LLMs fine-tuning and prompt design on models performances, showing that dialectal prompts generally outperformed MSA and English prompts after fine-tuning. Transfer learning, on the other hand, demonstrated limited effectiveness. While models pre-trained on MSA and Levantine/Gulf showed some ability to adapt to other regionally specific varieties (e.g, Maghrebi), their performances remained significantly lower compared to in-dataset configuration.

We believe our study will open the door to future directions in prompting LLMs for low-resource languages, beyond English or translated prompts.

Acknowledgment

We would like to gratefully thank the anonymous reviewers and the ACs for providing useful suggestions for improving the paper. This work has been supported by the French government as part of France 2030. Funded by the European Union - Next Generation EU as part of the France Relance. It has also been partially supported by DesCartes: the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program. Our work has also benefited from the AI Interdisciplinary Institute ANITI under the Grant agreement n°ANR-23-IACL-0002.

Limitations

This work advances the development of more inclusive NLP models by addressing the challenges posed by under-represented Arabic dialects. We however believe that its implications extend beyond a single language. Indeed, the challenges we addressed (dialectal variation, script differences, and cross-linguistic adaptability) are prevalent in other morphologically rich and diglossic languages, such as Hindi-Urdu. Future research will focus on evaluating language-centric LLMs, expanding the scope to include a broader range of dialects, beyond Arabic. This will provide deeper insights into the impact of script variations and code-switching on LLM performance, further enhancing their adaptability.

We utilized various open-source large language models in our experiments. It is important to acknowledge that these LLMs can exhibit biases and may encounter issues concerning token limit. Therefore, a critical approach should be adopted when interpreting the experimental outcomes.

Given the very good scores achieved by OCA and TSAC datasets, we conducted a contamination check using the Contamination Database.[3] Even if we do not find any evidence of overlap with our evaluation datasets, potential data contamination remains and should be acknowledged. Indeed, limited transparency in pre-training datasets and the lack of studies on Arabic dataset contamination leave room for uncertainty which encourages further research in Arabic datasets contamination.

A final limitation is related to the blurred distinction between MSA and dialects and across dialects, as MSA linguistic expressions often appear in daily speech. Our study relied on publicly available datasets that do not explicitly separate MSA, dialects, or non-Arabic samples, which may have caused some misrepresentations. Moreover, the ALDi metric proved unreliable for highly dialectal texts. Developing better metrics for Arabic dialects identification is a timely direction to explore in the future.

Ethics Statement

The data used for conducting the experiments are composed of texts taken from datasets publicly available to the research community.

Very few instances from the datasets we consider (in particular those for emotion and irony detection) may contain offensive or abusive language. In these cases (less than 0.5%), and to comply with the policies, LLMs generally refrain from generating any response prioritizing safe and respectful interactions, therefore the neutral class is assigned by default.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, et al. 2024. Larabench: Benchmarking arabic ai with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520.

AhmedElmogtaba Abdelmoniem Ali Abdelaziz, Ashraf Hatim Elneima, and Kareem Darwish. 2024. LLM-based MT data creation: Dialectal to MSA translation shared task. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 112–116, Torino, Italia. ELRA and ICCL.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Alaa Alharbi and Mark Lee. 2019. Crisis detection from arabic tweets. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pages 72–79.

[3] https://huggingface.co/spaces/CONDA-Workshop/Data-Contamination-Database

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for nli-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1318–1326, New York, NY, USA. Association for Computing Machinery.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2024. Darijabert: a step forward in nlp for the written moroccan dialect. *International Journal of Data Science and Analytics*, pages 1–13.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. 2019. Idat@fire2019: Overview of the track on irony detection in arabic tweets. In *Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15*, pages 10–13.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023a. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

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

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Amr Keleg, Walid Magdy, and Sharon Goldwater. 2024. Estimating the level of dialectness predicts interannotator agreement in multi-dialect Arabic datasets. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–789, Bangkok, Thailand. Association for Computational Linguistics.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. Solar 10.7b: Scaling large language models with simple yet effective depth upscaling. *Preprint*, arXiv:2312.15166.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Al-mubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5622–5640, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Pingsheng Liu, Zhengjie Huang, Xiechi Zhang, Linlin Wang, Gerard de Melo, Xin Lin, Liang Pang, and Liang He. 2023. A disentangled-attention based framework with persona-aware prompt learning for dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13255–13263.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. KDD '24, page 5487–5496, New York, NY, USA. Association for Computing Machinery.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5139, Bangkok, Thailand. Association for Computational Linguistics.

Salima Medhaffar, Fethi Bougares, Yannick Estève, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the third Arabic natural language processing workshop*, pages 55–61.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Leila Moudjari, Karima Akli-Astouati, and Farah Benamara. 2020. An Algerian corpus and an annotation platform for opinion and emotion analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1202–1210, Marseille, France. European Language Resources Association.

Leila Moudjari, Farah Benamara, and Karima Akli-Astouati. 2021. Multi-level embeddings for processing arabic social media contents. *Computer Speech & Language*, 70:101240.

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

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.

Viktória Ondrejová and Marek Šuppa. 2024. Can LLMs handle low-resource dialects? a case study on translation and common sense reasoning in šariš. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 130–139, Mexico City, Mexico. Association for Computational Linguistics.

Benedikt Perak, Slobodan Beliga, and Ana Meštrović. 2024. Incorporating dialect understanding into LLM using RAG and prompt engineering techniques for causal commonsense reasoning. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 220–229, Mexico City, Mexico. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Mohammed Rushdi-Saleh, M Teresa Martín-Valdivia, L Alfonso Ureña-López, and José M Perea-Ortega. 2011. Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, 62(10):2045–2054.

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.

Bhavani Shankar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. In-context mixing (ICM): Code-mixed prompts for multilingual LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4162–4176, Bangkok, Thailand. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Mohammad Ali Yaghan. 2008. Arabizi: A contemporary style of arabic slang. *Design issues*, 24:39–52.

Chiyu Zhang, Khai Doan, Qisheng Liao, and Muhammad Abdul-Mageed. 2023. The skipped beat: A study of sociopragmatic understanding in LLMs for 64 languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2630–2662, Singapore. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

A Subjective Datasets

A.1 Detailed Description

We detail below the datasets used in this study.

OCA$_{MSA}$ (Rushdi-Saleh et al., 2011): An MSA movie reviews dataset containing 500 reviews split into 250 positive and 250 negative instances collected from different Arabic web pages and blogs and automatically annotated using author's rating.

Sem17$_{MSA+Mixed}$ (Rosenthal et al., 2017): The dataset was used in the SemEval2017 Task 4, subtask A on identifying the overall sentiment of a tweet. The corpus is split into train and test sets containing respectively $3,355$ and $6,100$ tweets. The corpus does not target a given dialect.

Twifil-sent$_{Dz}$ and **Twifil-emo**$_{Dz}$(Moudjari et al., 2020): This is the largest Algerian dataset of about 11,000 tweets where each tweet has been manually annotated by crowd-sourcing for sentiment analysis (Twifil-sent$_{Dz}$) (positive, negative and neutral using a majority vote), as well as emotion (Twifil-emo$_{Dz}$) relying on a taxonomy of 10 emotions (*Happiness, Anger, Disgust, Fear, Sadness, Surprise, Trust, Love, Anticipation, and* Neutral).

TSAC$_{Tn}$ (Medhaffar et al., 2017): A Tunisian Sentiment Analysis Corpus of about 17k Facebook comments manually annotated into $8,215$ positive and $8,845$ negative.

Sem18$_{MSA+Mixed}$ (Mohammad et al., 2018): The SemEval-2018 Task 1 Affect in Tweets about predicting emotion intensity. We make use of the Emotion Classification (E-C) task dataset[4] which contains tweets collected in 2017 and manually annotated into 11 emotions categories (*Anger, Anticipation, Disgust, Fear, Joy, Love, Optimism, Pessimism, Sadness, Surprise*, and *Trust*).

IDAT$_{MSA+Mixed}$ (Ghanem et al., 2019): It contains tweets about different political issues and events related to the Middle East that was held during the years 2011 to 2018. Tweets are written with formal Arabic (MSA) and different Arabic language varieties: Egypt, Gulf, Levantine, and Maghrebi dialects. Each tweet has been manually annotated into Ironic or Not-Ironic.

Flood$_{Lv+Gl}$ (Alharbi and Lee, 2019), an Arabic-Twitter-Corpus-for-Flood-Detection The corpus includes 4,037 human-labeled Arabic Twitter messages for four high-risk flood events that occurred in 2018: Jordan floods, Kuwait floods, Qurayyat floods and Al-Lith floods. The selected events took place in different areas of the Arab world: Jordan, Kuwait, northern Saudi Arabia and western Saudi Arabia, respectively. The tweets were labeled based on relatedness to the crisis and information type (Other useful information, Affected individuals, Caution and advice, Donations and volunteering, Infrastructure and utilities damage, Not applicable, Sympathy and emotional support).

A.2 Dataset Statistics

Table 3 shows statistics about the datasets we consider here. Classification tasks are either binary or multiclass with a maximum of 11 classes for emotion detection. Four datasets target only dialects: Twifil-sent$_{Dz}$, TSAC$_{Tn}$, Twifil-emo$_{Dz}$, and Flood$_{Lv+Gl}$, while three are a mix between various dialects and MSA. We also consider one dataset in MSA only: OCA$_{MSA}$.

Figure 1 further analyzes the characteristics of our datasets in terms of the used script: Arabic, or Latin/Arabizi and mixed context in the training sets. We observe the presence of Latin and mixed script in all datasets, Flood$_{Lv+Gl}$ having the highest number of instances in Arabic script.

[4] https://huggingface.co/datasets/SemEvalWorkshop/sem_eval_2018_task_1

Dataset	Sem17$_{MSA+Mixed}$	Twifil-sent$_{Dz}$	OCA$_{MSA}$	TSAC$_{Tn}$	Sem18$_{MSA+Mixed}$	Twifil-emo$_{Dz}$	IDAT$_{MSA+Mixed}$	Flood$_{Lv+Gl}$
Train	21.20	19.28	32.40	27.67	51.24	19.15	38.17	16.66
Test	18.32	20.40	54.40	29.03	52.87	19.81	37.04	16.59

Table 2: ALDi scores in our datasets.

Task	Dataset	Source	Classes	Train	Test
SA	OCA$_{MSA}$	IMDB	2 (positive, negative)	400	100
	Sem17$_{MSA+Mixed}$	tweets	3 (positive, negative, neutral)	3,355	6,100
	Twifil-sent$_{Dz}$	tweets	3 (positive, negative, neutral)	7,144	2,069
	TSAC$_{Tn}$	facebook	2 (positive, negative)	13,665	1,981
ER	Sem18$_{MSA+Mixed}$	tweets	11(pessimism, fear, optimism, trust, anticipation, joy, anger, sadness, disgust, love, surprise)	4,034	1,009
	Twifil-emo$_{Dz}$	tweets	10 (happiness, anger, disgust, fear, sadness, Surprise, Trust, Love, Anticipation, and Neutral)	3,885	1,148
ID	IDAT$_{MSA+Mixed}$	tweets	2 (ironic, non-ironic)	4,024	1,006
CM	Flood$_{Lv+Gl}$	tweets	7 (Affected individuals, Caution and advice, Donations and volunteering, Infrastructure and utilities damage, Not applicable, Other useful information, Sympathy and emotional support)	3,223	806

Table 3: Datasets used in this study.

We also analyzed datasets in term of dialect/MSA distribution. Table 2 presents the average ALDi score for each train and test set. Manual inspection revealed inconsistencies, particularly with non-Arabic scripts (e.g., Arabizi), where ALDi fails to reflect dialectal content accurately. For instance:

- MSA example written in arabizi: "*hatha almakan jamil jidan*" (This place is very beautiful) → ALDi=0.116.
- Tunsian example in Arabizi: "*chbih yethaz wyet7at akeka 5o zina esghir hhhhhhh mas5fo*" (What's with Zina's little brother acting all hyped up? Hahaha, poor thing.") → ALDi=0.0522, which incorrectly suggests MSA.

These inconsistencies indicate that ALDi scores have to be improved: If MSA was very high in the datasets, then MSA prompts should have given the best results, which is not the case. However, the ALDi scores reinforce our hypothesis that fine-tuning elevates the model's ability to better adapt to the unique characteristics of dialectal data.

A.3 Train/Test Split

We used the canonical train/test splits provided by the original papers for most datasets to ensure consistency with prior research (see Table 3). To preserve split integrity, no additional shuffling or resampling was performed. For datasets without predefined splits (OCA and Flood), we applied train/test split with stratification and a fixed seed (125) to ensure balanced classes and reproducibility.

B Models

B.1 Models Description

We make use of the following open source models:

– **AraBERTv2** (Antoun et al., 2020): is a family of pre-trained transformer-based language models specifically designed for Arabic text based on BERT architecture and is optimized for various Arabic language processing tasks, such as text classification, named entity recognition, and sentiment analysis. We tested different version, and we report the best results with bert-base-arabert-v02 and bert-base-arabert-v02-twitter(fine-tuned on 60M Arabic tweets).

– **CAMeLBERT** (Inoue et al., 2021). Another collection of pre-trained BERT-based models tailored for Arabic NLP tasks. The results reported here are based on `bert-base-arabic-camelbert-mix`, which was pre-trained on a diverse corpus comprising Modern Standard Arabic (MSA), dialectal Arabic, and Classical Arabic.

– **DarijaBERT** (Gaanoun et al., 2024) is a BERT-based model designed for the Moroccan Arabic dialect, "Darija." It follows the BERT-base architecture without the Next Sentence Prediction

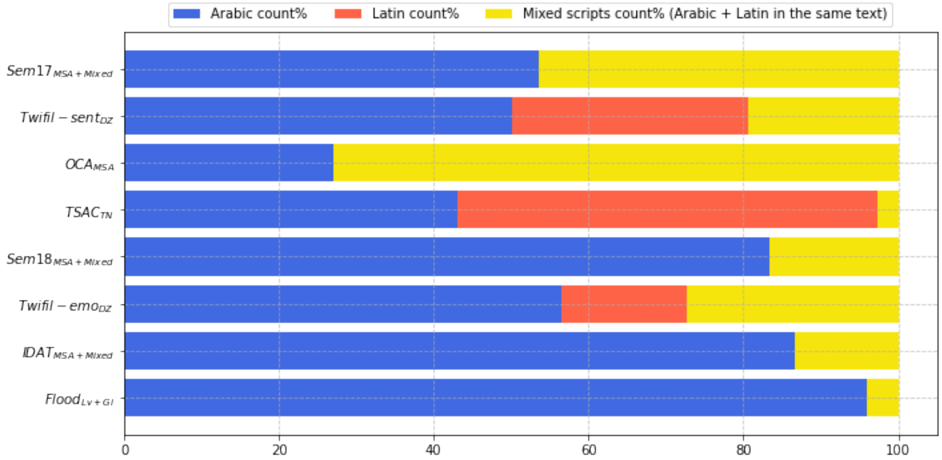

Figure 1: Distribution of Latin and Arabic scripts across datasets.

(NSP) objective. This Arabizi-specific version of DarijaBERT was trained on 4.6 million sequences of Darija written in Latin letters, using a dataset sourced from YouTube comments. Notably, the Arabizi model outperformed the other two variants from the same family.

– **AceGPT** (Huang et al., 2023b) is a collection of Arabic-centric LLMs specifically designed to advance Arabic language understanding and generation. The models are developed using three key strategies: (1) localized pre-training on extensive Arabic corpora, (2) localized fine-tuning with Arabic-specific questions paired with GPT-4-generated Arabic responses, and (3) reinforcement learning with AI feedback (RLAIF) to enhance performance. In this work, we utilize **AceGPT**$_{8b}$ and **AceGPT-chat**$_{8b}$, which are Arabic-focused LLMs built upon the LLaMA-3 architecture. These models are optimized to handle both MSA and Arabic dialects, making them well-suited for various Arabic natural language processing tasks, such as text classification, generation, and understanding.

– **Jais-Family** (Sengupta et al., 2023). It consists of a series of bilingual English-Arabic LLMs developed to excel in Arabic while maintaining strong English capabilities. These models are designed to handle various tasks, including text generation, comprehension, and summarization.

– **LLaMA3**$_{8b}$ and **LLaMA3-Instruct**$_{8b}$ (Touvron et al., 2023; AI@Meta, 2024) are part of the Meta LLaMA 3 family, a collection of pretrained and instruction-tuned generative LLMs. Pretrained on over 15 trillion tokens from publicly available sources. Instruction-tuned versions are optimized for dialogue tasks.

– **Mistral**$_{7b}$ and **Mistral-Instruct**$_{7b}$ (Jiang et al., 2023). **Mistral**$_{7b}$ is a 7-billion parameter dense transformer trained on a diverse corpus. Instruction-tuned variants of Mistral$_{7b}$ are further optimized for chat-based and instruction-following tasks.

– **SOLAR-inst**$_{10.7b}$ (Kim et al., 2023) The model was trained on publicly available datasets. It is instruction-tuned using supervised fine-tuning on a mixture of task-oriented prompts to perform instruction-following tasks such as question answering and dialogue.

– **EmoLlama**$_{7b}$, **EmoLlama-Instruct**$_{7b}$ and **EmoLlama-chat**$_{7b}$ are part of the EmoLLMs project (Liu et al., 2024), a collection of models designed for comprehensive affective analysis. This project focuses on the emotional understanding of text, offering models that can classify sentiments, detect emotional intensity, and predict sentiment strength. These variants are fine-tuned on the Meta LLaMA2 models with 234K data based on various classification and regression tasks.

B.2 Hyper-parameters

All **transformer models** have been trained using the Adam optimizer with an epsilon value of $1e-8$, a fixed batch size of 16 for training and 128 for validation, a learning rate of $2e-5$, and a sequence length of 128 tokens. This configuration is designed to balance efficient optimization with stable performance across various tasks, as the Adam optimizer's epsilon parameter helps maintain stability by preventing division by very small numbers in the moment estimates.

The number of epochs varies across different datasets (see Table 4): for AraBERTv2, OCA$_{MSA}$ and Sem18$_{MSA+Mixed}$, the models were trained for 7 epochs, while IDAT$_{MSA+Mixed}$, Sem17$_{MSA+Mixed}$, and Twifil-sent$_{Dz}$ were trained for 5 epochs each and TSAC$_{Tn}$ for 6. For CAMeL-BERT, Flood$_{Lv+Gl}$ was trained for 4 epochs, and Twifil-emo$_{Dz}$ for 7 epochs. This configuration ensures stable performance and efficient optimization across a wide range of tasks.

LLMs non fine-tuned hyper-parameters are shown in Table 5. To reduce the verbosity and keep the models focused on our scheme, we set the temperature to 0 to encourage deterministic responses. However, for the Llama models, which do not accept a temperature of 0, we adjusted it to 0.001.

For **fine-tuning**, we applied Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2024) to each of the previously mentioned open-source models (the hyper-parameters are summarized in Table 6). The choice of QLoRA is primarily due to its efficiency in fine-tuning large models on resource-constrained hardware while maintaining strong performance (we used a single GPU with 80GB of VRAM).

Each model was fine-tuned on the full training set of each dataset, and performance was assessed on a development set. The results are reported on the test set.

C Prompts

Figure 2 presents the prompting formats employed throughout our experiments, illustrating variations across English, MSA, and several Arabic dialects in both Arabic script and Arabizi. The first frame outlines the general prompt template structure applied consistently across instruction-tuning, inference, and evaluation stages.

We adopt a one-class-shot In-Context Learning prompting to control hallucination, using a concise prompting template with a single example of a single class and a 0 temperature, as extended prompts did not yield better results which was pointed out in other studies (Abdelaziz et al., 2024). This approach provided competitive performance on par with transformers across various tasks and datasets.

We provide the best performing prompts, namely fine-tuned one-class-shot ICL. We only illustrate them for sentiment analysis per language/dialect as they are the same for the emotion and irony detection, the only difference being the role and task

Prompting Template Used for Fine-Tuning
<|im_start|>
System: A) {prompt_task}
<|im_end|>
<|im_start|>
User: B) {user_text}
<|im_end|>
<|im_start|>
Assistant: C) {call_for_action}: {answer}
<|im_end|>

English Prompt
A) You are an Arabic sentiment text detector. Classify the text as *positive*, *negative*, or *neutral*. Return only the label.

B) Give only the label for this text: **[text]**

C) This text is mostly: **[label]**

MSA Prompt
A) أنت كاشف المشاعر في النصوص العربية. حدد التصنيف: محايد، إيجابي، أو سلبي.

B) أعط فقط الشعور بهذا النص: [text]

C) هذا النص في الغالب: [label]

Levantine (Arabic Script)
A) إنت مُصنّف مشاعر للنصوص، بتستلم نص عربي وبتعطيه تصنيف: محايد، إيجابي، أو سلبي.

B) أعطيني بس التصنيف لهالنص: [text]

C) هالنص علأغلب: [label]

Levantine (Arabizi Script)
A) Enta mosannif masha3er lilnosoos, btistalim nas 3arabi w bta3ti tasnif: neutral, positive aw negative.

B) A3tini bas eltasnif la halnas: **[text]**

C) Hal nass 3al aghlab: **[label]**

Algerian (Arabizi Script)
A) Ntaya détecteur ta3 les sentiments fi les texts ta3 l3rbiya w tgeneri l'étiquette: neutre, positive wla negative.

B) A3tini ghir le label ta3 had le texte: **[text]**

C) Had le texte ghalibane: **[label]**

Tunisian (Arabic Script)
A) إنت ديتيكتور تاع المشاعر في النصوص. وتعتي تصنيف: محايد، إيجابي، أو سلبي.

B) أعطيني غير التصنيف تاع هاد التكست: [text]

C) هذا التكست في الغالب: [label]

Figure 2: Prompt Design Across Dialects.

Dataset	**Sem17**$_{MSA+Mixed}$	**Twifil-sent**$_{Dz}$	**OCA**$_{MSA}$	**TSAC**$_{Tn}$	**Sem18**$_{MSA+Mixed}$	**Twifil-emo**$_{Dz}$	**IDAT**$_{MSA+Mixed}$	**Flood**$_{Lv+Gl}$
Epoch	5	5	7	6	7	7	5	4

Table 4: Epoch number for each dataset.

Parameter	**Value**
num_return_sequences	1
top_p	1
max_new_tokens	10
temperature	0
do_sample	False

Table 5: Parameters used for testing LLM models.

Parameter	**Value**
per_device_train_batch_size	8
per_device_eval_batch_size	8
optimizer	paged_adamw_32bit
learning rate	2e-4
logging_steps	1
gradient_accumulation_steps	1
epochs	1
fp16	True
Lora attention dimension (rank)	64
lora_alpha	16
lora_dropout	0.1
load_in_4bit	True
bnb_4bit_quant_type	nf4
torch_dtype	bfloat16

Table 6: Parameters used for fine-tunning the models with QLoRA.

description. For crisis management, we additionally enhance the prompts with the definition of each label.

D Detailed Results

D.1 Results per LLM and Tasks

Table 11 details all our results per LLMs for each dataset, langue/dialect and its associated scripts. A more synthetic view of the results is given in Table 7 and its corresponding plot in Figure 3 to visually compare performance drops before and after fine-tuning, across different models and datasets. Table 8 on the other hand gives a comparison of the best fine-tuned model across all datasets before and after removing MSA samples following their ALDi scores.

It is interesting to note that fine-tuned LLMs outperform transformers in all tasks, except crisis management and emotion detection in Twifil-emo$_{Dz}$ with a small decrease of around 3%. For 6 datasets out of 8, the best performance was obtained with AraBERTv2, CAMeLBERT being more productive for Flood$_{Lv+Gl}$, and Twifil-emo$_{Dz}$ datasets. DarijaBert was far behind the first two models. For example, for Twifil-emo$_{Dz}$ F1-score=13.74 (vs. 38.06) and Twifil-sent$_{Dz}$ F1-score=65.87 (vs. 74.46).

The results of OCA$_{MSA}$ and TSAC$_{Tn}$ are quite intriguing. We believe this can be attributed to the fact that both datasets (IMDB movie reviews and Facebook comments respectively) contain longer texts compared to the others, which are primarily sourced from Twitter. The increased text length likely provided the models with more context to infer the class. Additionally, both datasets focus on binary classification, which may have further simplified the task by reducing the complexity of decision boundaries.

Another interesting trend is observed: datasets composed primarily of texts written in Arabic script tend to achieve the best results when prompted in Arabizi (e.g., TSACTn, Twifil-emoDz). We hypothesize that this may be due to the fact that LLMs are predominantly trained on English data, and multilingual prompting helps bridge the gap and enhance performance (Huang et al., 2023a). This hypothesis is further supported by the observation that Lvabz contains English words, whereas DZ$_{abz}$ incorporates French, highlighting the impact of linguistic overlap in model adaptation.

Finally, small instruction and chat-tuned LLMs fine-tuned on specific tasks can outperform larger models, confirming recent findings (Zhang et al., 2023; Lu et al., 2024). This challenges the assumption that performance scales smoothly with model size, as earlier studies suggested (Black et al., 2022; Wei et al., 2022).

D.2 Transfer Learning Results

Tables 9 and 10 present the best results for both `Cross-dataset` transfer learning and `Fine-tuning transfer`, with the latter highlighted in yellow.

We observe that when a dataset performs best with a specific script, transfer learning from a model fine-tuned with a different script tends to yield lower results. For example, Twifil-sent$_{Dz}$ achieves its highest performance with AceGPT8b

	Sem17$_{MSA+Mixed}$		**Twifil-sent**$_{Dz}$		**OCA**$_{MSA}$		**TSAC**$_{Tn}$		**Sem18**$_{MSA+Mixed}$		**Twifil-emo**$_{Dz}$		**IDAT**$_{MSA+Mixed}$		**Flood**$_{Lv+Gl}$	
Fine-tuned ?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
Emollama$_{7b}$	Eng	Eng	Eng	Eng	Lv_{ar}	Dz_{abz}	MSA	MSA	Eng	MSA	Eng	Lv_{abz}	Lv_{abz}	Eng	Lv_{abz}	Tn_{ar}
Emollama-chat$_{7b}$	Eng	Dz_{abz}	Dz_{abz}	Tn_{ar}	Dz_{abz}	Dz_{abz}	Lv_{abz}	MSA	Eng	MSA	Eng	Lv_{abz}	Eng	Dz_{abz}	Lv_{abz}	Eng
Emollama-chat$_{13b}$	Eng	Eng	Eng	Lv_{abz}	Dz_{abz}	Tn_{ar}	Dz_{abz}	Lv_{ar}	MSA	Dz_{abz}	Eng	Lv_{abz}	Dz_{abz}	Lv_{ar}	Lv_{abz}	MSA*
Llama3$_{8b}$	Eng	Eng	Eng	Tn_{ar}	Lv_{abz}	MSA	Eng	Dz_{abz}	Lv_{abz}	Lv_{abz}	Eng	MSA	Lv_{ar}	MSA	Lv_{abz}	Eng
Llama3-inst$_{8b}$	MSA	Lv_{abz}	$Lv_{ar}{}^{*}$	Lv_{ar}	MSA*	MSA*	Dz_{abz}	Eng	Eng*	MSA	Eng*	$Lv_{abz}{}^{*}$	Lv_{abz}	Lv_{ar}	Lv_{abz}	MSA
Mistral$_{7B}$	Eng	Dz_{abz}	Dz_{abz}	Lv_{ar}	Lv_{ar}	MSA	Lv_{ar}	Dz_{abz}	MSA	Dz_{abz}	Eng	Lv_{abz}	Dz_{abz}	Tn_{ar}	Lv_{abz}	Lv_{ar}
Mistral-Inst$_{7b}$	Dz_{abz}	Dz_{abz}	Dz_{abz}	Eng	Dz_{abz}	MSA	$Dz_{abz}{}^{*}$	Tn_{ar}	Dz_{abz}	Eng	Eng	Dz_{abz}	Eng	Eng	Eng*	Lv_{abz}
SOLAR-inst$_{10.7b}$	Eng	Lv_{abz}	Eng	Lv_{ar}	Tn_{ar}	Dz_{abz}	Dz_{abz}	MSA	Eng	Lv_{abz}	Eng	Dz_{abz}	Eng*	Tn_{ar}	Eng	Lv_{ar}
AceGPT$_{8b}$	Eng	$Tn_{ar}{}^{*}$	Lv_{ar}	Eng*	Lv_{ar}	MSA*	Eng	Tn_{ar}	Lv_{ar}	$Lv_{abz}{}^{*}$	Eng	Eng	Lv_{ar}	$Lv_{ar}{}^{*}$	Eng	Dz_{abz}
AceGPT-chat$_{8b}$	MSA*	Tn_{ar}	Dz_{abz}	Dz_{abz}	Dz_{abz}	MSA*	Eng	$Dz_{abz}{}^{*}$	Eng	Tn_{ar}	MSA	MSA	Eng	Dz_{abz}	MSA	MSA
Jais-family$_{1p3b}$	Eng	MSA	Eng	$Lv_{abz}{}^{*}$	MSA	$Lv_{abz}{}^{*}$	$Lv_{abz}{}^{*}$	$Lv_{abz}{}^{*}$	MSA	$Lv_{abz}{}^{*}$	MSA	MSA	Lv_{ar}	Tn_{ar}	MSA	Eng

Table 7: Results per models showing outperforming language/script before and after fine-tuning in different colors: Grey for English, pink for MSA, green for Algerian Arabizi, orange for Tunisian Arabic, light and dark blue for Levantine Arabic and Arabizi respectively. Each entry represents the best performance across different prompting methods for each model. Stars in the cells indicate best configuration per dataset.

TASK	**SA**				**ER**		**ID**	**CM**
Models/Datasets	**Sem17**$_{MSA+Mixed}$	**Twifil-sent**$_{Dz}$	**OCA**$_{MSA}$	**TSAC**$_{Tn}$	**Sem18**$_{MSA+Mixed}$	**Twifil-emo**$_{Dz}$	**IDAT**$_{MSA+Mixed}$	**Flood**$_{Lv+Gl}$
Best macro-F1 scores (all instances)	85.23	89.59	99.98	99.34	32.41	35.81	94.89	67.57
% of instances with ALDi>0	37.01	8.22	47.00	1.72	4.26	9.41	19.78	18.61
Best macro-F1 scores on instances with ALDi>0	84.49	89.79	100.00	99.33	31.21	35.17	93.28	64.35
% of instances with ALDi>0.1	67.24	76.62	91	99.36	31.71	42.42	43.54	65.14
Best macro-F1 scores on instances with ALDi>0.1	80.68	89.17	100.00	36.4	23.59	32.31	90.74	44.83

Table 8: Comparison of the best fine-tuned model across all datasets before and after removing MSA samples following their ALDi scores set to two thresholds: $ALDi > 0$ and $ALDi > 0.1$ indicating that a sample is expected to be in dialect.

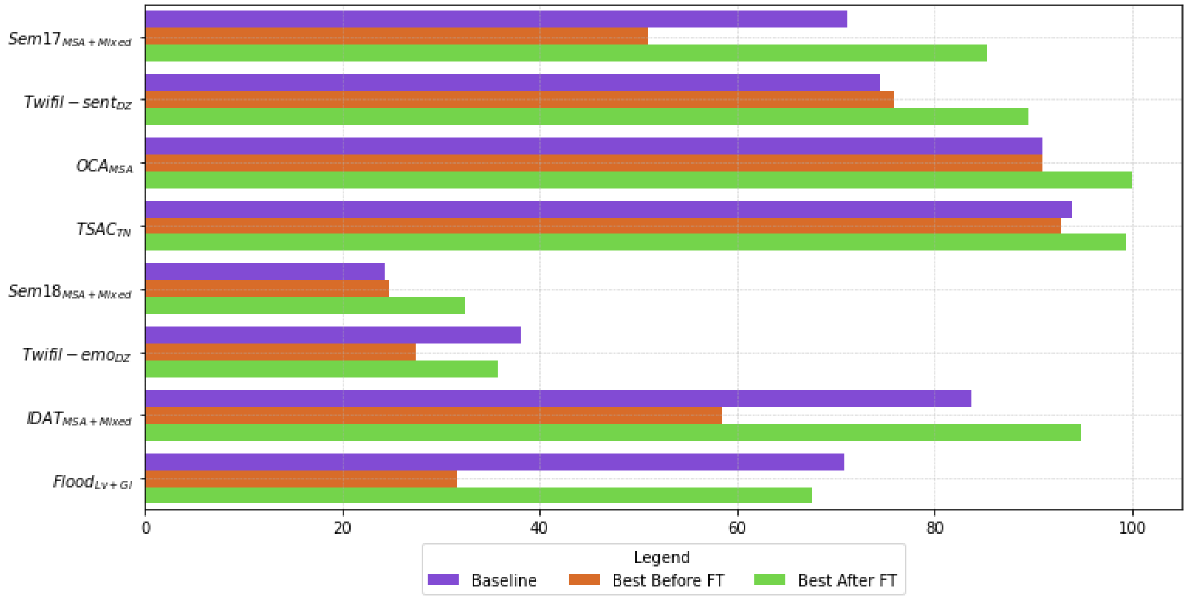

Figure 3: Performance comparison of LLMs across datasets: Baseline vs. Pre- and Post-Fine-Tuning.

Dataset	Best in-dataset FT	Best dialect transfer		
		Sem17	Twifil-Sent	TSAC
Sem17$_{MSA+Mixed}$	85.23 (Tn$_{ar}$)	43.58 (Lv$_{ar}$)	18.61 (MSA)	18.61 (Lv$_{ar}$)
Twifil-Sent$_{Dz}$	87.13 (Dz$_{abz}$)	35.13 (Lv$_{ar}$)	28.62 (Lv$_{ar}$)	17.82
TSAC$_{Tn}$	99.34 (Dz$_{abz}$)	54.48 (Dz$_{abz}$)	55.61 (Dz$_{abz}$)	45.93 (MSA)

Table 9: Sentiment Analysis transfer learning results. `Cross-dataset`: Best transfer learning vs. in-dataset performances in terms of macro F1-scores. `Fine-tuning transfer`: highlighted in yellow.

Dataset	Best in-dataset FT	Best dialect transfer	
		Sem18	Twifil-emo
Sem18$_{MSA+Mixed}$	32.41 (Lv$_{abz}$)	22.01 (Dz$_{abz}$)	6.56 (Tn$_{ar}$)
Twifil-emo$_{Dz}$	35.81 (Lv$_{abz}$)	35.13 (Lv$_{ar}$)	9.78 (Dz$_{abz}$)

Table 10: Emotion detection transfer learning results. `Cross-dataset`: Best transfer learning vs. in-dataset performances in terms of macro F1-scores. `Fine-tuning transfer`: highlighted in yellow.

using Dz$_{abz}$. However, in `Cross-dataset` transfer, the best results come from TSAC$_{Tn}$'s top-performing model, AceGPT-chat$8b$ with DZ$_{abz}$, showing a significant 37% performance gap compared to Sem17$_{MSA+Mixed}$'s best model, AceGPT$8b$ with Tn$_{ar}$.

Additionally, when these models encounter confusion, they tend to default to the base model language, typically MSA or Lv$_{abz}$, as observed in cases such as Sem17$_{MSA+Mixed}$.

A similar trend is observed in ED, where models fine-tuned on datasets with a predominant script struggle to generalize effectively when transferred to datasets using a different script.

Dataset		**Sem17**$_{MSA+Mixed}$	**Twifil-sent**$_{Dz}$	**OCA**$_{MSA}$	**TSAC**$_{Tn}$	**Sem18**$_{MSA+Mixed}$	**Twifil-emo**$_{Dz}$	**IDAT**$_{MSA+Mixed}$	**Flood**$_{Lv+Gl}$
Best Transformer/Baseline		71.22	74.46	91.00	93.90	24.23	38.06	83.75	70.87
Best open source LLM NFT		42.18	75.92	90.99	88.59	24.77	27.44	58.45	38.05
Best open source LLM NFT Lang		Eng	Lv$_{ar}$	MSA	Eng	Eng	Eng	Eng	Eng
Emollama$_{7b}$	Eng	**43.09**	**79.22**	76.72	62.51	20.30	21.02	**78.20**	60.81
	MSA	40.72	76.62	69.07	**94.92**	**26.75**	17.50	75.36	**62.51**
	Dz$_{abz}$	**43.01**	76.75	**81.97**	94.35	18.27	20.02	75.68	61.82
	Tn$_{ar}$	40.53	76.11	64.91	**94.72**	17.43	17.85	76.71	**62.96**
	Lv$_{abz}$	40.30	77.00	33.04	62.75	17.68	**23.99**	74.66	61.83
	Lv$_{ar}$	36.16	74.44	53.85	94.27	20.00	18.55	76.19	60.19
Emollama-chat$_{7b}$	Eng	43.80	76.45	81.93	94.39	17.63	18.06	76.67	**62.68**
	MSA	42.40	78.00	83.84	**94.88**	**20.98**	20.52	75.46	58.06
	Dz$_{abz}$	**44.31**	77.81	**85.99**	**94.88**	18.22	22.01	**78.42**	61.84
	Tn$_{ar}$	43.05	**78.77**	79.87	93.70	17.77	18.34	76.64	61.92
	Lv$_{abz}$	43.83	75.54	67.36	93.78	16.21	**24.32**	75.15	61.46
	Lv$_{ar}$	44.10	76.45	32.52	94.19	19.44	19.44	75.79	58.61
Emollama-chat$_{13b}$	Eng	**44.18**	77.06	66.37	94.06	18.81	20.56	76.18	67.04
	MSA	43.12	75.63	83.00	94.47	20.47	23.60	77.06	**67.57**
	Dz$_{abz}$	43.36	78.66	84.96	**95.08**	**21.07**	25.52	78.27	62.85
	Tn$_{ar}$	43.91	**79.65**	**87.00**	94.51	18.29	23.19	77.86	64.19
	Lv$_{abz}$	43.59	**79.86**	54.43	63.13	19.08	**28.21**	76.66	62.30
	Lv$_{ar}$	43.91	77.75	73.48	**95.41**	16.35	27.05	**79.22**	64.00
Llama3$_{8b}$	Eng	**43.50**	74.79	97.00	**94.43**	16.87	27.78	35.75	**62.10**
	MSA	34.98	78.82	**98.00**	92.76	14.75	**33.38**	**80.29**	58.14
	Dz$_{abz}$	40.58	**79.24**	**98.00**	**94.47**	14.02	27.85	78.81	43.62
	Tn$_{ar}$	43.39	**79.80**	34.21	93.09	17.07	31.79	75.54	61.32
	Lv$_{abz}$	42.96	77.84	68.75	93.00	**17.98**	32.60	78.76	**62.02**
	Lv$_{ar}$	43.34	**79.55**	88.95	93.46	16.01	28.51	79.74	61.35
Llama3-inst$_{8b}$	Eng	34.93	76.84	66.08	**94.92**	15.74	25.72	78.67	54.48
	MSA	43.30	65.89	**99.98**	92.15	**18.53**	30.29	70.82	**66.91**
	Dz$_{abz}$	43.42	68.30	99.00	91.60	11.98	24.23	79.07	61.69
	Tn$_{ar}$	43.70	60.08	99.00	93.86	17.77	23.92	78.72	16.52
	Lv$_{abz}$	**44.22**	73.94	91.99	94.06	16.11	**35.81**	77.17	65.96
	Lv$_{ar}$	42.31	**81.00**	**99.98**	94.51	16.80	22.27	**80.35**	58.91
Mistral$_{7b}$	Eng	33.21	53.81	87.68	89.80	16.18	24.02	**77.97**	56.76
	MSA	33.12	69.15	**94.00**	61.73	16.39	21.54	77.34	58.24
	Dz$_{abz}$	**38.84**	75.22	85.53	**93.13**	**19.60**	23.53	75.57	54.84
	Tn$_{ar}$	35.61	73.66	34.21	22.82	15.32	19.88	77.71	63.13
	Lv$_{abz}$	30.80	67.36	98.00	12.28	17.64	**27.73**	76.44	62.64
	Lv$_{ar}$	36.83	**76.03**	**93.94**	91.02	18.24	14.43	76.01	**66.57**
Mistral-inst$_{7b}$	Eng	21.76	**76.30**	82.22	90.63	**18.71**	19.10	**78.22**	56.06
	MSA	30.47	59.86	**95.43**	89.80	17.50	22.90	76.27	34.24
	Dz$_{abz}$	**35.54**	75.17	84.44	91.74	17.11	**26.73**	77.16	60.19
	Tn$_{ar}$	32.44	65.19	78.38	**93.33**	13.29	17.27	75.99	**61.31**
	Lv$_{abz}$	30.77	**76.29**	66.08	92.72	18.15	25.16	77.48	**61.50**
	Lv$_{ar}$	33.58	73.72	86.61	91.49	15.89	14.70	76.47	39.33
SOLAR-inst$_{10.7b}$	Eng	38.94	75.74	83.77	89.95	18.50	25.81	78.42	60.82
	MSA	33.95	73.97	92.98	**93.45**	18.03	22.42	78.44	65.69
	Dz$_{abz}$	37.90	78.94	**97.00**	91.34	17.68	**27.76**	78.40	**66.56**
	Tn$_{ar}$	34.00	75.99	82.61	91.69	17.96	23.17	**78.92**	63.99
	Lv$_{abz}$	**40.61**	76.18	82.49	90.76	**20.16**	**27.64**	77.90	60.36
	Lv$_{ar}$	33.68	**79.00**	86.84	90.59	16.75	24.13	75.41	**67.06**
AceGPT$_{8b}$	Eng	80.36	**89.59**	98.00	97.98	30.66	6.37	82.22	**7.27**
	MSA	81.39	86.77	**99.98**	98.89	30.20	6.26	87.34	3.43
	Dz$_{abz}$	79.23	85.88	97.00	98.03	29.54	6.26	77.52	**7.27**
	Tn$_{ar}$	**85.23**	85.23	**99.98**	**99.29**	30.02	6.26	77.30	3.43
	Lv$_{abz}$	78.36	78.50	34.21	96.21	**32.41**	3.59	83.66	3.43
	Lv$_{ar}$	80.53	84.29	**99.98**	98.59	31.77	6.26	**94.89**	3.43
AceGPT-chat$_{8b}$	Eng	79.48	84.24	98.00	97.98	30.24	6.26	84.54	7.27
	MSA	79.53	81.34	**99.98**	98.79	26.68	**6.29**	84.43	**32.37**
	Dz$_{abz}$	82.07	**87.13**	94.00	**99.34**	27.40	6.26	**85.08**	7.27
	Tn$_{ar}$	**82.48**	80.50	**99.98**	97.78	**31.87**	6.26	81.09	14.36
	Lv$_{abz}$	81.99	78.59	88.75	94.59	19.79	3.59	81.77	3.43
	Lv$_{ar}$	79.17	85.46	**99.98**	98.18	27.97	6.26	82.11	3.43
Jais-family$_{1p3b}$	Eng	13.26	20.52	32.43	**43.19**	**1.56**	**6.26**	34.21	**7.27**
	MSA	**18.61**	14.00	32.43	33.43	**1.56**	**6.26**	34.21	3.43
	Dz$_{abz}$	13.26	14.00	32.43	33.43	**1.56**	**6.26**	34.21	**7.27**
	Tn$_{ar}$	**18.61**	14.00	32.43	33.43	**1.56**	**6.26**	**52.61**	3.43
	Lv$_{abz}$	**18.61**	**32.20**	**52.49**	**43.19**	**1.56**	3.63	34.21	3.43
	Lv$_{ar}$	**18.61**	14.00	32.43	33.43	**1.56**	**6.26**	34.21	3.43

Table 11: Comparative performance of models across tasks, datasets, and dialects: Best performances are highlighted in green, worst in red, and each model's top scores are in bold. The first section of the table presents the best results for baseline and non-fine-tuned (NFT) models. The second and detailed section is dedicated to fine-tuned results.