

ResPL: Reasoning-Based Pseudo-Labels for Sarcasm Detection in Low-Resource Arabic Dialects

Ali Almutairi¹, Shoaib Jameel², Aditya Joshi¹,
Gelareh Mohammadi¹, Imran Razzak³

¹University of New South Wales, Australia, ²University of Southampton, UK

³MBZUAI, UAE

ali.almutairi@unsw.edu.au

Abstract

Sarcasm classification on Arabic dialects acutely suffer from low quantity and noisy labelled data. While few-label classification techniques typically employ pseudo-labelling, current pseudo-labelling methods are prone to noise accumulation. We introduce a novel framework called ResPL, leveraging Large Language Model (LLM) reasoning to generate higher-quality and interpretable pseudo-labels via a prompting strategy that engages the LLM in a Chain-of-Thoughts (CoT) process. Our method refines these reasoning-based labels through thematic categorisation and utilises them in an Incremental Learning (IL) paradigm to enrich the final model training with limited gold labels. We validate our approach on two dialectal sarcasm datasets using a range of base models and previously reported approaches. We report significant improvements over SOTA F-scores, affirming the value of reasoning-grounded pseudo-labels for low-resource Arabic dialect settings. Code, data and trained models are available at <https://github.com/z556190/ResPL>.

with a shortage of annotated datasets, computational tools, a nascent research community and limited institutional infrastructure and underfunded research frameworks. Despite that, Arabic is an ancient Semitic language with early attestations in first-millennium-BCE inscriptions that later developed into diverse regional and sub-regional dialects. (UNESCO, 2023). It can be divided into Modern Standard Arabic (MSA) and various local dialects, which reflect changes in the language and are used by over 422 million speakers worldwide (Faheem et al., 2024). The web, particularly social media, is a significant platform for Arabic content creation, primarily consisting of text-based posts. This textual data enables systematic analysis through computational linguistics methods, supporting tasks such as text classification, sentiment analysis, and discourse understanding. Nonetheless, the progression of Arabic research in NLP has been relatively slow, particularly in the context of few-label. The limited availability of labelled data further impedes the advancement of Arabic NLP, restricting the development of essential applications and affecting the broader NLP landscape.

1 Introduction

Technological advancements have spurred research in text analysis and classification, a crucial aspect in fields like healthcare and web security. For example, mechanisms to verify hate speech (Khurana et al., 2025) and algorithms to classify phishing attempts (Alshdadi, 2024) have been developed to address specific challenges.

However, text classification (TC) faces challenges that may hinder its effectiveness in Low-resource (LR) languages. Key obstacles include: (1) Datasets with limited labels expressed as few labelled datasets, and (2) the lack of effective pre-processing and language-specific tools. In contrast with languages with abundant resources, such as English and Chinese, Arabic is categorised as an LR language (Faheem et al., 2024), grappling

The challenges associated with the scarcity of labelled and unlabeled datasets have been explored from various perspectives, particularly concerning computational algorithms during both the data collection and processing stages. In the data collection phase, crowdsourcing (Yung and Demberg, 2025) may serve as a viable solution for labelling data in specific contexts. This approach can lead to excessive workload, negatively affecting output quality and increasing the risk of human error. Furthermore, multiple text data augmentation techniques have emerged, including Easy Data Augmentation (EDA) (Wei and Zou, 2019) and methods based on LLMs (Li et al., 2024). Notably, (Li et al., 2024) has leveraged LLMs to generate higher-quality observations through tailored instructions tailored for few-shot scenarios.

In the processing stage, several significant methodologies have been developed to enhance data quality. First, active learning (Margatina et al., 2023) involves a human intermediary who plays a crucial role in optimising data quality. However, this approach may necessitate the presence of specialists, which is not always feasible for all tasks. Second, semi-supervised (Yusuf et al., 2022) learning has the potential to automate the active learning process, thereby mitigating the need for annotators. Third, self-learning techniques generate Pseudo-Labels (PLs) and progressively refine them over time. Fourthly, an IL framework, exemplified by Flick (Almutairi et al., 2025), utilises clustering techniques on embeddings of unlabelled data, thereby generating PLs. These PLs are subsequently employed to initialise the CLS-FT model by integrating the weights derived from a previously trained version of the PL-FT model on the generated PLs. Nevertheless, it carries the risk of accumulating noise within the dataset. Additionally, the existing methodologies can often be ambiguous, particularly for linguistic researchers. Lastly, LLMs, especially those employing reasoning capabilities, can, to some extent, assist in few-shot learning tasks (Hasanaath et al., 2025), although they present significant operational costs and availability in the long run.

In light of the challenges identified above, this study investigates whether pseudo-labels that are grounded in explicit reasoning are more useful for few-label learning than pseudo-labels that carry only a class decision. We propose ResPL, a three-stage framework in which a reasoning-capable teacher LLM produces, for every unlabelled text, both a pseudo-label and a short rationale; these rationales are then abstracted into a compact inventory of semantic themes; and the resulting supervision is consumed by two successive student encoders, so that pseudo-label training and gold-label training never share a single optimisation stage. Distilling from a stronger teacher into a small encoder keeps privacy exposure, inference cost and API dependence low at deployment time, while separating the two students limits the propagation of pseudo-label noise into the final classifier. The formalism is given in Section 3 and the pipeline in Section 4. We evaluate ResPL on Arabic dialectal sarcasm detection, a task where the pragmatic cue is frequently carried by dialect- and culture-specific signals and where labelled data is particu-

larly scarce.

Our contributions can be summarised as follows:

- The development and preprocessing of two datasets that incorporate reasoning and categorise them into related themes, which will be made publicly available.
- The fine-tuning of two models for dialectal sarcasm detection in mono- and multi-dialect in few-shot settings where data is limited, which will also be publicly available.
- The evaluation of two Arabic datasets in dialectal sarcasm detection domains within a constrained data context.
- The establishment of a viable framework for few-label methods tailored to LR settings, particularly for under-resourced languages like Arabic.

2 Related Work

The scarcity of Arabic data in LR settings has been discussed through various methodologies focused on data acquisition and advanced algorithms. To mitigate this issue, strategies for obtaining additional data and efficiently managing annotators can be employed. However, challenges remain, including: (1) difficulty in collecting the necessary data; (2) potential inconsistencies in annotation quality due to high workloads; (3) lengthy annotation processes from large volumes of raw data; and (4) data privacy constraints. As a result, synthetic and augmented data serve as promising alternatives, although the labelling challenge persists. In terms of techniques, semi-supervised, weakly-supervised, and self-supervised learning algorithms have been implemented to integrate a limited number of labelled instances with unlabeled data, gradually enhancing the labelling budget through the generation of PLs. For instance, AlYami and Al-Zaidy (2022) employed a semi-supervised approach to effectively tackle the limitation of labelled data in the context of Arabic dialects, utilising two specific datasets: the Arabic Dialect Short Text and the Arabic Dialect Dictionary. Their integration of the AraRoBERTa model yielded promising outcomes. Moreover, Semi-supervised approaches have also proven highly effective for sarcasm detection despite data scarcity (Ameur et al., 2023). Nevertheless, it is crucial to acknowledge that accuracy alone may not serve as a sufficient evaluation metric, particularly in scenarios where dataset imbal-

ance is present. Despite employing sophisticated semi-supervised frameworks like semi-Supervised Generative Adversarial Network (SS-GAN) to enhance Arabic dialect identification, ? reported sub-optimal results under constrained data conditions, suggesting that SS-GAN approaches alone may not fully resolve the issue of model generalisability. Abdelali et al. (2024) explored few-shot prompting with GPT-4 for Arabic sarcasm detection, noting that performance remains limited when restricted to minimal observations. However, the proprietary nature of this model necessitates a subscription, presenting potential risks associated with long-term availability and feasibility. Recent advancements have also seen the application of reasoning models like DeepSeek R1 and GPT o4-mini to Arabic dialect and sarcasm detection, proving effective even in contexts with limited observations (Hasanaath et al., 2025). However, it is worth noting that these impressive results from open source reasoning models were derived from very large models, which demand significant computational resources for both training and inference.

These methodologies often suffer from noise during training, induced by the concurrent use of PLs and a limited number of true labels, which may introduce errors in the model’s progression. Consequently, a pivotal shift in these methodologies has involved the employment of clustering techniques to generate PLs and the segregation of the training phases for PLs and true labels. Shnarch et al. (2022) introduced the Cluster and Tune framework, which has gained recognition as a leading approach within the academic community. Additionally, Alsuhaibani et al. (2024) developed an enhanced version of this framework, termed ID-oFew, which outperformed its precursor on the same evaluative benchmarks. However, implementing multiple clusters has resulted in diminished model efficiency and increased operational costs. In response, an alternative framework called Flick (Almutairi et al., 2025) has been proposed, consisting of three key stages: (1) generating PLs through clustering the embeddings of unlabeled data; (2) refining these clusters to the top-k candidates with the aid of a language-specific Pre-Trained Model (PTM) that enhances performance; and (3) fine-tuning a small number of labelled examples using a language-related PTM (Cls-FT), initialising weights based on the prior stage.

While Flick has achieved SOTA results in LR

contexts with limited labelling (Almutairi et al., 2025), there remains room for improvement in the quality of PLs. The existing Pseudo-Labeling (PL) methodology is relatively rudimentary, and the rationale behind assigning PLs lacks clarity. Therefore, integrating the CoT reasoning model with the primary model could potentially enhance pseudo-label quality and elucidate the underlying reasoning for researchers. Nonetheless, since the reasons stem from individual data points, reliance on the parent model may be costly and inconsistently accessible. Moreover, obtained PLs might be leveraged by a thematic process to enrich the content of the few-labels in the final training stage.

3 Problem formulation

The problem is mathematically formulated as follows.

Let $\mathcal{U} = \{x_i\}_{i=1}^N$ denote the unlabelled pool, where each x_i is a piece of raw text and N is the number of training texts whose gold labels are withheld. Let $\mathcal{L} = \{(x_j, y_j)\}_{j=1}^M$ denote the small gold-labelled set available for the final training stage, where $y_j \in \mathcal{Y}$ is the human annotation of x_j . The index i ranges only over $\mathcal{U}$ and the index j ranges only over $\mathcal{L}$, and the two sets are disjoint ($\mathcal{U} \cap \mathcal{L} = \emptyset$). For the two sarcasm benchmarks used here $\mathcal{Y} = \{0, 1\}$, with 1 denoting sarcastic.

$$\mathcal{U} = \{x_i\}_{i=1}^N, \quad \mathcal{L} = \{(x_j, y_j)\}_{j=1}^M, \quad (1)$$

$$M = 100, \quad y_j \in \{0, 1\}.$$

To generate more reliable pseudo labels $\hat{y}_i$ for each x_i , we establish a connection between the raw text x_i and the reasoning r_i with $\hat{y}_i$ through a prompt template P directed at a teacher LLM G . This process yields a more dependable pseudo label $\hat{y}_i$, grounded in the reasoning r_i , derived from internal reasoning chains.

$$(\hat{y}_i, r_i) \leftarrow G(x_i; p), r_i \in R \quad (2)$$

However, the set of reasons $R = \{r_1, r_2, \dots, r_i\}$ exhibits high variance and inconsistencies. Consequently, clustering them into related themes t_i helps mitigate noise and enhances the training process for the Student B BERT-Based Model (BBM). A theme is defined as a group of semantically related textual data points. Each theme t_i is associated with its corresponding reasoning r_i through the teacher LLM G by inputting the reasons R and

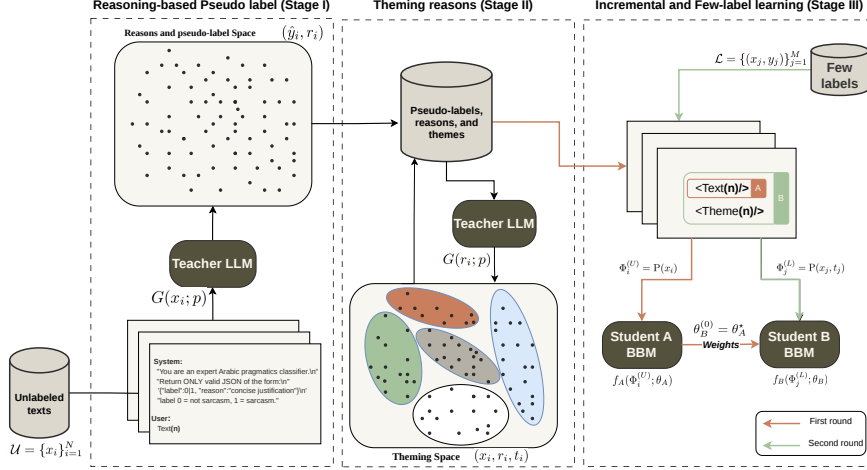


Figure 1: Our novel ResPL framework.

prompt P . The resulting set of extracted features, referred to as Theme T , minimises the noise from the reason set R and improves consistency.

$$t_i \leftarrow G(r_i; p), t_i \in \{1, \dots, T_d\} \quad (3)$$

Subsequently, we crafted prompt templates $\Phi_i^{(U)}$ and $\Phi_j^{(L)}$ to be input into the related model in the incremental and final training stages, respectively. Prompt template $\Phi_i^{(U)}$ incorporates x_i and is fed into a Student A (smaller BBM) f_A in the Incremental layer to initialise the weights for training in a later stage, which is evaluated using a loss equation (eq.6) presented below.

$$\Phi_i^{(U)} = P(x_i) \quad (4)$$

Finally, the weights obtained from the previous step are employed to initialise the weights (eq.7) for training a successive Student B (smaller BBM) f_B using prompt template $\Phi_j^{(L)}$, which consists of x_j, t_i as follows:

$$\Phi_j^{(L)} = P(x_j, t_j) \quad (5)$$

and is assessed against the actual labels as indicated in eq.8 below. We used Cross-Entropy (CE) as a cost function. $P(\cdot)$ is a transfer function where it transfer the format of the raw text to the following format: <text> raw text </s> <theme> theme name </s>.

$$\hat{\mathbf{p}}_i = f_A(\Phi_i^{(U)}; \theta_A), \quad Loss_A = \frac{1}{N} \sum_i CE(\hat{\mathbf{p}}_i, \hat{y}_i) \quad (6)$$

$$\theta_B^{(0)} = \theta_A^* \quad (7)$$

$$\hat{\mathbf{q}}_j = f_B(\Phi_j^{(L)}; \theta_B), \quad Loss_B = \frac{1}{M} \sum_j CE(\hat{\mathbf{q}}_j, y_j) \quad (8)$$

4 Methodology

This section outlines the key components of our methodology. Figure 1 illustrates our methodology.

4.1 Knowledge distillation

The utilisation of a powerful LLM, such as GPT-o3, to generate and preprocess textual data enhances the content used to train a less robust LLM, leading to improved performance in both the Incremental and final phases of model development. By transferring high-quality data from the teacher LLM to the student one, the subsequent performance is refined, positively impacting the reliability, availability, sustainability, and overall performance of the proposed framework. Furthermore, this knowledge will be applied to train Student A BBM by PLs, which will, in turn, initialise the Student B BBM for handling few-label data. We instantiate the teacher with GPT-o3, but the framework assumes only that the teacher can return a label together with a rationale in structured form (Stage I) and can group short texts into a bounded inventory (Stage II). Any reasoning-capable LLM satisfying this interface can be substituted; the trade-offs of doing so are discussed in the Limitations.

Reasoning-based Pseudo label While LLMs possess significant semantic understanding, generating direct PLs through LLMs may lead to less reliable outcomes due to factors such as lack of context, brittleness, and opacity. To improve the reliability of these PLs, we obligate the LLM to provide reasoning that supports its decisions for specific observations. This implicit connection between the LLM’s final answer and the rationale for each data point compels the model to engage in an internal reasoning process, drawing from context before arriving at a final decision backed by reasoning. By requesting a rationale, the LLM is prompted to create an interpretability pathway that utilises latent knowledge, enhancing the strength of the resulting PLs.

Theming reasons To reduce the high variance and noise present in the context window of the reasons, and to enhance the context for the training phase of the student B, we extract a finite set of themes from the reasons space by prompting strategy to the teacher LLM. A theme is defined as a group of semantically related textual data points. Each data point is then assigned to a theme through a semantically comprehensive computation using a robust LLM. Theming is performed by the teacher in a single pass over the pooled reason set; it is an LLM-driven semantic grouping, not embedding-space clustering as in ClusterTune, IDoFew or Flick. We do not fix T_d in advance; the teacher is constrained to return a mutually exclusive, collectively exhaustive inventory of 8–20 themes (Appendix A.2) and chooses the granularity itself. In practice it returned 8 themes for Ar-Sarcasm and 7 for Sa’7r; the full inventories, with English glosses, are given in Table 6. The knowledge obtained from this more reliable LLM will improve the performance of the student B, as it incorporates higher-quality insights from the teacher LLM. This structured contextual information streamlines the training process of the later BBM by providing more focused content. As the text and theme, and a few real labels are processed in a structured manner, we will employ an IL approach to optimise the student B model for few-shot learning.

4.2 Incremental learning and Few-label

In stage 3, we employed two student BBMs to separate the pseudo-label training from the few-label training. This approach helps mitigate the risk of error accumulation that can occur when combining

pseudo and real labels in a single model. Additionally, we utilised the pseudo-label training capabilities of Student A BBM to initialise the weights for Student B BBM, as it shares some similarities with the task at hand. This strategy, combined with the few labels, facilitates a more effective continuation and adjustment of the training process.

4.2.1 Student A in Incremental learning stage

In the context of IL, a preliminary model (student A BBM) is developed prior to the final model (student B BBM), facilitating a continuous evolutionary process aimed at recognising patterns analogous to those employed in the subsequent model. This initial training phase is instrumental in ensuring that the final model achieves a level of proficiency that is adequate for processing a limited dataset effectively.

4.2.2 Student B in Few-label learning stage

The weights derived from the Student A model will subsequently be utilised in the Student B model to facilitate the training of 100 real labelled data instances within the same model class. Moreover, the thematic tags enrich the context window alongside the text. Our framework has been applied across two Arabic datasets, demonstrating its effectiveness in both cross-domain and LR contexts.

4.3 Prompt Strategy

To effectively transfer knowledge from Teacher LLMs, our prompt template strategy incorporates context, tasks, instructions, and constraints. This approach guides the reasoning behaviour of the LLM before it generates output. Then, we leverage this new knowledge by theming related sematical texts. Later, we use this knowledge for smaller models’ enhancements. We have designed three distinct prompt templates. First, we instruct the Teacher LLM to produce PLs supported by reasoning. This prompts the Teacher LLM to engage in CoT process prior to generating these PLs. Concretely, the Stage I instruction is:

System Prompt

You are an expert Arabic sarcasm classifier.
Return ONLY valid JSON of the form:
{ "label": 0|1, "reason": "concise justification" }
label 0 = not sarcasm, 1 = sarcasm.

Emitting the label and the justification inside a single JSON object forces the teacher to commit to a rationale in the same forward pass that produces the decision, rather than rationalising a label

Dataset	Train	Eval	Test	Setting
Ar-Sarc.	6,005	2,432	2,110	Multidialectal
Sa'7r	4,058	3,458	1,186	Monodialectal

Table 1: Dataset splits and dialect settings.

after the fact. Full templates for all three stages are given in Appendix A. Next, we build on this newfound knowledge by categorising the reasons into related themes by prompting the Teacher LLM. More information can be found in prompt template in Appendix A.2. Finally, we utilise text and theme tags to transfer the raw text to incorporate the new information as additive knowledge alongside raw text in the Student B BBM. Refer to prompt templates in Appendix A.3 and A.4 for the Student A and Student B BBMs, respectively.

5 Results and Discussion

Dataset Two distinct publicly available Arabic datasets were used to assess the ResPL framework, including (1)Ar-Sarcasm (Abu Farha and Magdy, 2020) (2) The Irony of Saudi Dialect (Sa'7r) (Al-Mazrui et al., 2022). The ArSarcasm dataset served as the basis for the Ar-Sarcasm dataset. Every dataset includes sets for training, evaluation, and testing. Each dataset’s details are summarised in Table 1. Lastly, 100 balanced labelled observations were used to train the Student B model to mimic LR settings. For detailed dataset sizes and token statistics, see Appendix E.

We evaluate ResPL on two Arabic text-classification benchmarks, Ar-Sarcasm and Sa'7r, using macro-F1. All models, including baselines, are fine-tuned under an identical 100-label budget and training protocol to ensure a fair comparison. In Appendix C, Table 5 reports the hyperparameters used by ResPL and all re-run baselines. Table 2 presents a comparison of ResPL against 11 Arabic and English BERT-based models (BBMs), as well as 6 SOTA few-label methods originally designed for English and re-evaluated here in both their original form and with an AraBERTv2 backbone.

Dialect-exclusive fine-tuning generalises poorly. Under 100-label SFT, the dialect-exclusivity pattern becomes even more pronounced. EgyBERT, trained exclusively on Egyptian tweets and forum text, achieves the single highest score on Ar-Sarcasm among all BBMs (64.29), consistent with Egyptian content comprising approximately 47.5% of the sarcastic tweets in that dataset (Abu Farha

Model	Ar-Sarcasm	Sa'7r	Mean
<i>Zero-shot (no fine-tuning)</i>			
Arabic-BERT	45.77	32.63	39.20
AraBERTv2	47.56	32.92	40.24
AraBERT-Twitter (large)	46.76	38.52	42.64
MARBERTv2	50.40	51.74	51.07
<i>Supervised Fine-tuning on BERT-Based Model (BBMs)</i>			
Arabic-BERT	56.71	49.70	53.20
ARBERTv2	60.42	64.82	62.62
AraBERT-Twitter (large)	60.97	64.42	62.70
AlcLaMv2	63.51	62.55	63.03
CamelBERT-mix	57.95	50.98	54.47
CamelBERT-MSA-DID	55.24	61.05	58.14
MARBERTv2	56.28	62.59	59.44
EgyBERT	64.29	38.72	51.51
SaudiBERT	60.52	61.40	60.96
BERT (English)	54.07	53.48	53.77
DistilBERT (English)	51.95	45.68	48.81
mBERT	48.13	51.28	49.70
<i>Few-label SOTA Methods</i>			
Cluster&Tune *	44.86	46.45	45.66
Cluster&Tune	48.33	65.11	56.72
IDoFew *	49.81	64.28	57.05
IDoFew	62.40	66.18	64.29
FLICK *	<u>69.24</u>	<u>67.96</u>	<u>68.60</u>
FLICK	62.80	66.69	64.75
ResPL	72.39	69.98	71.19

Table 2: Macro-F1 (%) on Ar-Sarcasm (multi-dialect) and Sa'7r (Saudi dialect). "*" indicates original published hyperparameters; all other baselines use the same budget and training parameters as ResPL. **Bold** = best per column; underline = second best.

and Magdy, 2020). On Sa'7r, however, EgyBERT collapses to 38.72, a cross-dialect drop of 25.57 points, because its representations are tuned to the Egyptian dialect that is absent from Saudi discourse. More broadly, several Arabic BBMs achieve relatively balanced performance across the two dialectally diverse sarcasm datasets, suggesting that their pre-training learns representations that transfer across different Arabic varieties. These results expose a fundamental limitation of the dialect-exclusive and non-exclusive pre-training paradigm: a model that excels at one Arabic dialectal sarcasm variety tends to fail at another. While pre-training appears to provide a degree of cross-dialect generalisation, the relatively modest gains observed under 100-label SFT suggest that standard few-label fine-tuning may not be sufficient. This observation points to the potential value of developing more effective fine-tuning strategies for LR dialects.

Multi-dialect and dialect-mixed models are more robust. Across the SFT BBMs group, the most consistent cross-dataset performers are models with broad dialectal coverage. ARBERTv2 (60.42/64.82) and AraBERT-Twitter (60.97/64.42) both achieve higher Sa’7r scores than on Ar-Sarcasm, suggesting that their broader pre-training distributions generalise better to Saudi dialect once even a small number of task labels are provided. MARBERTv2’s Twitter-heavy pre-training similarly yields a Sa’7r advantage under SFT (56.28/62.59), despite underperforming EgyBERT on Ar-Sarcasm. AlcLaMv2, trained from scratch on a 13 GB dialectal social-media corpus, achieves the strongest mean across all BBMs (63.51/62.55; mean 63.03), demonstrating that dialect-diverse yet domain-matched pre-training can be more effective than scale alone. Notably, CamelBERT-MSA-DID, a model fine-tuned for dialect identification on Twitter, shows a Sa’7r advantage of +5.81 points over Ar-Sarcasm (55.24/61.05), supporting the hypothesis that explicit dialect awareness in the feature space improves Saudi sarcasm detection even when that awareness was trained for a different task. English-only and multi-lingual models (BERT, DistilBERT, mBERT) are mostly below their Arabic counterparts on both datasets, suggesting that linguistic mismatch is a substantial barrier regardless of model scale. ResPL achieves +8.10 points over EgyBERT on Ar-Sarcasm and +5.16 points over ARBERTv2 on Sa’7r, demonstrating that improvements in the fine-tuning process can enhance the limitations of conventional few-label adaptation. Notably, these gains are obtained using a general-purpose Arabic BERT backbone rather than a dialect-specialised model.

Few-label SOTA methods are sensitive to dialectal mismatch. The backbone swap between two languages (English and Arabic) critically determines whether a SOTA few-label method can handle dialectal sarcasm. Cluster&Tune* with its original English-oriented backbone reaches only 44.86/46.45; switching to ARBERTv2 boosts Sa’7r by +18.66 points (to 65.11) while Ar-Sarcasm improves only marginally (+3.47), confirming that the clustering step benefits disproportionately from an Arabic-aware backbone on the Saudi dialect task. IDoFew follows a similar pattern: its ARBERTv2 variant gains +1.90 on Sa’7r over the original while gaining +12.59 on Ar-Sarcasm, reflecting that ID-

Model	Ar-Sarcasm	Sa’7r	Mean
ALLaM-7B-Instruct	54.91	52.89	53.90
LLaMA 3.1-8B	40.38	45.60	42.99
LLaMA 3.2-1B-Instruct	53.14	36.23	44.69
AceGPT-13B	54.24	49.63	51.94
Qwen2.5-72B-Instruct	14.05	34.11	24.08
GPT-4o	<u>67.14</u>	<u>65.18</u>	<u>66.16</u>
ResPL	72.39	69.98	71.19

Table 3: Macro-F1 (%) comparing ResPL with fine-tuned LLMs. **Bold** = best; underline = second best.

oFew’s instance-selection strategy might capture cross-dialect sarcasm variance more readily in the multi-dialect corpus. FLICK*, the highest baseline (69.24/67.96) among them, is the only SOTA method to maintain sub-2-point parity across both datasets under its original configuration, a result attributable to its design targetted Arabic language and its feature-space clustering operating over a richer label-informed representation. Under our parameter budget, FLICK drops on Ar-Sarcasm and on Sa’7r. ResPL surpasses FLICK* by +3.15 points on Ar-Sarcasm and +2.02 points on Sa’7r, establishing a new state of the art on both dialectal sarcasm benchmarks under the 100-label constraint.

Table 3 compares ResPL against six LLMs fine-tuned under the same 100-label budget and training protocol. This comparison is designed to examine whether LLMs exhibit differential sensitivity to the two dialectal sarcasm settings and evaluate whether a compact BERT-based model, trained via our continual learning pipeline, can compete with parameter-heavy LLMs.

LLMs are not uniformly sensitive to Arabic dialect in sarcasm. Unlike BBMs, whose per-dataset gaps follow a predictable pattern tied to pre-training dialect coverage, LLMs under few-label fine-tuning exhibit erratic cross-dialect behaviour. LLaMA 3.2-1B-Instruct scores 53.14 on Ar-Sarcasm but collapses to 36.23 on Sa’7r ($\Delta = -16.91$), a steeper Saudi-dialect drop than even Arabic-BERT at zero-shot. This collapse likely reflects the combination of its English-dominant pre-training and the severely constrained fine-tuning signal: 100 Saudi dialect sarcasm samples are insufficient to re-orient a model whose representations of Arabic are both sparse and MSA-biased. Qwen2.5-72B-Instruct presents the most dramatic case: it scores only 14.05 on Ar-Sarcasm, well

Configuration	Ar-Sarcasm	Sa'7r	Mean
Zero-shot	47.56	32.92	40.24
Only Student B ^N	60.42	64.82	62.62
Only Student B ^T	55.05	59.84	57.45
Only Student A ^{NR}	65.86	57.29	61.58
Only Student A ^R	70.81	69.93	70.37
Full ResPL ^N	63.05	67.36	65.20
Full ResPL^T	72.39	69.98	71.19

Table 4: Ablation study (Macro-F1, %). Each row removes or modifies one ResPL component. ^R: reasoning-based pseudo-labels; ^{NR}: non-reasoning pseudo-labels; ^T: theme tags enabled; ^N: no theme tags. Best results are shown in **bold**.

below random for the sarcastic class in a class-imbalanced binary task, yet rises to 34.11 on Sa'7r. This inversion suggests that fine-tuning a large instruction-tuned model on 100 multi-dialect examples catastrophically disrupts its internal representation of cross-dialect sarcasm, whereas the more homogeneous Saudi dialect signal allows some residual alignment. Arabic-aware LLMs (ALLaM-7B, AceGPT-13B) show narrower cross-dialect gaps -2.02 and -4.61 respectively), consistent with their Arabic pre-training providing more stable dialectal grounding. GPT-4o fine-tuned is the best LLM overall (67.14/65.18; mean 66.16) and maintains near-parity across dialects (-1.96), suggesting that its proprietary fine-tuning pipeline may include regularisation that limits dialectal overfitting under low-resource conditions. ResPL surpasses GPT-4o fine-tuned by $+5.25$ on Ar-Sarcasm and $+4.80$ points on Sa'7r, while using a publicly available small encoder model, enabling strong performance across both dialectal sarcasm settings without relying on a closed API or models with massive parameter counts.

5.1 Ablation study

To isolate the contribution of each component in ResPL, we conduct a systematic ablation over 2 benchmarks. Table 4 reports macro-F1 scores for 7 configurations: the full method and 6 degraded variants obtained by isolating individual pipeline stages. **Firstly**, reasoning-based PLs appear to be the most consequential component, particularly for Saudi dialect. Student A with reasoning PLs outperforms its non-reasoning counterpart by $+4.95$ points on Ar-Sarcasm and $+12.64$ points on Sa'7r ($57.29 \rightarrow 69.93$): trigger chain-of-thought reason-

ing by the teacher LLM can surface the pragmatic intent behind each utterance in ways that a plain label cannot, though the exact mechanism warrants further investigation. **Secondly**, theme tags are beneficial when combined with Student A's reasoning-grounded representations. Injecting themes into plain SFT (Student B only, with themes) reduces performance relative to the no-theme SFT baseline (-5.37 on Ar-Sarcasm; -4.98 on Sa'7r), suggesting that theme addition introduces noise unless the model has already been exposed to reasoning-based PLs supervision. Conversely, the full pipeline with theme tags gains $+9.34$ on Ar-Sarcasm over the no-theme variant ($63.05 \rightarrow 72.39$), while the gain on Sa'7r is more modest ($+2.62$), possibly because Sa'7r's narrower topical scope makes additional thematic grouping less informative. **Finally**, the two components appear complementary across dialectal sarcasm settings: reasoning PLs provide the larger relative gain on the monodialectal Saudi corpus, whereas theme tags provide the larger relative gain on the topically diverse multi-dialect corpus, suggesting each addresses a different source of within-class variance.

6 Conclusion

In this study, we investigated the use of a reasoning-capable LLM to generate pseudo-labels and accompanying rationales for low-resource Arabic sarcasm detection. The results suggest that reasoning-based pseudo-labels can improve the effectiveness of the proposed ResPL framework across both monodialectal and multi-dialectal Arabic sarcasm classification under limited labelled data. In addition, the generated rationales provide explicit explanations for why a particular pseudo-label was assigned, offering greater transparency into the pseudo-labels generation process than label-only annotations. We also observed that incorporating these reasoning-based PLs may contribute to higher-quality supervision signals, although the precise mechanisms behind these improvements require further investigation. Across the evaluated datasets, ResPL outperformed the compared baselines, including prior few-label methods for Arabic sarcasm detection. Future research could quantify how far the framework's gains survive substitution of an open-weight reasoning teacher, explore more synthetic data generation strategies, and investigate stronger intermediate-learning approaches for few-label scenarios.

Limitations

Several limitations warrant consideration. First, ResPL instantiates the teacher with GPT-o3, a closed-source model, which constrains reproducibility, imposes a one-off cost linear in $|U|$, and excludes practitioners without API access; we mitigate this by releasing the complete Stage I output, all pseudo-labels, rationales and theme assignments, together with the fine-tuned student checkpoints, so that Stages II and III can be reproduced and our reported scores verified without a single API call. Nothing in the framework requires a proprietary teacher, and substituting an open-weight reasoning model is a drop-in replacement at the interface level, though we did not run this substitution and do not claim our gains would survive it, since Table 4 shows rationale quality to be the component to which the framework is most sensitive. Second, our evaluation covers only Egyptian-dominant and Saudi dialectal settings; Iraqi and Sudanese varieties are absent, and because Arabic sarcasm cues are frequently dialect- and culture-specific (Appendix D), transfer to those varieties, and to other low-resource languages, remains unassessed. Third, the size of the theme inventory is delegated to the teacher within a permitted range rather than tuned, so we cannot report how sharply performance depends on T_d . Fourth, all results come from a single seed (0), fixed identically for ResPL and every re-run baseline. Finally, we did not test augmented or synthetic unlabelled data during the incremental phase, and the framework remains text-only, whereas social-media sarcasm is often carried by images, emoji and quoted context.

References

- Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 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, St. Julian’s, Malta. Association for Computational Linguistics.
- Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. [ARBERT & MARBERT: Deep bidirectional transformers for Arabic](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.
- Ibrahim Abu Farha and Walid Magdy. 2020. [From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset](#). In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.
- Murtadha Ahmed, Saghir Alfasly, Bo Wen, Jamal Addeen, Mohammed Ahmed, and Yunfeng Liu. 2024. [AlclAM: Arabic dialect language model](#). In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 153–159, Bangkok, Thailand. Association for Computational Linguistics.
- Halah AlMazrui, Najla AlHazzani, Amaal AlDawood, Lama AlAwlaqi, Noura AlReshoudi, Hend AlKhalifa, and Luluh AlDhubayi. 2022. [Sa’7r: A saudi dialect irony dataset](#). In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur’an QA and Fine-Grained Hate Speech Detection*, pages 60–70, Marseille, France. European Language Resources Association.
- Ali Almutairi, Abdullah Alsuhaibani, Shoaib Jameel, Usman Naseem, Gelareh Mohammadi, and Imran Razzak. 2025. [Flick: Few labels text classification using k-aware intermediate learning in multi-task low-resource languages](#).
- Abdulrahman A. Alshdadi. 2024. [LSTM-PSO: NLP-based model for detecting phishing attacks](#). In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 70–79, Lancaster, UK. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.
- Abdullah Alsuhaibani, Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2024. [Idofew: Intermediate training using dual-clustering in language models for few labels text classification](#). In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM ’24*, page 18–27, New York, NY, USA. Association for Computing Machinery.
- Reem AlYami and Rabah Al-Zaidy. 2022. [Weakly and semi-supervised learning for Arabic text classification using monodialectal language models](#). In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 260–272, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Asma Ameer, Sana Hamdi, and Sadok Ben Yahia. 2023. [Domain adaptation approach for arabic sar-](#)

- casm detection in hotel reviews based on hybrid learning. *Procedia Computer Science*, 225:3898–3908. 27th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES 2023).
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. **AraBERT: Transformer-based model for Arabic language understanding**. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.
- M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaiyan, Ali Alammari, Zaki Alawami, Abdulmohsen AlThubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairish, Areeb Alowisheq, and Haidar Khan. 2025. **ALLam: Large language models for arabic and english**. In *The Thirteenth International Conference on Learning Representations*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. **BERT: pre-training of deep bidirectional transformers for language understanding**. *CoRR*, abs/1810.04805.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Mohamed Atta Faheem, Khaled Tawfik Wassif, Hanaa Bayomi, and Sherif Mahdy Abdou. 2024. **Improving neural machine translation for low resource languages through non-parallel corpora: a case study of egyptian dialect to modern standard arabic translation**. *Scientific Reports*, 14:2265.
- Ahmed Hasanaath, Aisha Alansari, Ahmed Ashraf, Chafik Salmane, Hamzah Luqman, and Saad Ezzini. 2025. **Arareasoner: Evaluating reasoning-based llms for arabic nlp**.
- 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*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Urja Khurana, Eric Nalisnick, and Antske Fokkens. 2025. **DefVerify: Do hate speech models reflect their dataset’s definition?** In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4341–4358, Abu Dhabi, UAE. Association for Computational Linguistics.
- Yichuan Li, Kaize Ding, Jianling Wang, and Kyumin Lee. 2024. **Empowering large language models for textual data augmentation**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12734–12751, Bangkok, Thailand. Association for Computational Linguistics.
- Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. **Active learning principles for in-context learning with large language models**. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034, Singapore. Association for Computational Linguistics.
- Meta. 2024. **Introducing meta llama 3: The most capable openly available llm to date**.
- OpenAI. 2025. **GPT Models (April 14 version)**. Large language model.
- Faisal Qarah. 2024a. **Egybert: A large language model pretrained on egyptian dialect corpora**. *arXiv preprint arXiv:2408.03524*.
- Faisal Qarah. 2024b. **Saudibert: A large language model pretrained on saudi dialect corpora**. *arXiv preprint arXiv:2405.06239*.
- Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. **KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media**. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. **Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter**.
- Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2022. **Cluster & tune: Boost cold start performance in text classification**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7639–7653, Dublin, Ireland. Association for Computational Linguistics.
- Qwen Team. 2024. **Qwen2.5: A party of foundation models**.
- UNESCO. 2023. **Arabian chronicles in stone: The jabal ikmah inscriptions**. <https://www.unesco.org/en/memory-world/arabian-chronicles-stone-jabal-ikmah>. Accessed: 22 September 2025.
- Jason Wei and Kai Zou. 2019. **EDA: Easy data augmentation techniques for boosting performance on text**

classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Frances Yung and Vera Demberg. 2025. [On crowdsourcing task design for discourse relation annotation](#). In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 12–19, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Mahmoud Yusuf, Marwan Torki, and Nagwa El-Makky. 2022. [Arabic dialect identification with a few labeled examples using generative adversarial networks](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 196–204, Online only. Association for Computational Linguistics.

Jianqing Zhu, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Mosen Alharthi, Bang An, Juncai He, Xiangbo Wu, Fei Yu, Junying Chen, Zhuoheng Ma, Yuhao Du, Yan Hu, He Zhang, Emad A. Alghamdi, Lian Zhang, Ruoyu Sun, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Second language (arabic) acquisition of llms via progressive vocabulary expansion.

A Prompt templates

A.1 Reason-based PLs prompt template 1

A.1.1 For sarcasm datasets

System Prompt

You are an expert Arabic sarcasm classifier.
Return ONLY valid JSON of the form:
{ "label": 0|1, "reason": "concise justification" }
label 0 = not sarcasm, 1 = sarcasm.

A.2 Theming reasons prompt template 2

A.2.1 For sarcasm datasets

System Prompt

You are an expert in taxonomy design for NLP tasks. Your job is to read a set of short “reasons” people gave for labeling Arabic texts into sarcasm or not sarcasm, and group those reasons into a compact, non-overlapping set of THEMES.

- A theme = a group of semantically related texts.
- Aim for 8–20 themes (whatever is most natural).
- Themes must be mutually exclusive and collectively exhaustive.

Return ONLY valid JSON matching this schema:

```
{
  "themes": [
```

```
{
  "id": "<INT>",
  "name": "<SHORT TITLE>",
  "description": "<1-sentence explanation>",
  "example_reasons": [
    "<verbatim reason 1>",
    "<verbatim reason 2>",
    "...
  ]
}
```

A.3 Student A BBM prompt template 3

System Prompt

```
<|text|> raw_text </s>
```

A.4 Student B BBM prompt template 4

System Prompt

```
<|text|> raw_text </s> <|theme|> theme_name </s>
```

B Baseline Models

PTMs Models: We conducted a comparison of our framework with well-known publicly accessible language models such as BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), and mBERT (Devlin et al., 2018).

Arabic-PTMs Models: Our research also incorporated ArabicBERT. The models considered include ArabicBERT (Safaya et al., 2020), AraBERTv2 (Antoun et al., 2020), AraBERT-Twitter (Antoun et al., 2020), AlcLaMv2 (Ahmed et al., 2024), CamelBERT-Mix (Inoue et al., 2021), CamelBERT-MSA-Mix (Inoue et al., 2021), ARBERTv2 and MARBERTv2 (Abdul-Mageed et al., 2021), EGYBERT (Qarah, 2024a), and SaudiBERT (Qarah, 2024b).

LLMs: We compared our ResPL framework with fine-tuned LLMs using supervised fine-tuning (SFT), recognising their effectiveness in managing complex tasks. For this purpose, we select AllAM-7B (Bari et al., 2025), LLaMA 3.1-8B and LLaMA 3.2-1B-Instruct (Meta, 2024), AceGPT (Zhu et al., 2024), Qwen-72B-instruct (Team, 2024), and GPT-4o (OpenAI, 2025) as representative models in our comparison.

Intermediate frameworks: To assess the performance of our intermediate stage, we compared our model’s results with those of Cluster & Tune (Shnarch et al., 2022), IDoFew (Alsuhaibani et al., 2024), and Flick (Almutairi et al., 2025), all of which utilise an inter-training task.

Dataset	Stage	Epochs	LR	BS
Ar-Sarcasm	Student A	1	2×10^{-5}	8
	Student B	10	5×10^{-5}	8
Sa'7r	Student A	3	2×10^{-5}	64
	Student B	20	3×10^{-5}	64

Table 5: Hyperparameters for ResPL and all re-run baselines. Student A = pseudo-label fine-tuning; Student B = classification fine-tuning. Both stages use the Adam optimiser with seed 0. LR = Learning rate and BS = Batch Size.

C Parameter settings

All ResPL experiments use ARBERTv2 as the backbone and GPT-o3 (OpenAI, 2025) as the teacher reasoning LLM for pseudo-label generation and theme extraction. For the non-reasoning pseudo-label ablation, we additionally employed GPT-3.5 Turbo (OpenAI, 2025) to generate pseudo-labels without reasoning, allowing us to assess the contribution of reasoning-based pseudo-label generation. The two training stages, Student A (pseudo-label fine-tuning) and Student B (classification fine-tuning) are run sequentially with the Adam optimiser, a fixed random seed of 0, and dataset-specific hyperparameters reported in Table 5. All BBM, LLM baselines and SOTA re-runs share the same per-dataset hyperparameters as ResPL for a fair comparison; only rows marked $\star$ retain their originally published training configurations. LLM baselines are fine-tuned on the same 100-label budget, where GPT-4o is fine-tuned over the API, and the open-source LLMs are fine-tuned locally.

D Error analysis

In our subsequent investigations, we conducted a randomised sampling of 30 mislabelled observations. These misclassifications revealed distinct patterns across various tasks and datasets. In the sarcasm datasets: (1) Ar-Sarcasm and (2) Sa'7r, it was observed that the smaller LLM frequently misinterpreted the use of metaphorical devices. This confusion arises from the dual nature of metaphors, which can serve as serious expressions intended to clarify meaning or as devices employed to convey sarcasm. Notably, there were instances where the smaller LLM accurately predicted the correct label, indicating potential issues in the annotation process for certain observations. It is crucial to consider

that cultural differences in the expression of sarcasm may contribute to these discrepancies, particularly as the Ar-Sarcasm dataset is predominantly derived from Egyptian sources, while the Sa'7r dataset is primarily based on Saudi contexts. Moreover, the lack of evident contextual clues makes it challenging to accurately assess the intended meaning, which is a critical factor in the annotative process affecting classification outcomes. Further analysis indicated that the smaller LLM encounters difficulties in differentiating between sarcastic and non-sarcastic text within specific Arabic dialects, such as Saudi Arabic. In this context, words commonly used in a religious context, like Text¹: الحمد لله والشكر, are sometimes applied to convey sarcasm, which presents subtle cultural nuances that the smaller LLM struggles to capture.

E Datasets statistics

We report dataset sizes and token statistics across all splits in Tables 7–8.

¹Text: "Praise and thanks be to God."

ID	Arabic Theme	English Translation
Ar-Sarcasm		
1	العدوان المباشر دون سخرية	Direct hostility without sarcasm
2	مؤثرات النبرة القائمة على الشكل	Form-based tone markers
3	الاستفسارات والطلبات الحقيقية	Genuine questions or requests
4	المبالغة والتحويل غير المعقول	Hyperbole and implausible exaggeration
5	السخرية من خلال عكس المعنى	Sarcasm through meaning reversal
6	النبرة الحرفية الواقعية	Literal and factual tone
7	السخرية باستخدام السؤال البلاغي	Sarcasm through rhetorical questions
8	المشاعر الإيجابية أو العاطفية الصادقة	Genuine positive or emotional sentiment
Sa'7r		
1	المبالغة والتحويل	Hyperbole and exaggeration
2	التصريح المباشر بالمزاح	Explicit joking statements
3	التعبير الديني أو الامتنان الصادق	Religious expressions or sincere gratitude
4	عبارة واقعية محايدة	Neutral factual statement
5	مزاح خفيف وودّي	Light-hearted and friendly humour
6	مؤثرات النبرة أو الرموز التعبيرية	Tone markers or emojis
7	الفكاهة السوداء والكوميديا السوداء	Dark humour / black comedy

Table 6: Examples of themes extracted by the teacher LLM for the sarcasm datasets. Arabic themes are shown with English translations.

Split	#Rows	Tok. Mean	Tok. Med.	Tok. Max	Tok. Total	Word Mean	Char Mean
Pseudo-Labeled	3958	50.67	41	175	200571	19.49	111.88
Real-Labeled	100	50.31	42	151	5031	19.87	113.85
Eval	3458	34.05	25	143	117751	14.27	79.59
Test	1186	34.59	26	181	41019	14.31	80.62

Table 7: Statistics of the Sa'7r dataset splits.

Split	#Rows	Tok. Mean	Tok. Med.	Tok. Max	Tok. Total	Word Mean	Char Mean
Pseudo-Labeled	5905	66.50	68	182	392659	21.68	157.65
Real-Labeled	100	65.71	67	163	6571	21.50	154.83
Eval	2432	46.10	48	156	112109	14.80	104.04
Test	2110	45.63	47	135	96272	14.82	103.56

Table 8: Statistics of the Ar-Sarcasm dataset splits.