# Bridging Dialectal Gaps in Arabic Medical LLMs through Model Merging

**Ahmed Ibrahim, Abdullah Hosseini, Hoda Helmy, Wafa Lakhdhar, and Ahmed Serag**
AI Innovation Lab, Weill Cornell Medicine - Qatar, Doha, Qatar
{azi4002, abh4006, hoh4002, wal4005, afs4002}@qatar-med.cornell.edu

## Abstract

The linguistic fragmentation of Arabic, with over 30 dialects exhibiting low mutual intelligibility, presents a critical challenge for deploying natural language processing (NLP) in healthcare. Conventional fine-tuning of large language models (LLMs) for each dialect is computationally prohibitive and operationally unsustainable. In this study, we explore model merging as a scalable alternative by integrating three pre-trained LLMs—a medical domain expert, an Egyptian Arabic model, and a Moroccan Darija model—into a unified system without additional fine-tuning. We introduce a novel evaluation framework that assesses both dialectal fidelity via dual evaluation: LLM-based automated scoring and human assessments by native speakers. Our results demonstrate that the merged model effectively handles cross-dialect medical scenarios, such as interpreting Moroccan Darija inputs for Egyptian Arabic-speaking clinicians, while maintaining high clinical relevance. The merging process reduced computational cost by over 60% compared to per-dialect fine-tuning, highlighting its viability for resource-constrained settings. This work offers a promising path for building dialect-aware medical LLMs at scale, with implications for broader deployment across linguistically diverse regions.

## 1 Introduction

The Arabic language landscape, characterized by profound linguistic fragmentation into numerous regional dialects, presents a formidable challenge for Natural Language Processing (NLP), particularly in high-stakes domains like healthcare (Alasmari, 2025; Inoue et al., 2022). While Modern Standard Arabic (MSA) serves a unifying function, daily communication—including critical patient-clinician interactions—occurs predominantly in local dialects. These dialects, such as Egyptian Arabic and Moroccan Darija, often exhibit stark phonological and lexical divergence, severely limiting mutual intelligibility across geographical distances (Trentman and Shiri, 2020). This fragmentation creates tangible and potentially dangerous communication barriers within healthcare systems: patients describing symptoms in their native dialect may be misunderstood by clinicians unfamiliar with its nuances, leading to misdiagnosis, ineffective treatment, or delayed care (Shoufan and Alameri, 2015).

Addressing this challenge through conventional Large Language Model (LLM) finetuning is fraught with difficulty (Wu et al., 2025; Ibrahim et al., 2025a,b). Training separate, specialized medical language models for each major dialect is prohibitively resource-intensive, requiring vast amounts of annotated dialectal medical data and significant computational power for each variant even with quantization methods (Hu et al., 2022; Brown et al., 2020). This approach is fundamentally unscalable given the sheer number of Arabic dialects and the continuous resource constraints faced in many regions. Consequently, there is an urgent need for efficient and scalable methodologies that can bridge dialectal gaps in specialized domains without the burden of training and maintaining numerous individual models.

This paper investigates a solution to this problem: leveraging model merging techniques (Brunet et al., 2006; Xu et al., 2024) to consolidate specialized capabilities into a single, unified model. We explore the feasibility of integrating pre-trained LLMs possessing distinct expertise—specifically, an Egyptian Arabic dialect expert, a Moroccan Darija expert, and a general medical-domain model, without resorting to further fine-tuning. Our core research question is: *Can model merging yield a single, resource-efficient language model capable of robustly handling critical cross-dialect medical communication tasks?*

We adopt a rigorous validation strategy combining automated evaluation with human assessment.



*Proceedings of The Third Arabic Natural Language Processing Conference*, pages 338–346

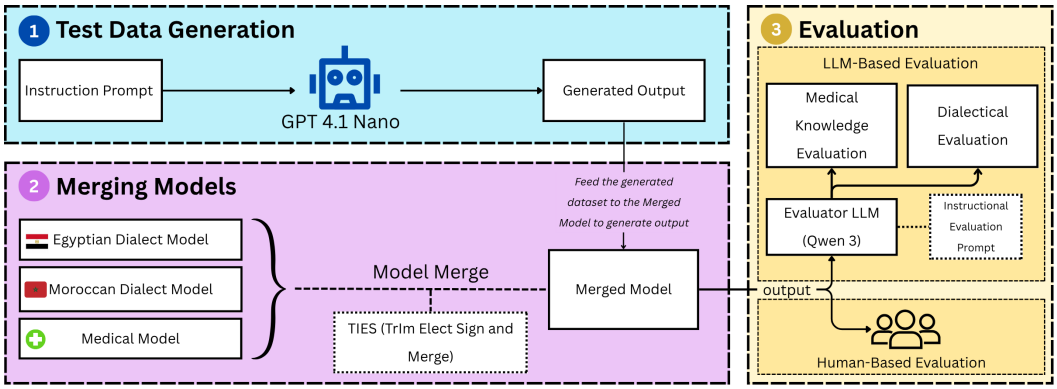


Figure 1: The framework consists of three stages: (1) Test data generation using GPT 4.1 Nano to produce dialect-specific medical symptom descriptions; (2) Model merging via the TIES algorithm, integrating Egyptian Arabic, Moroccan Darija, and medical domain LLMs into a unified model; and (3) Dual evaluation of the merged model through automated (LLM-based) and human-based assessments, focusing on both medical accuracy and dialectal comprehension.

To support this evaluation, we construct a dataset of patient symptom descriptions spanning Egyptian Arabic, Moroccan Darija, and MSA. Quantitative metrics are used to assess general model performance, while human evaluations—conducted by native speakers of the respective dialects—focus on practical utility.

## 2 Related Work

Our work intersects three key areas: dialectal Arabic NLP, medical language processing in Arabic, and model merging techniques for adapting large language models.

### 2.1 Arabic Dialects in NLP

The Arabic language landscape is characterized by diglossia, with MSA coexisting alongside over 30 regional dialects, such as Egyptian Arabic, Moroccan Darija, among others. These dialects differ substantially in phonology, lexicon, and syntax, often to the extent that they are mutually unintelligible (Kwaik et al., 2018; Al-Wer and de Jong, 2017; Salameh et al., 2018). This linguistic diversity presents a major obstacle for NLP systems, particularly in complex tasks such as intent classification and symptom extraction.

The challenge is especially acute in healthcare contexts, where patients frequently describe symptoms using their native dialects, which may be unfamiliar to clinicians. This misalignment can introduce significant communication barriers, leading to misunderstanding and clinical risk (Ellahham, 2021; Zhang et al., 2022).

These challenges highlight the urgent need for Arabic medical NLP resources that account for dialectal diversity, motivating a closer look at existing datasets and their limitations in supporting real-world clinical applications.

### 2.2 Medical NLP in Arabic

Medical NLP in Arabic remains underdeveloped compared to high-resource languages, primarily due to the scarcity of annotated clinical datasets—particularly those that capture dialectal variation. While most existing research focuses on MSA, real-world patient communication often occurs in regional dialects, reducing the practical effectiveness of MSA-centric models in clinical settings.

Recent initiatives have begun to address this gap. The Arabic Healthcare Dataset (AHD) (Al-Majmar et al., 2024), derived from Altibbi, provides a large-scale collection of question–answer pairs across diverse medical categories. However, dialect-rich medical corpora remain limited. Social media resources such as ArCOV-19 (Haouari et al., 2021) offer health-related content spanning multiple arab countries, but lack clinical precision. Other efforts include dialect-focused corpora like the Shami corpus for Levantine Arabic (Abu Kwaik et al., 2018), which support dialectal NLP tasks but are not tailored to the medical domain.

These limitations underscore the need for alternative approaches that are both resource-efficient and dialect-aware, motivating our exploration of model merging for scalable Arabic medical NLP.

### 2.3 Emergence of Model Merging Techniques

Recent advances in model merging have established it as a critical paradigm for consolidating specialized capabilities from multiple pre-trained models into a unified framework without additional training. This approach directly addresses scalability challenges in multilingual NLP by enabling efficient integration of domain-specific and dialect-specific expertise (Yang et al., 2024). Techniques such as Fisher-weighted averaging (Matena and Raffel, 2022) and TIES-Merging (Yadav et al., 2023) allow the integration of multiple pre-trained models—for example, dialect-specific experts and general-purpose medical LLMs—into a unified framework that retains their respective strengths. These approaches offer a scalable alternative to traditional fine-tuning pipelines, particularly in low-resource or fragmented language settings like Arabic.

While concrete numbers may vary by task and setup, these methods have repeatedly demonstrated efficiency gains—such as reducing compute and storage compared to training separate models—without compromising on performance. This makes them compelling for constructing single, robust Arabic medical LLMs that effectively handle multiple dialects and domains without expensive per-dialect pre-training and finetuning pipelines.

## 3 Methodology

### 3.1 Base Models

All models used in this study are based on the Gemma 2B architecture. We integrate three specialized variants representing complementary expertise in medical and dialectal domains:

- **Medical Domain Expert**: `OpenMeditron/Meditron3-Gemma2-2B` is a clinical language model co-developed with clinicians and humanitarian practitioners. It is trained with an emphasis on equitable representation, contextual diversity, and alignment with evidence-based medical guidelines—particularly for low-resource settings and underserved populations.
- **Egyptian Arabic Specialist**: A custom Gemma 2B model fine-tuned on the `MBZUAI-Paris/Egyptian-SFT-Mixture` dataset. This model was developed specifically to fill the gap in Egyptian dialect models based on the Gemma 2B architecture. The fine-tuning process focuses on capturing the phonological, syntactic, and lexical characteristics unique to Egyptian Arabic, which are not adequately represented in standard Arabic models.
- **Moroccan Darija Specialist**: `MBZUAI-Paris/Atlas-Chat-2B` is an instruction-tuned model designed for Moroccan Darija as part of the Jais project. It is optimized for a range of generative tasks including question answering, summarization, and translation. The model is designed to be lightweight and suitable for deployment in resource-constrained environments.

All three models share the same tokenizer and vocabulary inherited from the base Gemma 2B architecture. This architectural consistency ensured full vocabulary coverage across both dialectal variations and medical terminology, eliminating any risk of out-of-vocabulary degradation or tokenization mismatches during the merging process.

### 3.2 TIES-Based Model Merging

Our primary merging strategy follows the **TIES** (Trim, Elect Sign, and Merge) methodology (Yadav et al., 2023), a zero-shot model merging technique designed to mitigate task interference when combining multiple fine-tuned models. TIES creates a unified multitask model by aligning significant directional updates across task-specific models without requiring further training or access to original training data.

To implement this, we used MergeKit[1], an open-source framework that supports flexible model merging strategies, including TIES. MergeKit is a toolkit designed for assembling and merging large language models. It supports an extensive range of model architectures and implements numerous merging algorithms such as TIES, SLERP, task arithmetic, and Fisher-weighted averaging.

The process involves three key stages:

1. **Trim (Sparsification)**: For each task-specific model (e.g., dialect specialists), we compute a *task vector* as the parameter difference from a reference model, in our case the medical base:

$$\tau_i = \theta_{\text{dialect}_i} - \theta_{\text{med}}$$

[1]https://www.arcee.ai/product/mergekit

These task vectors are then sparsified by retaining only the top-$k$ parameters by magnitude (we use a density of 0.6, corresponding to $k = 20\%$) to emphasize impactful updates and reduce potential conflicts from noise or overfitting.

2. **Elect Sign**: Among the retained (nonzero) parameter updates, directional disagreements can still occur. In this step, TIES resolves sign conflicts by electing the consensus direction. A parameter's sign is retained only if at least $70\%$ of the models agree on the direction of the update, ensuring robustness across tasks.

3. **Merge**: Finally, the aggregated parameter updates are merged back into the base model. Only updates with elected signs contribute to the merged model, while trimmed or conflicted parameters default to zero. The final update rule is:

$$\theta_{\text{merged}} = \theta_{\text{med}} + \lambda \sum_i w_i \cdot \tau_i^{\text{sparse}}$$

where $\lambda$ is a global scaling factor and $w_i$ is the weight assigned to each model (set in our configuration as $w_i = 0.6$ for dialect models and $w_i = 0.4$ for the medical model).

```
models:
  - model: MBZUAI-Paris/Atlas-Chat-2B
    parameters:
      density: 0.6
      weight: 0.6
  - model: AITheChillGuy/Egyptian-Chat-2
    B
    parameters:
      density: 0.6
      weight: 0.6
  - model: OpenMeditron/Meditron3-Gemma2
    -2B
    parameters:
      density: 0.6
      weight: 0.4
merge_method: ties
base_model: google/gemma-2-2b-it
parameters:
  normalize: true
  int8_mask: true
dtype: float16
```

Figure 2: YAML configuration for TIES merging via MergeKit. Weights balance dialect specialization (0.6) against medical domain knowledge (0.4), with uniform density (0.6) for parameter sparsification.

### 3.3 Evaluation

**Test Dataset** To address the critical shortage of dialect-rich medical datasets, we generated a specialized evaluation set using `gpt-4.1-nano`. The generation process followed a structured system prompt (illustrated in Figure 3) designed to ensure clinical plausibility, dialectal accuracy, and consistency across Egyptian Arabic, Moroccan Darija, and MSA. An example of the generated test data is shown in Figure 4.

**Prompt Design Principles** The system prompt enforced four core generation constraints:

1. **Linguistic purity**: Strict separation between MSA and dialect outputs
2. **Clinical focus**: Symptom descriptions only
3. **Demographic Variation**: Differences in representation across age and gender groups.
4. **Tone control**: Neutral, descriptive patient narratives

**Dialectal Adaptation Protocol** For dialect generation, we modified the prompt's language specification while preserving clinical constraints:

- **Egyptian Arabic**: "Use authentic Egyptian colloquial Arabic"
- **Moroccan Darija**: "Use authentic Moroccan Darija expressions"
- Maintained identical content requirements across all variants

**Dataset Composition** The final corpus contains 900 clinically valid symptom descriptions:

- **MSA**: 300 examples
- **Egyptian Arabic**: 300 examples
- **Moroccan Darija**: 300 examples

**Metrics** To evaluate the quality and reliability of the merged model, we adopted a two-pronged evaluation framework combining LLM-based assessment and human judgment:

- **LLM-based Evaluation**: We used `Qwen 3 Base`—a strong Arabic-capable foundation model ranked highly on the Hugging Face Open LLM Leaderboard—to provide

```
        "role": "system",
        "content": (
         "You are an Arabic-speaking medical professional tasked with
generating realistic patient statements in Modern Standard Arabic (MSA) for
clinical training. Follow these rules:\n"
         "1. Language:\n"
         "•Use clear, simple Modern Standard Arabic (no dialects, no medical
jargon).\n\n"
         "2. Content:\n"
         "•Only describe symptoms, concerns, or contextual details.\n"
         "•Avoid direct questions (e.g., \"Is this serious?\", \"Should I get tested?
\", \"Should I go to a doctor?\").\n"
         "•Include:\n"
         "-Symptom details (location, duration, severity).\n"
         "-Triggers, alleviating factors, or family history.\n"
         "-Emotional/practical impact (e.g., anxiety, work disruption).\n\n"
         "3. Demographics: Vary scenarios (adults, children, elderly, pregnant
women).\n\n"
         "4. Tone: Neutral, descriptive, and natural — as if a patient is calmly
describing their condition to a physician without seeking advice."
        )
       },
       {
        "role": "user",
        "content": (
         "Generate one patient statement that describes symptoms and
concerns without asking if I should visit a doctor. "
         "Here is the previous statement for context: {}\n"
         "Follow these examples:\n"
         "•مثال ١: \"أشعر بألم في أسفل الظهر يمتد إلى الساق اليمنى منذ أسبوعين، ويزداد عند
الجلوس لفترات طويلة. لم تتحسن الحالة مع استخدام المسكنات العادية.\"\n"
         "•مثال ٢: \"ظهرت لي طفح جلدي أحمر على الذراع بعد استخدام نوع جديد من الصابون.
الحكة شديدة وتؤثر على نومي.\"\n"
         "•مثال ٣: \"أعاني من تعب مستمر منذ ثلاثة أشهر، رغم حصولي على نوم كافٍ، وأواجه
صعوبة في التركيز في العمل.\"\n"
```

Figure 3: Prompt for data generation. Identical content rules applied to all dialects with language specifications modified for MSA, Egyptian Arabic, and Moroccan Darija versions.

automated, dialect-sensitive evaluation. The model was prompted to rate responses along two axes:

- *Dialectal Fidelity* (1–5): Assesses the consistency, authenticity, and appropriate use of the target dialect in the generated response.
- *Medical Competence* (1–5): Evaluates the clinical accuracy, relevance, and appropriateness of the response.

For each dialect, 300 representative prompts were used. Scores were assigned based on predefined rubrics (see Figure 5 for the full prompt template).

- **Human Evaluation**: To assess the real-world quality of the merged model's outputs, we conducted evaluations with native speakers of Egyptian Arabic and Moroccan Darija. Using a set of 30 examples, each evaluator reviewed responses across two key dimensions:
  - *Dialectal Naturalness and Fidelity*: Raters judged how fluent, idiomatically accurate, and culturally authentic the responses were in their respective dialects.
  - *Overall Coherence and Appropriateness*: Evaluators assessed whether responses demonstrated general medical knowledge, internal coherence, and alignment with the input prompts.

## 4 Results

### 4.1 LLM-Based Evaluation

Table 1 reports average scores for LLM-based dialectal fidelity, while Table 2 presents average scores for LLM-based medical competence, both evaluated using the `Qwen 3 Base` evaluator over 300 prompts per dialect.

In both Table 1 and Table 2, rows represent the dialect of the test prompt, while columns correspond to the model being evaluated—MSA, Egyptian, Darija, and the Merged model. Each model was evaluated across all three dialects.

| Prompt | MSA | Egyptian | Darija | Merged |
|---|---|---|---|---|
| MSA | 4.78 | 4.64 | 3.34 | 4.89 |
| Egyptian Arabic | 2.14 | 4.40 | 1.96 | 3.91 |
| Moroccan Darija | 1.91 | 2.35 | 4.02 | 3.82 |

Table 1: Averaged LLM-based Dialectal Fidelity scores (1–5). Rows indicate the dialect of the test prompts; columns represent the model being evaluated.

| Prompt | MSA | Egyptian | Darija | Merged |
|---|---|---|---|---|
| MSA | 4.12 | 3.68 | 3.90 | 4.02 |
| Egyptian Arabic | 2.28 | 4.32 | 2.03 | 3.88 |
| Moroccan Darija | 1.83 | 1.77 | 4.05 | 3.85 |

Table 2: Averaged LLM-based Medical Competence scores (1–5). Rows indicate the dialect of the test prompts; columns represent the model being evaluated.

### 4.2 Human Evaluation

Table 3 presents the results of the human-based evaluation, in which native speakers rated the naturalness and coherence of generated outputs on a 5-point Likert scale. Evaluators assessed outputs across MSA, Egyptian Arabic, and Moroccan Darija.

أشعر بألم في الصدر يمتد إلى الظهر منذ خمسة أيام، ويبدأ بعد ممارسة الأنشطة البدنية الخفيفة.
الألم يزداد مع التنفس العميق ويخف عندما أستريح. لم ألاحظ وجود ضيق في التنفس أو دوار، ولكن الأعراض تؤثر على قدرتي على ممارسة يومياتي بشكل طبيعي.

أنا بقالي حوالي أسبوع، حاسس بوجع بسيط في نص ظهري من الناحية اليسرى، وبيختفي وبيجي، وأوقات بيكون خفيف وأوقات بزيد. الوجع ده بيزيد لما بقوم من مكان طويل أو بعد ما أكون قاعد لوقت طويل، وكمان بحس كأن في ضغط أو ثقل جوا الظهر والكتف، وده بيعمل لي شوية توتر وقلق على صحتي، خاصة إني دايمًا عندي مشكلة في السكر وضغط الدم، فده بيخلي بالي يتوتر أكتر لما بحس بالحاجات دي.

أنا حاس براسي من تحت ف رقبتها، وهاد الألم كيجي ويختفي على مدار النهار. أوقات كيكون حاد، وكيعطل لي حركة ديال راسي، بحال إلى ممنوع نمدها أو نملها، وهاد الشي كيعمل لي إزعاج كبير. وزيد على ذلك، كحس أوقات بتنمل ف يدي، خصوصًا فاليسار، وهاد الشي كيخلي يدي تتعب بسرعة وما نقدرش نخدم مزيان. ديما كنقلق من هاد الحالة إذا زادت وتطورت، وكنحس براسي كنعاني من تعب فالجسم، وتركيزي حتى هو كيضعف.

Figure 4: Examples of generated test data in MSA (left), Egyptian Arabic (middle), and Moroccan Darija (right)

```
You are tasked with evaluating the quality of a language model's response to a simulated patient statement.
The evaluation has two major objectives:

1. **Dialectical Fidelity**: Does the model generate responses in the requested Arabic dialect with consistency and authenticity?
2. **Medical Competence**: Does the model demonstrate medically sound reasoning and an appropriate level of expertise?

**Requested Dialect**: {dialect}
**Patient Statement**: {question}
**Model Response**: {answer}

Please evaluate the response based on the following criteria:

**Dialectical Fidelity** (Score: 1-5)
- 5: Consistently uses the requested dialect; rich, natural, idiomatic usage
- 4: Mostly uses the correct dialect with minor inconsistencies
- 3: Noticeable mixing of dialects but still understandable
- 2: Major dialectal mismatches or frequent intrusion of unrelated forms
- 1: Ignores the requested dialect entirely

**Medical Competence** (Score: 1-5)
- 5: Accurate, contextually rich, medically plausible response
- 4: Minor gaps or simplifications, but overall medically reasonable
- 3: Several issues in reasoning, missing details
- 2: Contains inaccurate or misleading content
- 1: Contains factual errors or harmful advice

Please provide a score for each criterion and a brief justification.
```

Figure 5: Evaluation prompt used for LLM-based scoring. Qwen 3 Base was instructed to assess each model response for dialectical fidelity and medical competence using a consistent rubric across MSA, Egyptian Arabic, and Moroccan Darija.

| | MSA | Egyptian | Moroccan |
|---|---|---|---|
| **Quality (1-5)** | 4.91 | 4.87 | 4.20 |

Table 3: Averaged Human evaluation (naturalness & coherence).

### 4.3 Example Model Outputs

To qualitatively assess the merged model's performance, we present sample outputs when the prompt is in one dialect and the response is expected in a different dialect (see Figure 6).

## 5 Discussion

The results of this study demonstrate the significant potential of model merging as a scalable solution for dialect-rich NLP applications in specialized domains. By integrating dialectal specialists (Egyptian Arabic, Moroccan Darija) with a medical domain model using the TIES-merging technique, we produced a unified model capable of handling cross-dialect medical communication with minimal computational overhead. This directly addresses the limitations of conventional fine-tuning approaches, which are impractical given the extent of linguistic fragmentation in Arabic.

Our LLM-based evaluation revealed that the merged model maintained robust performance across dialect boundaries, achieving dialectal fidelity scores between 3.82 and 4.89, and medical competence scores between 3.85 and 4.02 across test scenarios. Notably, the model could accurately interpret Moroccan Darija symptom descriptions for Egyptian Arabic-trained clinicians—a critical cross-dialect use case where conventional models often fail. This capability directly mitigates real-world communication barriers in healthcare settings where mutual dialect intelligibility is limited.

The resource efficiency of our approach is also noteworthy. The TIES-merging process completed in about 10 minutes on a single L4 GPU, using 9.3 GB of memory, and reduced storage requirements by 67% compared to maintaining separate specialized models. This lightweight computational profile makes dialect-sensitive medical NLP feasible in resource-constrained environments, where per-dialect fine-tuning would be prohibitively expensive. Such efficiency could democratize access to specialized NLP technologies in under-resourced regions.

Human evaluations further validated the real-world applicability of the merged model. Native speakers rated the model's naturalness and coherence at $\mu = 4.87$ for Egyptian Arabic and $\mu = 4.20$ for Moroccan Darija, confirming authentic linguis-


Figure 6: Example showing the input in green and output generated from the merged model. (A) Question in MSA and answer in Egyptian Arabic, (B) Question in Egyptian Arabic and answer in Moroccan Darija, and (C) Question in Moroccan Darija and answer in MSA.

tic adaptation. This underscores the model's ability to retain medical knowledge while fluently adapting to diverse dialects—supporting the notion that linguistic form and domain content can be effectively disentangled in the merging process, consistent with findings from recent parameter-efficient multitask learning literature. Subject matter experts observed that the model preserves natural phrasing and medical accuracy within the dialectal context. It also successfully interprets input in one dialect and reformulates the medical explanation in the target dialect.

### 5.1 Practical Implications

This work supports three key advancements for Arabic NLP in healthcare: First, it enables the deployment of a *single* unified medical NLP system that can serve diverse Arabic-speaking populations without maintaining multiple dialect-specific models. Second, model merging simplifies system updates—new dialects can be incorporated by merging in additional specialist models without retraining the full architecture. Third, this methodology offers a template for extending scalable model merging to other fragmented domains such as legal or educational NLP, where specialized dialect handling is equally critical.

## 6 Conclusion

This study establishes model merging as a viable paradigm for overcoming Arabic's dialectal fragmentation in high-stakes healthcare NLP. By consolidating specialized capabilities into a unified and resource-efficient model, we bridge critical communication gaps while substantially reducing computational demands. As Arabic NLP continues to evolve, such scalable approaches will be essential for enabling equitable and inclusive language technology across the linguistically diverse Arab world.

## 7 Limitations

Despite promising results, our dialectal coverage is limited to Egyptian Arabic and Moroccan Darija; incorporating additional varieties such as Levantine or Gulf Arabic would offer a more comprehensive test of the approach's scalability. While test data generation helped address data scarcity, real-world patient utterances are likely to exhibit greater variability and noise than those present in our controlled corpus. Additionally, our evaluation primarily focused on clinician-facing comprehension. Future work should explore patient-facing generation tasks, such as producing dialect-specific

medical advice, to better understand the model's bidirectional utility.

## References


Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A corpus of Levantine Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nashwan Ahmed Al-Majmar, Hezam Gawbah, and Akram Alsubari. 2024. Ahd: Arabic healthcare dataset. *Data in Brief*, 56:110855.

Enam Al-Wer and Rudolf de Jong. 2017. Dialects of arabic. *The handbook of dialectology*, pages 523–534.

Ashwag Alasmari. 2025. A scoping review of arabic natural language processing for mental health. *Healthcare*, 13(9).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Greg Brunet, Marsha Chechik, Steve Easterbrook, Shiva Nejati, Nan Niu, and Mehrdad Sabetzadeh. 2006. A manifesto for model merging. In *Proceedings of the 2006 international workshop on Global integrated model management*, pages 5–12.

Samer Ellahham. 2021. Communication in health care: Impact of language and accent on health care safety, quality, and patient experience. *American journal of medical quality : the official journal of the American College of Medical Quality*, Publish Ahead of Print.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021. ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 82–91, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Ahmed Ibrahim, Abdullah Hosseini, Salma Ibrahim, Aamenah Sattar, and Ahmed Serag. 2025a. D3: A small language model for drug-drug interaction prediction and comparison with large language models. *Machine Learning with Applications*, 20:100658.

Ahmed Ibrahim, Abdullah Khalili, Maryam Arabi, Aamenah Sattar, Abdullah Hosseini, and Ahmed Serag. 2025b. Mera: Medical electronic records assistant. *Machine Learning and Knowledge Extraction*, 7(3).

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1708–1719, Dublin, Ireland. Association for Computational Linguistics.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. A lexical distance study of arabic dialects. *Procedia computer science*, 142:2–13.

Michael Matena and Colin Raffel. 2022. Merging models with fisher-weighted averaging. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the 27th international conference on computational linguistics*, pages 1332–1344.

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China. Association for Computational Linguistics.

Emma Trentman and Sonia Shiri. 2020. The mutual intelligibility of arabic dialects: Implications for the language classroom. *Critical Multilingualism Studies*, 8(1):104–134.

Xiao-Kun Wu, Min Chen, Wanyi Li, Rui Wang, Limeng Lu, Jia Liu, Kai Hwang, Yixue Hao, Yanru Pan, Qingguo Meng, and 1 others. 2025. Llm fine-tuning: Concepts, opportunities, and challenges. *Big Data and Cognitive Computing*, 9(4):87.

Zhengqi Xu, Ke Yuan, Huiqiong Wang, Yong Wang, Mingli Song, and Jie Song. 2024. Training-free pretrained model merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5915–5925.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Ties-merging: resolving interference when merging models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *Preprint*, arXiv:2408.07666.

Dangui Zhang, Zichun Jiang, Yu Xie, Weiming Wu, Yixuan Zhao, Anqi Huang, Tumei Li, and William Ba-Thein. 2022. Linguistic barriers and healthcare in china: Chaoshan vs. mandarin. *BMC Health Services Research*, 22(1):376.