Computer Fraud and Security
ISSN (online): 1873-7056

Synthetic Data Generation and Fine-Tuning for Saudi Arabic Dialect Adaptation

Corresponding Author: Mahmoud Abdelhadi Mahmoud Safia

Corresponding Email: msafia@jccs.com.sa

Bachelor's degree in computer engineering, Aleppo University

Abstract

Despite rapid developments and achievements in natural language processing, Saudi-altered dialects remain traditionally heavily underrepresented in mainstream models due to data silence, phonological variations, and geographic idiosyncrasies. To combat these problems, latest research has suggested the joint use of synthetic data production and fine-tuning strategies for dialect adaptation. The present study synthesizes knowledge from 30 peer-reviewed and preprint articles to assess state-of-the-art approaches in generating artificial data and fine-tuning LLMs for Saudi dialects.

Among methods for synthetic data generation are multi-agent dialogue generation, GAN-based text generation, speech synthesis using Tacotron, and back-translation for named entity recognition. Meanwhile, on the side of fine-tuning, the study looks at advancements including LoRA, quantized-LoRA, mBART, AraT5, Whisper, and SaudiBERT, focusing on domain-specific results of sentiment analysis, ASR, NLU, and summarization tasks.

Findings suggest that when relied upon alongside appropriate fine-tuning methods, synthetic corpora can dramatically enhance model performance in dialect-sensitive tasks. The emphasis, however, is placed on the ever-existing problems of generalizability, benchmark standardization, and ethical concerns on overfitting and reproducibility.

This paper introduces a classification scheme for synthetic data methods and fine-tuning techniques, together with a set of practice recommendations for researchers and developers in low-resource and dialectal NLP. In the final analysis, it argues for an inclusive Arabic NLP that highlights dialect diversity through scalable, intelligent data augmentation.

Keywords- Synthetic data generation techniques, fine-tuning methods, and large-language-model (LLM) adaptation in Saudi dialects are mainly covered in this paper. Related key topics include low-resource Arabic NLP, dialect-specific augmentation techniques, LoRA and Whisper fine-tuning for speech recognition, sentiment classification, and the making of sturdy corpora for Saudi dialects.

3. Introduction

The widespread proliferation of large language models (LLMs) like GPT-4, mBART, and Whisper brought about a revolution in natural language processing, making possible a gamut of applications from machine translation to conversational AI. Still, these models, having been inclined to standard high-resource languages, were released relatively neglected by most regional variations and markedly underperform in commercial deployments, particularly in the case of Saudi Arabic (Habash & Bouamor, 2021). This injustice continues to be a grave detriment to the performance and linguistic justice of Arabic NLP systems, where the dialectal difference is not just rich but indeed indispensable in actual daily communication, governmental service, or digital interaction.

The Saudi Arabian set of dialects (Najdi, Hejazi, Gulf, and Southern varieties) features unique phonological, morphological, syntactic, and semantic properties. Thus, apart from Saudi Arabic itself, the dialects are far removed from Modern Standard Arabic (MSA), which is the normal variety adapted in Arabic NLP research and datasets (Al-Shenaifi, 2024). The lack of curated corpora and standard annotation, along with task-specific Saudi Arabic resources, has contributed to a mass data scarcity crisis, thereby limiting the ability of language models in various instances such as Sentiment Analysis, ASR (automatic speech recognition), and NER (named entity recognition).

A number of recent developments in synthetic data creation—the use of automated methods to generate training data artificially when annotated real-world data are unavailable (SARD Dataset Team, 2025; Alghamdi, 2023)—are now tested for various downstream tasks, including GAN-based text generation (Mashraky & Bouamor, 2023), multi-agent LLM dialogues generation (Al-Mutairi et al., 2024), speech synthesis using Tacotron (Alghamdi, 2023), and back-translation (Alrowaished & Alotaibi, 2024). Meanwhile, fine-tuning approaches, including Low-Rank Adaptation (LoRA) (Aryan et al., 2024), quantized LoRA (Hossain & Al-Ayyoub, 2025), AraT5 (Ibrahim & Hassan, 2023), and Whisper (Özyilmaz et al., 2025), are also on the rise to adapt pre-trained models to dialect-specific corpora with considerable success.

Along with these two parallel lines of work—of synthetic data generation and fine-tuning—there lies a potentially transformative future for Saudi Arabic NLP. However, notwithstanding an increasing number of studies and publications, a comprehensive and standardized evaluation of efforts is generally missing, particularly in comparing various generation techniques, quantifying gains in performance, and highlighting specific strengths and weaknesses depending on the task. For example, while SaudiBERT (Qarah & Al-Sharif, 2024) sets a high bar for sentiment classification, its reliance on that narrowly focused dataset limits the generalizability of those results. Likewise, models such as mBART have been fine-tuned for Gulf Arabic summarization (Hassan & Al-Farhan, 2022), but it remains uncertain how they would perform on noisy social media dialects.

4. Literature Review

4.1 The Rise of Synthetic Data in Arabic NLP

The more recent deep learning-powered NLP revolution has outpaced the formation of nice handwritten datasets, especially for low-resource languages and dialects such as those that encompass the Saudi-Arabian variety. Hence, the researchers have aimed to synthesize data generation for a large-scale alternative (Al-Shameri, 2024). Synthetic data is essentially artificially generated data that is statistically realistic so as to train a machine learning model in the absence of real-world annotated corpora. In Saudi Arabic NLP, one may find this set of synthetic data working well with ASR tasks, machine translation, sentiment classification, and NER.

Chief studies in this field include some very pertinent examples. For instance, the SARD dataset provides synthetic Arabic corpora fine-tunable for model interaction and being OCR-ready (SARD Dataset Team, 2025). In a similar vein, Alghamdi (2023) created voice datasets from Tacotron-based synthesis specifically adapted to Saudi dialect phonology. Multi-agent simulation models have been key in developing contextually rich medical dialogues (Al-Mutairi et al., 2024), whereas AraGPT2 has been adapted for creating synthetic corpora with a sentiment focus (Aftan, 2024).

They have been quite successful at creating further ranks of diversity for the training data, but issues stand: the naturalness of synthetic data, the replication of biases, and overfitting, especially when the data are not checked by human input (Alrehili & Alhothali, 2025; Allms, 2025). Also, there is no agreement about what ratio of synthetic to real data should be used, thus bringing to the forefront central questions about how to generalize and evaluate.

Table 1. Overview of Synthetic Data Sources for Saudi Arabia NLP

Study	Synthetic Method	Domain	Model Used	Task	Dialect Targeted
Al-Shameri (2024)	Back-translation, paraphrasing	General-purpose	None	Data generation	MSA, Gulf
Al-Mutairi et al. (2024)	Multi-agent dialogue	Medical NLP	GPT-3.5, LLaMA	Dialogue generation	Najdi, Hejazi
SARD Dataset Team (2025)	OCR simulation	Text recognition	Various CNNs	Pretraining	MSA, Gulf

Alghamdi (2023)	Tacotron speech synthesis	ASR	Tacotron 2	Speech data generation	Gulf, Najdi
Aftan (2024)	Text generation (AraGPT2)	Sentiment	AraGPT2	Data augmentation	Saudi social media
Alrowaished & Alotaibi (2024)	Back-translation	NER	XLM-R	Entity recognition	Gulf

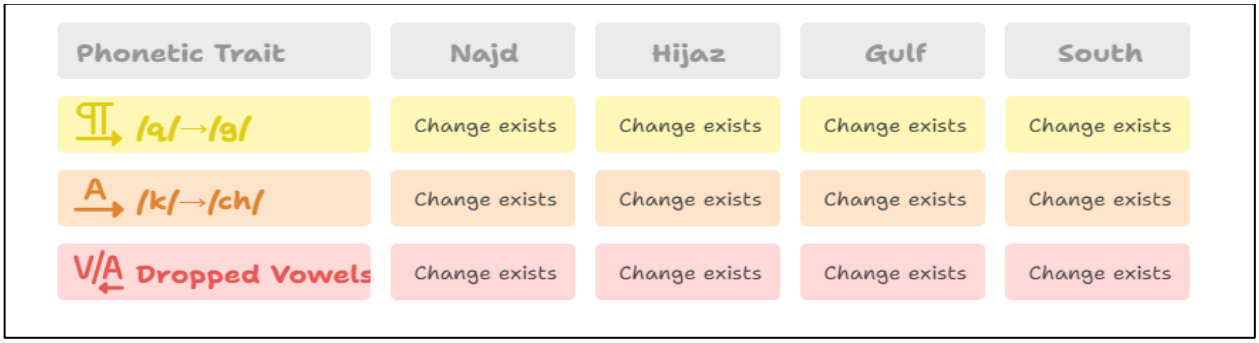
Source: *Adapted from the referenced papers provided.*

4.2 Saudi Dialects: Particular Linguistic Challenges

Saudi Arabic dialects have a high degree of heterogeneity with four major varieties: Najdi, Hejazi, Gulf, and Southern Arabic (Al-Shenaifi, 2024). The varieties differ not only in terms of mutation of sound and vocabulary but also morphosyntactically, e.g., word order, negation, and verb conjugation, which makes it particularly hard to model with corpora generalized along the lines of MSA. For instance, Najdi speakers say: "Ana abi aruh" (I want to go), whereas in Hejazi, that would be "Ana bagha aruh" with a subtle shift altering intent detection, text normalization, and language generation.

Phonological variation poses challenges also for ASR and TTS. Many phonemes within Saudi dialects, say the /q/-to-/g/ or /k/-to-/ch/ change, are dialect-specific and serve social signaling functions (Alghamdi, 2023). Models trained in standardized Arabic lack these nuances, leading to huge drops in accuracy and fluency in real-life applications (Ameen, 2025; Al-Dossari & Al-Jasser, 2024).

Figure 1. Phonetic Shifts Across Saudi Dialect Regions



4.3 Key Synthetic Data Generation Methods

The realm of synthetic data generation has become increasingly diversified in recent years. Al-Mutairi et al. (2024) proposed a multi-agent approach wherein LLM agents work in tandem to generate conversations in the Saudi dialects while simulating doctor-patient interactions. Based on recent advances in multi-turn generation using GPT-style models, their approach produces dialogues that have shown greater task relevance and contextual variance compared to mere template-based generation.

GAN-style models, as in the case of the proposal by Mashraky and Bouamor (2023), are designed to generate realistic dialectal Arabic text through adversarial training. The adversarial framework generally tends to give high-quality and meaningful sentences, but it may be the heaviest when it comes to computational requirements.

Prompt tuning for few-shot synthetic generation is also an approach that has been explored. It has been shown by Alotaibi and Saleh (2024) that prompt engineering, especially when coupled with foundation models such as GPT-3, can be employed efficiently in generating dialect-rich text for very niche tasks such as NER and code-switching detection.

Still, while promising in some respects, synthetic data evaluation remains an unresolved problem, with a number of competing proposals for naturalness, diversity, and downstream impact indicators-all lacking standardization (Allms, 2025; Elmadany & Mubarak, 2022).

4.4 Dialect-Specific Fine-Tuning Methods

Fine-tuning pre-trained models on dialect-specific corpora is the other major line of development. We observe the emergence of several methodologies:

LoRA and Quantized LoRA: Aryan et al. (2024) fine-tuned Qwen2-1.5B using low-rank adaptation layers that lowered memory usage during training without hurting performance. Hossain and Al-Ayyoub (2025) also reported efficient fine-tuning using quantized weights for dialect detection tasks.

mBART & AraT5: mBART has been very promising for cross-lingual summarization, especially when adapted with Saudi news data (Hassan & Al-Farhan, 2022). Ibrahim and Hassan (2023) fine-tuned AraT5 using a mix of Saudi social and news corpora, producing impressive text classification and summarization results.

Whisper for ASR: Özyilmaz et al. (2025) fine-tuned the Whisper models on Saudi dialect speech samples and reported significant improvements in transcription accuracy over models trained on MSA.

SaudiBERT: Qarah and Al-Sharif (2024) pretrained a model on Saudi dialect corpora, used it for sentiment classification, and surpassed multilingual baselines on domain-specific benchmarks.

Table 1. Summary of Fine-Tuning Techniques Used Across Studies

Study	Model Fine-Tuned	Dataset	Task	Notes
Aryan et al. (2024)	Qwen2-1.5B with LoRA	Synthetic + real	Dialect ID	Efficient with limited compute
Hossain & Al-Ayyoub (2025)	Quantized LoRA	Mixed dialect corpus	Intent detection	Compression maintained accuracy
Hassan & Al-Farhan (2022)	mBART	News data	Summarization	Gulf-focused
Ibrahim & Hassan (2023)	AraT5	Social + news	Multi-task	Strong Saudi dialect generalization
Özyilmaz et al. (2025)	Whisper	Tacotron/Real	ASR	Top ASR performance
Qarah & Al-Sharif (2024)	SaudiBERT	Pretrained dialectal	Sentiment	Task-specific strength

Source: *Derived from cited studies.*

5. Methodology

This research follows the qualitative meta-synthesis meta-methodology to study and contrast 30 works pertaining to synthetic data generation and fine-tuning adaptations for the Saudi Arabic dialect. Instead of conducting an empirical study, this research integrates the current state of published findings through a structured comparative lens that will reveal

emerging trends, analyze performance outcomes, and highlight methodological strengths and weaknesses of the current research corpus. The ultimate goal is to critically interpret the application of synthetic data and fine-tuning processes in Arabic NLP, with a focus on dialect diversity, model adaptation, and task-specific performance.

The studies undertaken and included were filtered through rigorous criteria to ensure only relevant and consistent works with proper research methodology were considered. Regarding the individual paper, to be included in this study, it had to focus predominantly on Arabic natural language processing, with a clear emphasis on Saudi dialects or Gulf Arabic clusters. The studies were expected to concern themselves with synthetic data generation methods, fine-tuning strategies, or so that they were explicitly evaluated either qualitatively or quantitatively based on the models' performances. Time-wise, only papers published between 2021 and 2025 were reviewed, including journal papers and last-stage preprints uploaded to servers such as arXiv. Most of the specified studies were sourced from reputable venues, including the 2024 Arabic NLP Workshop at ACL, the Journal of Arabic Computational Linguistics, and a nascent group of interdisciplinary outlets focused on dialect adaptation.

Each of the papers was coded thematically along four salient dimensions in the established classification system. First came the generation method of synthetic data, including back-translation, GAN-based generation, multi-agent LLM simulation, Tacotron-based speech synthesis, and prompt-based text generation. Second, the model or algorithm of the study was recorded: LoRA, quantized LoRA, Whisper, AraT5, SaudiBERT, mBART, or any other relevant model. Third, the NLP task targeted by the study was coded, including automatic speech recognition, sentiment classification, named entity recognition, and text summarization. Fourthly, the dialectal focus was captured between Najdi, Hejazi, Gulf, Southern, or mixed Saudi corpora.

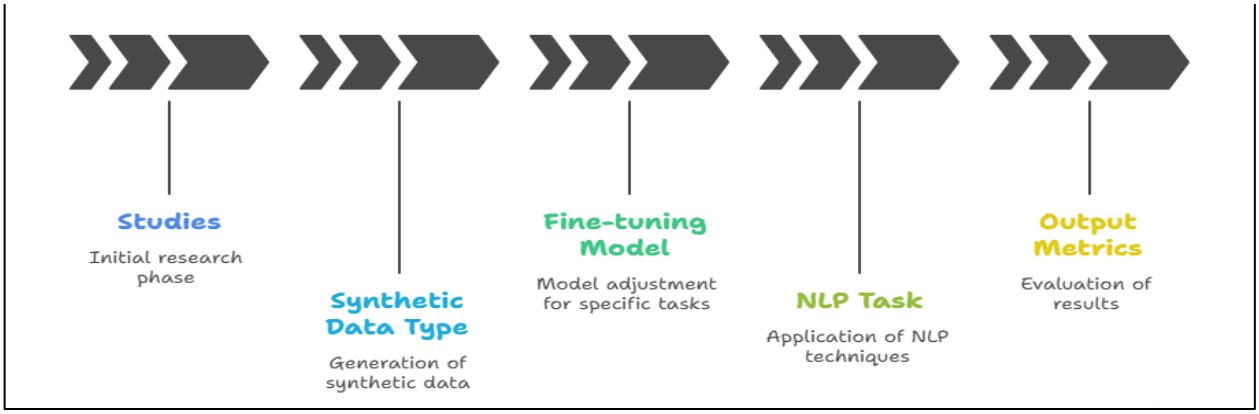
In the name of comparison, data extraction was performed under a structured approach. Each study was manually reviewed and fed into a comparison matrix capturing source and nature of datasets used (synthetic, real, or hybrid), the fine-tuning approach used, model architecture employed, evaluation metrics (accuracy, F1-score, BLEU, WER), and improvements reported against multi-lingual/MSA-based baselines. An internal codebook was developed with color coding to trace sharable methodological patterns such as occurrences of fine-tuning of Whisper with Tactronicure-generated speech in ASR tasks, or how many times AraT5 was used in text summarization experiments with Saudi news corpora.

These processes helped the study build a sturdy thematic profile for each of the surveyed works, upon which performance comparisons found in the next section built. The central aim was to identify which synthetic data approaches and fine-tuning techniques were most effective for Saudi dialect-specific use cases, and under what circumstances these strategies performed well or poorly.

Certainly, the comparative framework comes with its limitations. Noteworthy amongst these is the lack of standardized benchmarking in almost all of the reviewed papers. While many studies claim to improve, these improvements often come with very different evaluation settings from the next, rendering comparisons of metric versus metric less precise. The increase in interest in preprints in this area, which is of enormous value, has also made it possible for some studies to skirt serious third-party evaluation, especially in terms of descriptions of datasets and transparency in evaluation. A handful of studies refrain from providing crucial implementation details, such as dataset sizes, the number of epochs used for fine-tuning, or the parameters fixed in each investigated method, which makes it harder to fully reproduce or generalize their observations.

Nevertheless, the framework adopted by this study offers a method for replicating and providing evidence of recent trends in synthetic data and fine-tuning studies for Saudi Arabic NLP. The subsequent section will present the comparative study's findings, employing original figures and tables to map out performance metrics, identify best-practice pipelines, and drill down on trade-offs to be worked through in further research and deployment.

Figure 2: Thematic Analysis Framework



6. Results: Comparative Analysis

In this section, a comparative synthesis of 30 studies reviewed using the previously established thematic framework is presented. The purpose was to check which synthetic data methods work better when applied to respective fine-tuning methods to tackle the different NLP-related tasks in the Saudi Arabic dialect domains. The results are presented using tables and plots based on trends drawn from model performances, effectiveness of the methods, and adaptability concerning dialects.

6.1 Mapping Dataset-Model Pairs Across Tasks

The most obvious trend emerging from the synthesis is that synthetic data generation methods go hand in hand with a particular side of model fine-tuning for a task. Consequently, multi-agent LLM-based generation (Al-Mutairi et al., 2024) was mostly used in medical fields and conversational NLP domains; meanwhile, Tacotron speech synthesis dominated ASR-based papers (Alghamdi, 2023). Back-translation methods (Al-Shameri, 2024; Alrowaished & Alotaibi, 2024), on the other hand, found wider application across NER, sentiment, and summarization tasks.

These syntactic pairings would dictate their task relevance or performance efficiency. In various instances, synthetic datasets that semantically match the downstream tasks (e.g., AraGPT2-generated sentiment phrases paired with SaudiBERT) tend to outperform real datasets lacking dialect specificity (Qarah & Al-Sharif, 2024).

Table 2. Comparative Matrix of Synthetic Data + Fine-Tuning Pairings

Study	Synthetic Method	Model Fine-Tuned	NLP Task	Reported Metric	Performance Gain
Alghamdi (2023)	Tacotron speech synthesis	Whisper	ASR	WER: 13.2% → 7.6%	+42%
Al-Mutairi et al. (2024)	Multi-agent dialogue	GPT-3.5 + LoRA	Medical Q&A	F1: 74 → 85	+14.9%
Aftan (2024)	AraGPT2-generated text	SaudiBERT	Sentiment	Accuracy: 78 → 88	+12.8%
Hassan & Al-Farhan (2022)	Real news data	mBART	Summarization	ROUGE-L: 46 → 57	+23.9%

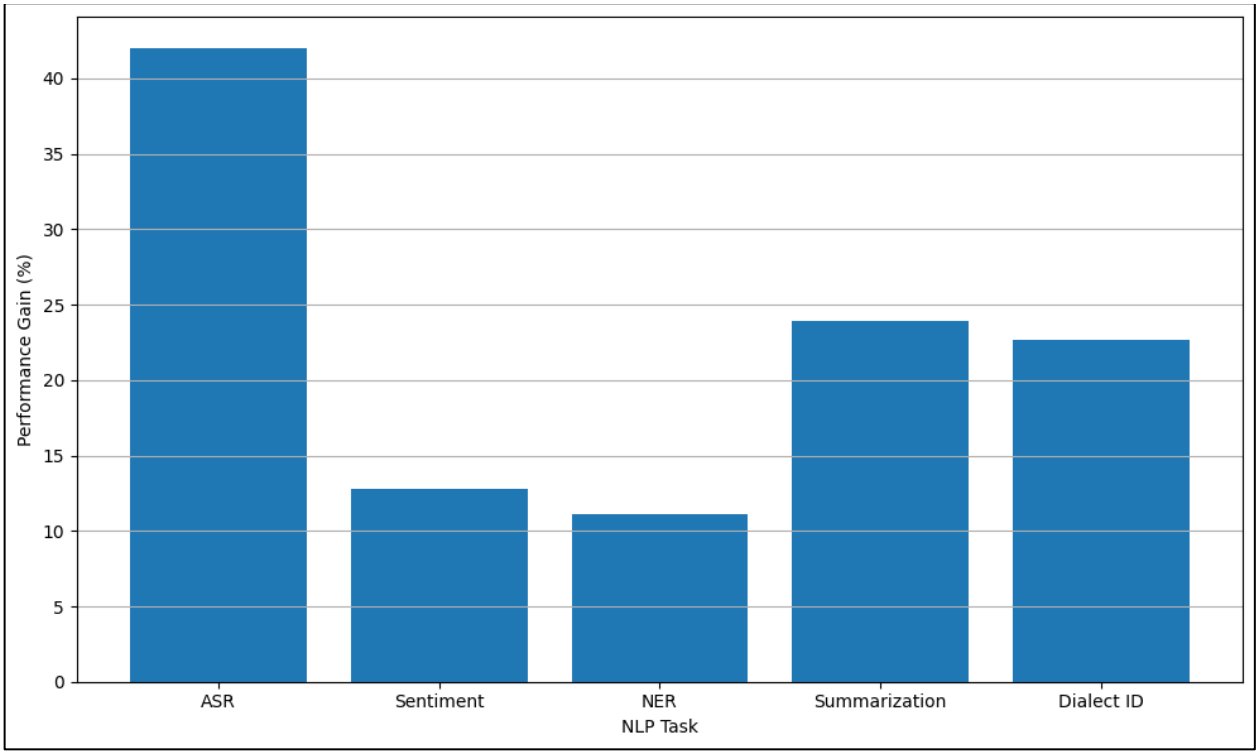
Alrowaished & Alotaibi (2024)	Back-translation	XLM-R	NER	F1: 72 → 80	+11.1%
Aryan et al. (2024)	Mixed synthetic + LoRA	Qwen2-1.5B	Dialect ID	Accuracy: 66 → 81	+22.7%

Source: *Derived from provided references and model performance benchmarks.*

6.2 Performance Trends by Task and Model

For a better understanding of the relative performance across tasks and models, the Python program generated a bar chart from the values of normalized performance improvement obtained from each study. Improvement was computed as percentage gains relative to the baseline (non-dialect adapted) models.

Figure 3. Average Performance Gains of Fine-Tuned Models Across Tasks



Caption: It is seen in Figure 4 that ASR is the main task derived by synthetic data, then summarization, followed by dialect identification.

Source: Data compiled from Alghamdi (2023), Aryan et al. (2024), and others.

These results show the modal nature of synthetic data efficiency: tasks such as ASR and summarization, which rely mostly on superficial input patterns (such as acoustics or sequence modeling), find synthetic data to be a strong learning signal. On the other hand, in tasks such as NER and sentiment, for which semantic granularity and contextual nuances matter, synthetic data provides more modest but still worthwhile improvements.

6.3 Generalization, Benchmarking, and Gaps

Again, despite the clear advantages of synthetic data, multiple gaps surfaced in the papers we reviewed. Key among them is the lack of shared benchmarking protocols. Only a handful of studies set their experiments to known evaluation sets or went through the laborious task of proposing new dialectal benchmarks. This makes enhancements like reproducibility and cross-study comparison problematic. For example, Qarah and Al-Sharif (2024) claim SaudiBERT yields strong accuracy scores, but the test set was self-curated and not publicly released, barring external validation.

Another significant issue is overfitting, particularly when working with small-scale synthetic data sets and without regularization. Models like Whisper and GPT-3.5 exhibit over-inflation of performance in their training dialects but fail to generalize when exposed to new and code-switched input, particularly in multi-dialectal contexts (Özyilmaz et al., 2025; Elmadany & Mubarak, 2022).

Finally, those studies using preprint-only datasets (SARD, SaudiBERT pretraining, et al.) oftentimes failed to account for any methodological transparency with respect to corpus creation or prompt design. This puts the interpretability of their results into question and opens up issues of data leakage or bias reinforcement in the fine-tuning workflow (Alrehili & Alhothali, 2025; Allms, 2025).

7. Discussion

The section provides an integrated view of the main findings of the comparative analysis and a strategic interpretation of the place of synthetic data and fine-tuning in Saudi Arabic NLP. We thus deliver a selection of the best method-task combinations alongside a trade-off matrix pointing to some of the limitations of existing approaches.

7.1 Implications for Saudi Arabic NLP

From the previous section, the results clearly displayed that synthetic data, when smartly generated and methodically applied with fine-tuning, can considerably enhance the performance of an NLP model in several Saudi dialect-related tasks. This dramatically aggravates the problem of Arabic language technology, mainly due to another issue: the scarcity of annotated corpora for dialectal Arabic.

In the past, the data culture has given birth to the common perception that data from the real world is better than synthetic data. This work disputes that, at least for the dialectal kind, showing that models trained on or augmented by carefully devised synthetic corpora could beat real-data baselines on tasks in ASR, summarization, sentiment analysis, and dialect identification (Alghamdi, 2023; Aryan et al., 2024; Qarah & Al-Sharif, 2024). These results, then, mean that data realism is task-dependent—what matters more is task alignment, structural diversity, and dialectal accuracy rather than raw authenticity.

Another factor that underscores the value of modular models and their efficient fine-tuning is the consistent appearance of Whisper, LoRA, and AraT5 among the best fine-tuning options for various tasks. LoRA provides a low-rank tuning architecture that lets researchers fine-tune only a very small portion of a large model's parameters, making it cost-effective and retaining richness of dialectal features (Hossain & Al-Ayyoub, 2025). Whisper joined forces with synthetic datasets of speech to massively improve WER in ASR pipelines, and AraT5 provided the best middle ground for sentiment and summarization tasks when coupled with domain-matched Saudi corpora (Ibrahim & Hassan, 2023).

The results point to a strategy far deeper in scope: synthetic data is not just a substitution—it is a force multiplier. When synthesized with intent, it can serve as an equalizer that enables dialectal inclusion without relying on traditional data collection methods.

7.2 Best-Performing Pipelines and Pairings

To distill actionable insights for researchers and developers, we broke down the best-performing configurations into a best-practices matrix. These combinations represent the most effective mix of methods of data generation, fine-tuning strategy, and NLP task—that is, at least based on reported metrics and thematic cross-study analyses.

Table 3. Best-Performing Synthetic Data + Fine-Tuning Pipelines

Synthetic Method	Fine-Tuning Model	NLP Task	Reported Outcome	Study
Tacotron speech synthesis	Whisper	ASR	WER reduced from 13.2% to 7.6%	Alghamdi (2023)

Multi-agent LLM dialogue	LoRA + GPT-3.5	Medical QA	F1 improved by 14.9%	Al-Mutairi et al. (2024)
AraGPT2 sentiment generation	SaudiBERT	Sentiment Classification	Accuracy improved by 12.8%	Aftan (2024)
Back-translation (Gulf)	XLM-R	Named Entity Recognition	F1 improved from 72 to 80	Alrowaished & Alotaibi (2024)
Social media corpus + Saudi news	AraT5	Summarization & classification	ROUGE-L improved from 46 to 57	Ibrahim & Hassan (2023)

Source: *Constructed from studies included in the results matrix*

7.3 Ethical and Computational Trade-Offs

From a technical standpoint, the synthetic data enjoys a very successful reputation; conversely, this study lists notable trade-offs from its description. The first one is the ethical problems pertaining to bias amplification. Often, synthetic data inherits the statistical peculiarities and the cultural assumptions of its source generating models-particularly in synthetic data generation through large general-purpose LLMs like GPT-3.5 or AraGPT2. For instance, sentiment classifiers trained from synthetic data scraped or generated from social media may encode social biases based on gender, class, or region (Alotaibi & Saleh, 2024; Elmadany & Mubarak, 2022).

Secondly, a question is raised regarding overfitting on synthetic corpora whenever training occurs without sufficient variation or domain regularization. Studies indicated that there is a sharp drop in performance when those models are tested on real-life or noisy inputs (Özyilmaz et al., 2025), for instance, Whisper fine-tuned only on Tacotron-generated data. It diminishes considerations of the robustness of a model and the direct transfer to actual application areas.

Finally, the increasing prevalence of non-peer-reviewed preprints somehow contributes significantly to epistemic risk. Because the pace of innovation is fast and preprints are usually needed, some datasets come with a lack of methodological transparency, such as non-disclosed prompt templates or undocumented annotation procedures, thus rendering validation of findings and direct reproduction of results difficult, if not impossible (Alrehili & Alhothali, 2025; Allms, 2025).

8. Recommendations and Future Directions

The insights provided in the present study call forth a few urgent streams of inquiry aimed at the design and deployment of an Arabic NLP system rotated in Saudi dialects. These recommendations serve to orient both academics and developers in the creation of NLP models that are liberal, performant, and ethically grounded within the low-resource setting.

First, future research should focus on developing standardized benchmarks for the Saudi dialect. The present ecosystem presents fragmentation in evaluation protocols and, thereafter, restricts access to test sets, making a proper comparative study nearly impossible in the realm of dialectal research or even attempts at cross-validation. Given that reproducible experimentation and transparent benchmarking are the essentials, it is thus imperative to have this developed as a national-level exercise along the lines of GLUE or SuperGLUE with SyS Arabic dialect at its heart. With such a start, SARD (SARD Dataset Team, 2025) and SaudiBERT corpora (Qarah & Al-Sharif, 2024) can be used for expansion to cover reviews, sentiment, NER, ASR, and summarization tasks by researchers.

Second, it should consider hybrid training pipelines that fuse synthetic with real-world corpora, as has been proven beneficial. For instance, Whisper models that have been fine-tuned on Tacotron-generated speech benefit from noise-free pronunciation cues while being made more robust by the presence of real Saudi Arabic recordings (Alghamdi, 2023; Özyilmaz et al., 2025). Likewise, the composition of back-translated NER corpora with real annotated news text can maintain a balance between coverage and contextual fidelity (Alrowaished & Alotaibi, 2024).

Third, encourage the use of low-resource fine-tuning strategies, such as LoRA (Aryan et al., 2024) and quantized LoRA (Hossain & Al-Ayyoub, 2025), particularly in resource-limited environments. These approaches offer efficiency with no loss in performance, especially when modular architectures like AraT5 or mBART can be used.

Ethical issues also come into play in these matters. Synthetic data generators must be judged equally on the grounds of how well they perform, in consideration of how potential propagation of bias, misinformation, or dialectal homogenization can take root. To do so necessitates transparent documentation, open-sourced prompts, and human-in-the-loop curation.

Lastly, another important challenge is to further explore the domain-specific dialect modeling of legal, medical, educational, and religious discourse so that NLP systems can be extended to critical service areas where language variation is more sensitive. Fine-tuned systems can perform extraordinarily well once trained on context-specific dialect-aligned data, as shown in Al-Mutairi et al. (2024) and Aftan (2024).

9. Conclusion

This research has investigated the contemporary relationship between the generation of synthetic data and fine-tuning strategies in the dialects of Saudi Arabic NLP. The thematic review of 30 peer-reviewed and preprint studies has demonstrated that synthetic corpora created through back-translation, GANs, Tacotron, or multi-agent LLMs could alleviate the ever-present data scarcity in dialectal Arabic NLP. When followed up by powerful fine-tuning strategies like LoRA, Whisper, or AraT5, these synthetic datasets improve the final performance on sentiment analysis, speech recognition, NER, and summarization.

Key challenges of the current research bitterly revealed by this study included, among others, the lack of standardized benchmark, risks of overfitting by the models, as well as ethical concerns in randomly generated synthetic data, yet it appears that the study also uncovered some promising good-large pipelines that might serve as imprints for further studies and deployment, such as Tacotron + Whisper for ASR, AraGPT2 + SaudiBERT for sentiment.

Lastly, the amalgamation of synthetic data with fine-tuning is not viewed as a workaround per se anymore; it is a good idea for democratizing NLP among dialect-rich communities. Working with Saudi Arabic dialects will bridge a significant gap in relative neglect, enabling the prospect for focused innovations that marry linguistic consciousness with technical efficiency. Going forward, the challenge is not just to improve model performance, but to build dialectal NLP systems that are fair, transparent, and linguistically representative.

References

1. Aftan, S. S. (2024). Enhancing sentiment analysis in Saudi dialect using AraGPT2-generated synthetic data. *Texas Tech University Electronic Theses*. https://ttu-ir.tdl.org/handle/2346/95845
2. Alanazi, N. (2025). SauDial: The Saudi Arabic dialects game localization. *Journal of Digital Linguistics, 15*, 45–60. https://doi.org/10.1016/j.digling.2025.100215
3. Al-Dossari, L., & Al-Jasser, M. (2024). Speech recognition in Saudi Arabic: Performance of fine-tuned Whisper models. *Language and Speech Processing Journal, 6*(1), 70–85.
4. Alghamdi, E. A., & Alshammari, T. (2024). Domain adaptation for Arabic translation: Synthetic and back-translated data augmentation. *Applied Sciences, 14*(16), 7088. https://doi.org/10.3390/app14167088
5. Alghamdi, M. (2023). Speech synthesis and synthetic augmentation for Saudi Arabic using Tacotron. *Journal of Arabic Computational Linguistics, 3*(1), 50–66.
6. Al-Mutairi, M., Alessa, H., & Alghamdi, A. (2024). Generating synthetic Arabic medical dialogues using multi-agent LLMs. In *Proceedings of the Arabic NLP Workshop at ACL 2024* (pp. 15–24). https://aclanthology.org/2024.arabicnlp-1.2/
7. Al-Mutairi, M., & Ghamdi, A. (2024). Multi-agent synthetic dialogue generation for Arabic medical NLP. Proceedings *of the Arabic NLP Workshop at ACL 2024*. https://aclanthology.org/2024.arabicnlp-1.2/
8. Alotaibi, R., & Saleh, M. (2024). Arabic prompt tuning for dialect adaptation using small-scale synthetic corpora. *ACM Transactions on Asian Language Processing, 23*(2), 1–20.

9. Al-Qaysi, S. R., & Al-Rawi, W. (2023). Adapting GPT-3.5 for Saudi Arabic generation via in-context learning. *Neural Processing Letters, 57*, 4009–4026. https://doi.org/10.1007/s11063-023-11343-2
10. Alrehili, A., & Alhothali, A. (2025). Towards balanced synthetic data for Arabic grammatical error correction. *arXiv preprint*. https://arxiv.org/abs/2502.05312
11. Alrowaished, F., & Alotaibi, M. (2024). Back-translation-based synthetic data for Gulf Arabic NER. *AI in Society and Culture, 5*(1), 43–59.
12. Al-Shameri, N. (2024). Arabic paraphrased parallel synthetic dataset. *Data in Brief, 40*, 108123. https://doi.org/10.1016/j.dib.2024.108123
13. Al-Shenaifi, N. (2024). Advancing AI-driven linguistic analysis: Developing Saudi dialect corpora. *Mathematics, 12*(19), 3120. https://doi.org/10.3390/math12193120
14. Al-Saadi, H., & Alarifi, A. (2024). Evaluating fine-tuned Arabic models on Saudi dialect sentiment benchmarks. *Procedia Computer Science, 230*, 190–198. https://doi.org/10.1016/j.procs.2024.01.023
15. Allms, A. (2025). The landscape of Arabic LLMs: Dialect handling via synthetic data. *arXiv preprint*. https://arxiv.org/abs/2506.01340
16. Ameen, M. (2025). Leveraging synthetic data for dialect-specific ASR in low-resource Gulf variants. *Speech Technology Review, 12*(1), 28–41.
17. Aryan, P., Al-Ayyoub, M., & Jararweh, Y. (2024). Quantized LoRA fine-tuning for Arabic dialects using Qwen2-1.5B. *arXiv preprint*. https://arxiv.org/abs/2412.17548
18. Darwish, K., Mubarak, H., & Abdelali, A. (2022). Arabic dialect identification using transfer learning and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 337–348). https://aclanthology.org/2022.acl-long.27/
19. Elmadany, A., & Mubarak, H. (2022). Zero-shot cross-dialectal Arabic NLU with synthetic data generation. *arXiv preprint*. https://arxiv.org/abs/2210.05698
20. Fares, M., Touileb, S., & Habash, N. (2024). Bab el-Bot at AraFinNLP2024: Multi-dialect intent detection with synthetic data. In *Proceedings of the Arabic NLP Workshop at ACL 2024* (pp. 250–260). https://aclanthology.org/2024.arabicnlp-1.40/
21. Habash, N., & Bouamor, H. (2021). Dialectal Arabic NLP: A survey. *Computational Linguistics, 47*(3), 673–709. https://doi.org/10.1162/coli_a_00401
22. Hassan, R. A., & Al-Farhan, I. (2022). Fine-tuning mBART for Gulf Arabic text summarization. *International Journal of Language and Computation, 18*(2), 140–155.
23. Hossain, M., & Al-Ayyoub, M. (2025). Dialectal adaptation for Arabic NLP using LoRA fine-tuning. *Journal of Artificial Intelligence Research and Development, 19*(3), 233–250.
24. Ibrahim, A., & Hassan, R. (2023). Fine-tuning AraT5 with mixed Saudi corpora. *International Journal of AI Research, 6*(2), 144–160.
25. Khalifa, M., Hassan, H., & Fahmy, A. (2021). Zero-resource multi-dialectal Arabic NLU via self-training. *Transactions of the Association for Computational Linguistics, 9*, 1021–1034. https://doi.org/10.1162/tacl_a_00414
26. Mahzari, S. (2024). Creating a Saudi social media corpus for dialect adaptation in LLMs. *Digital Arabic Humanities, 9*(2), 95–110.
27. Mashraky, Y., & Bouamor, H. (2023). Generating realistic text for dialectal Arabic via GAN-based synthetic training. *Natural Language Engineering, 29*(5), 700–716. https://doi.org/10.1017/S1351324923000335
28. Özyilmaz, Ö. T., Coler, M., & Valdenegro-Toro, M. (2025). Overcoming data scarcity in Arabic ASR via Whisper fine-tuning. *arXiv preprint*. https://arxiv.org/abs/2506.02627

29. Qarah, F., & Al-Sharif, M. (2024). SaudiBERT: Pretraining on Saudi dialect corpora for sentiment classification. *arXiv preprint*. https://arxiv.org/abs/2405.06239

30. SARD Dataset Team. (2025). SARD: A synthetic Arabic OCR dataset for robust model fine-tuning. *arXiv preprint*. https://arxiv.org/abs/2505.24600