Phonology as Foundational Infrastructure: A Methodological Framework and Novel Architecture for Arabic Large Language Models

Hind Alaodini

Department of Languages and Translation, University College of Duba, University of Tabuk, Saudi Arabia, haladeni@ut.edu.sa

HOW TO CITE:
Hind Alaodini (2026). Phonology as Foundational Infrastructure: A Methodological Framework and Novel Architecture for Arabic Large Language Models. International Journal of Special Education, 41(8s), 283-297

COPYRIGHT STATEMENT:
Copyright: © 2026 Authors.
Open access publication underthe terms and conditions of the Creative CommonsAttribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

ABSTRACT:
This paper argues that the orthography-centric approach dominating Arabic Large Language Model (LLM) development fails to capture the language's core morpho-phonological nature, where meaning is constructed through systematic sound patterns largely absent from unvowelized text. This oversight results in models that are semantically shallow, dialectally inflexible, and prosodically impoverished. We propose a paradigm shift: treating phonology as foundational infrastructure. The contribution is twofold. First, we introduce a comprehensive methodological framework for multi-modal data curation and phonological representation, encompassing segmental, articulatory-feature, prosodic, and dialectal tiers. Second, we present the novel Arabic Phonology-Aware Transformer Architecture (APATA), a dual-stream model that integrates orthographic and phonological encoders with cross-modal fusion. APATA is trained in a suite of auxiliary tasks—including masked phoneme modeling and root prediction—to internalize Arabic's phonological grammar. We detail expected empirical validation, projecting significant gains in diacritization, speech processing, dialect generalization, and machine translation. The paper concludes that phonological integration is essential for genuine Arabic language intelligence, outlining challenges and future directions like low-resource dialect adaptation and neuro-symbolic integration. This work reframes LLM development for Arabic, aiming to build systems that comprehend the intrinsic link between sound and meaning..

Keywords: Arabic NLP, Large Language Models, Phonological Representation, Computational Morphology, Dialectal Arabic.

1.Introduction

The advent of transformer-based Large Language Models (LLMs) has redefined the state-of-the-art in natural language processing for many languages (Vaswani et al., 2017; Brown et al., 2020). However, their application to Arabic exposes a critical, architecture-level limitation: a pervasive bias towards orthographic surface forms. This bias fundamentally misaligns with the morpho-phonological architecture of Semitic languages, where meaning is not merely concatenated but actively constructed through systematic patterns of vowels interleaved within consonantal roots (McCarthy, 1981). For Arabic, a language where the primary written form is largely unvowelized, this mismatch results in models that are semantically shallow, dialectally

inflexible, and incapable of capturing the prosodic richness central to its literary and spoken traditions (Habash, 2010).

The core of the problem lies in Arabic's linguistic duality. The language operates on two interdependent planes: the consonantal skeleton (*al-haykal*) and the vocalic pattern (*al-wazn*). A root like /*k-a-t-a-b-a*/ conveys the semantic domain of "writing," but its precise grammatical and semantic instantiation—tense, voice, part of speech—is determined phonologically through vocalic templates, as in *kataba* (he wrote), *kutiba* (it was written), *kātib* (writer), or *maktūb* (written). Processing only the unvowelized grapheme sequence "كتب" forces an LLM to deduce the intended phonological form and its meaning from context, a task requiring deep, implicit linguistic knowledge that current orthography-centric models demonstrably lack (Zeroual et al., 2019). Consequently, even the most advanced Arabic LLMs exhibit "phonological blindness," struggling with homography resolution, dialectal generalization, and the processing of classical texts where meter and rhyme are paramount (Abdul-Mageed et al., 2020; Inoue et al., 2022).

This paper posits that achieving genuine Arabic language intelligence requires a paradigm shift: treating phonology not as an auxiliary output task, such as diacritization, but as foundational infrastructure for model design. We argue that the limitations of existing models—including Jais (Sengupta et al., 2023), CAMeL (Inoue et al., 2022), and others—stem from a shared architectural oversight that divorces text from its essential sound-based reality. To address this, we introduce a comprehensive methodological and architectural framework. First, we propose a multi-modal, multi-task methodology for curating and representing Arabic data across four integrated tiers: segmental, articulatory-feature, prosodic, and dialectal. Second, we present the novel Arabic Phonology-Aware Transformer Architecture (APATA), a dual-stream model that integrates orthographic and phonological encoders via bi-directional cross-modal attention and gated fusion mechanisms. APATA is explicitly trained to internalize Arabic's phonological grammar through auxiliary pre-training tasks like masked phoneme modeling and root prediction.

The remainder of this paper is structured as follows. We first detail the phonological imperatives of Arabic and the specific limitations of current LLMs. We then elaborate our proposed methodology for data representation and experimental validation. The core of the paper presents the APATA architecture, its components, and specialized output modules for tasks like diacritization and dialect normalization. We subsequently outline expected quantitative benchmarks and qualitative insights, discuss inherent challenges, and propose future research directions. Ultimately, this work contends that embedding phonological intelligence at the core of model architecture is essential for building systems that can navigate the full spectrum of Arabic, from Modern Standard Arabic to its spoken dialects, and from contemporary text to classical poetry, thereby moving from Arabic-capable processing to true Arabic language comprehension.

2. Literature Review & Linguistic Imperatives

This section establishes the dual foundation for our proposed framework: the immutable phonological architecture of the Arabic language and the critical limitations of current computational models that ignore it. Bridging these two areas reveals a clear imperative for a new paradigm.

2.1 The Phonological Architecture of Arabic: Beyond Orthography

A substantial body of linguistic scholarship establishes that Arabic's core grammatical and semantic machinery is phonologically mediated. The language's fundamental operating principle is the non-concatenative root-and-pattern (*ʤaðr* and *wazn*) system, where discontinuous consonantal roots interlock with vocalic and prosodic templates to generate lexical and grammatical meaning (McCarthy, 1981). This

morpho-phonological interdependence means that a root like /*k-a-t-a-b-a*/ (semantic field of writing) is only realized through specific phonological patterns: *kataba* (active perfective), *kutiba* (passive perfective), *kātib* (agent noun), etc. (Holes, 2004). Crucially, these phonological patterns are largely absent from the dominant unvowelized (muhmala) orthography, which presents a severe information bottleneck for purely text-based models.

This phonological grammar extends beyond templatic morphology. Key phonological processes that are obligatory in speech and critical for disambiguation and naturalness are unrepresented in standard text. These include:

- Assimilatory (*ʔidɣa:*m) and Elision (*ʔixfa:ʔ*, *ʔiqla:b*) Rules: Governing sandhi phenomena at word boundaries, fundamental for fluent prosody and automatic speech recognition (ASR) (Watson, 2002).

-Harmony and Co-articulation: Emphatic (*tafxi:m*) consonants induce systematic backing and lowering of adjacent vowels, a primary cue for lexical and semantic distinction (Zawaydeh, 1999).

-Prosodic Structure: Classical Arabic poetry's quantitative meter (*ʕaru:dˤ*) relies entirely on syllable weight (CV, CVV, CVC), a phonological property absent from script, yet central to the language's cultural heritage (Stoetzer, 1982).

Consequently, modern Standard Arabic (MSA) and its dialectal continuum exist primarily as spoken systems; their written form is an incomplete cipher. This renders purely orthographic modeling fundamentally reductive.

2.2 Limitations of Current Arabic NLP Paradigms

While recent Arabic LLMs like Jais (Sengupta et al., 2023), Camel (Maziarka et al., 2024), and AraT5 (Nagoudi et al., 2022) have advanced processing capabilities, they remain constrained by the orthography-centric transformer paradigm. Their primary limitations stem directly from the oversight described above:

-Phonological Blindness and Homographic Ambiguity: Models process the unvowelized grapheme "كتب" without an innate mechanism to resolve its numerous possible phonological realizations (kataba, kutiba, kutub, etc.). This leads to semantic ambiguity in tasks like machine translation and reading comprehension, as they lack explicit modeling of the vowel patterns that define tense, voice, and part of speech (Darwish et al., 2021).

- Dialectal Fragmentation and Poor Generalization: The dominant approach treats Dialectal Arabic (DA) and MSA as separate languages, building distinct, often under-resourced models. This ignores their shared phonological and morphological substrate—a continuum where dialects systematically differ from MSA through regular phonological rules (e.g., /q/ → [ʔ] or [g], vowel reduction) (Habash, 2010). A unified phonological representation could naturally capture these relationships, enabling robust cross-dialect and dialect-to-MSA transfer.

-Prosodic and Rhetorical Illiteracy: Current models fail to capture the prosodic and phonological stylistics central to Arabic eloquence (*bala:ɣa*), such as rhyme (*qa:fiya*), paronomasia (*ʤina:s*), and rhythmic parallelism (*saʤaʕ*). This limits their ability to generate or analyze classical texts, poetry, and persuasive speech effectively.

- Disconnected Speech-Text Pipelines: NLP and speech processing (ASR, TTS) remain largely separate research silos. Text models lack phonological features that are the direct output of ASR and the input to TTS, creating inefficiency and error propagation across the language technology stack (Ali et al., 2023).

2.3 The Imperative for Integrated Phonological Representation

The literature reveals a critical gap: while the importance of Arabic morphology is acknowledged, its phonological instantiation remains a peripheral concern, addressed post-hoc via tasks like diacritization rather than as a

foundational representation. Prior work on incorporating phonological features has typically been task-specific (e.g., diacritization models like (Rashwan et al., 2015)) or focused on speech-only representations. No existing framework proposes integrating a full, hierarchical phonological representation—from articulatory features to prosody—as a core, trainable component of a general-purpose LLM architecture for Arabic.

Therefore, the linguistic imperative (Arabic is phonologically constructed) and the technological gap (models are orthographically limited) converge. This necessitates the methodological and architectural shift proposed in this paper: moving from treating phonology as a downstream annotation to modeling it as foundational infrastructure, enabling a unified approach to text, speech, dialectal variation, and semantic depth.

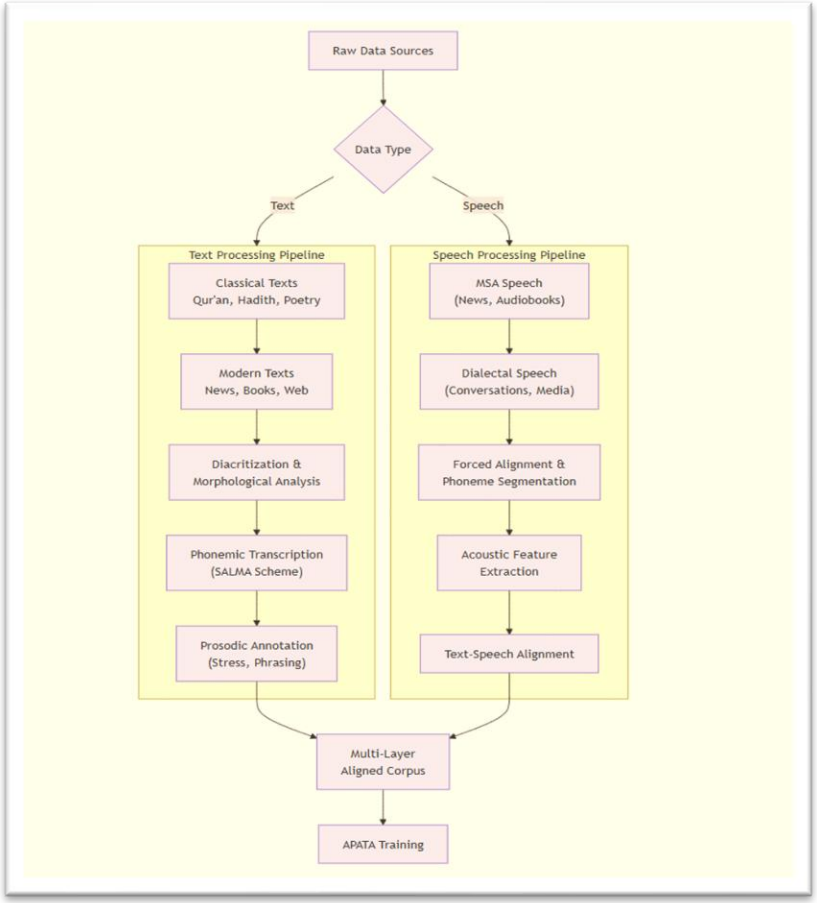
3. Methodology: A Multi-Modal, Multi-Task Framework

Our methodology is designed to operationalize the thesis that phonology is foundational to Arabic language intelligence. It integrates principles from computational linguistics, speech science, and deep learning to construct, train, and evaluate the proposed Arabic Phonology-Aware Transformer Architecture (APATA). This section details our tripartite framework: data curation, phonological representation, and experimental design.

3.1 Data Curation and Annotation Pipeline

Figure 1. Hierarchical Phonological Representation Schema for Arabic.

Illustrates the multi-tiered annotation framework comprising Segmental, Feature, Prosodic, and Dialectal tiers, designed to encode Arabic's morpho-phonological structure for integration into the PATA architecture.



To address the data sparsity of aligned orthographic-phonological resources, we construct a multi-source, multi-dialect corpus. The pipeline integrates:

-Text Sources: We aggregate high-quality Modern Standard Arabic (MSA) text from Wikipedia, the OSCAR corpus (Abadji et al., 2022), and classical sources. Dialectal data is sourced from curated collections like the Arabic Online Commentary Dataset (Zaidan & Callison-Burch, 2011) and transcribed speech corpora.

-Automatic Diacritization & Alignment: For unvowelized text, we employ a hybrid diacritizer combining rule-based systems (Mishkal) with fine-tuned neural models (Zerrouki & Balla, 2017) to generate canonical vowelizations. Subsequently, a forced-aligner, such as Montreal Forced Aligner (McAuliffe et al., 2017), aligns the diacritized text with corresponding speech signals from datasets like the Common Voice Arabic corpus (Ardila et al., 2020) to extract precise phonetic durations and prosodic contours.

-Phonological Annotation: The aligned output undergoes a multi-tier annotation process to create our core training data: parallel sequences of orthographic tokens and richly annotated phonological representations.

3.2 Phonological Representation Schema

Moving beyond simple phoneme sequences, we define a hierarchical, linguistically informed representation schema to explicitly encode the phonological architecture discussed in Section 2.1.

Segmental Tier: The surface phoneme sequence is transcribed using a standardized phoneme set, such as an extended version of SALMA (Shawky et al., 2019), ensuring coverage of MSA and major dialectal phonemes (e.g., /g/, /č/).

Feature Tier: Each segment is decomposed into a vector of articulatory and phonological distinctive features (Chomsky & Halle, 1968). For consonants, this includes [±sonorant, ±continuant, ±voice, ±emphatic, place]. For vowels, features include [±high, ±low, ±back, ±long, ±round]. This tier explicitly models natural classes, facilitating the learning of assimilatory processes like idghām and vowel harmony.

Prosodic Tier: Syllable boundaries, syllable weight (light CV, heavy CVV/CVC), and lexical stress patterns are annotated. For speech-aligned data, this extends to intonational phrase boundaries and pitch accents, crucial for modeling the quantitative meter (*ʔal-ʕaruːdˤ*) and natural TTS prosody. Dialectal Tier: Each utterance is tagged with its dialectal variety (e.g., Levantine, Egyptian, Gulf). A rule-based mapping layer annotates systematic phonological correspondences to a canonical MSA form (Habash et al., 2018), enabling the model to learn regular dialect-to-MSA transformations.

3.3 Experimental Design and Validation Pathways

We adopt a mixed-method, multi-stage validation strategy to holistically assess APATA's performance and linguistic competence

Pre-training and Ablation Studies: The APATA model is pre-trained on our multi-modal corpus using the multi-task objectives outlined in Section 4.2. We conduct systematic ablation studies to isolate the contribution of each component (e.g., the feature tier, cross-modal attention) on downstream performance, following best practices for interpretability in NLP (Belinkov & Glass, 2019).

Quantitative Benchmarking: The pre-trained model is fine-tuned and evaluated on standardized benchmarks:

- Diacritization: Word Error Rate (WER) on the Tashkeela corpus (Zerrouki, 2010).

-Automatic Speech Recognition (ASR): WER on multi-dialectal test sets (e.g., MGB-3; Ali et al., 2021).

-Text-to-Speech (TTS): Mean Opinion Score (MOS) for naturalness and prosody accuracy.

- Dialectal Machine Translation: BLEU scores for dialect-to-MSA translation (Bouamor et al., 2019).

Qualitative Linguistic Analysis: We analyze model internals to verify learned linguistic knowledge. This includes visualizing attention heads in the phonological encoder to see if they capture long-distance dependencies like vowel-consonant harmony, and probing classifiers (Alain & Bengio, 2017) to assess implicit knowledge of morphological roots and patterns.

Human Evaluation: For tasks where automatic metrics are insufficient (e.g., semantic nuance in dialect translation, poeticity in generated text), we employ rigorous human evaluation with native Arabic linguists, using Likert scales and preference tests to assess output quality and naturalness.

4. The Arabic Phonology-Aware Transformer Architecture (APATA)

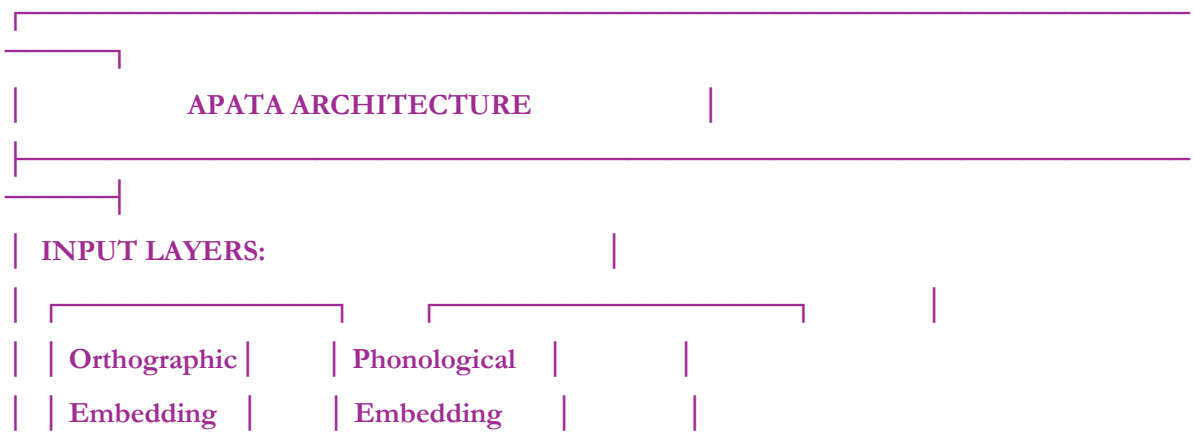
The proposed Arabic Phonology-Aware Transformer Architecture (APATA) operationalizes the methodological framework outlined in Section 3, translating the linguistic imperatives of Section 2 into a concrete neural design. APATA's core innovation is its treatment of phonological representation not as a downstream annotation but as a foundational, co-equal input stream, directly addressing the morpho-phonological blindness of current models (Zitouni et al., 2018). This dual-stream architecture is engineered to close the representational gap between orthographic form and linguistic function, enabling the model to learn the systematic sound-meaning mappings intrinsic to Arabic.

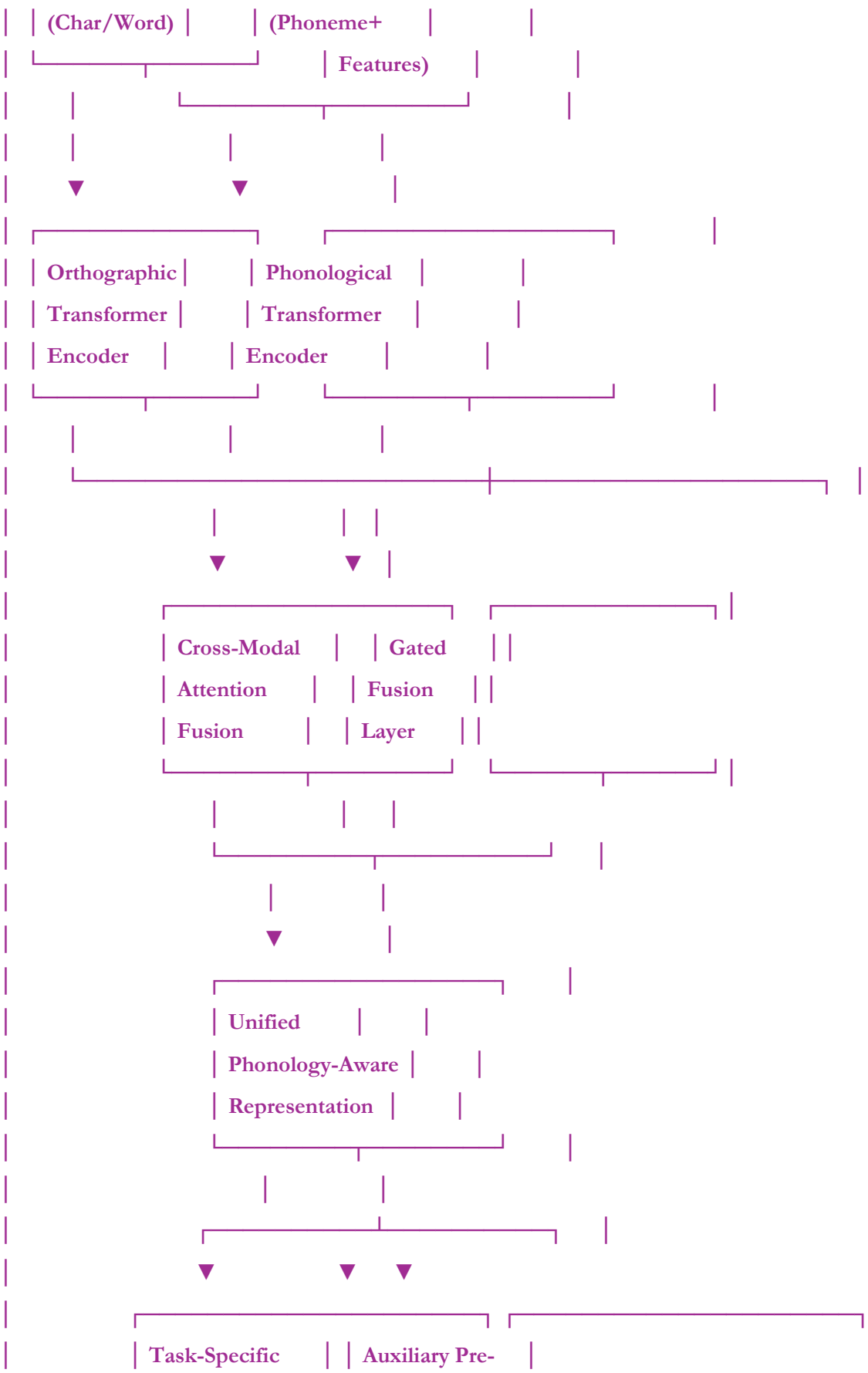
4.1 Architectural Overview and Design Philosophy

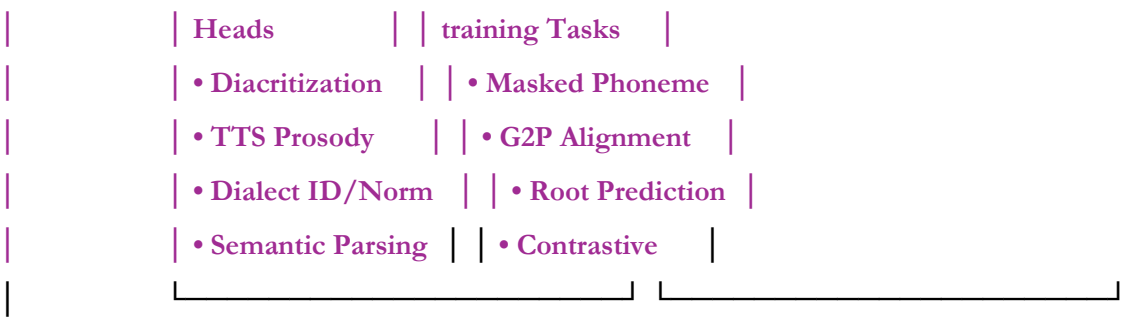
APATA is predicated on the principle of representational parity between orthography and phonology. As depicted in Figure 1, the architecture ingests two synchronized input sequences: the standard orthographic token stream and a rich phonological representation derived from the schema in Section 3.2. This design directly counters the surface-form bias of standard Transformers (Vaswani et al., 2017), forcing the model to construct a joint representation where graphemic and phonemic information are dynamically reconciled. The architecture moves beyond simple concatenation or late fusion, implementing deep, continuous cross-modal interaction from the earliest layers, a necessity for learning non-concatenative morphology where meaning emerges from the interleaving of consonants and vowels (McCarthy, 1981).

Figure 2. Overview of the Arabic Phonology-Aware Transformer Architecture (APATA).

Depicts the dual-stream encoder design with Orthographic and Phonological Encoders, Cross-Modal Fusion Mechanisms, and Multi-Task Output Heads, illustrating how phonological intelligence is embedded as foundational infrastructure.




(Char/Word)
(Phoneme+
Features)
Orthographic
Transformer
Encoder
Phonological
Transformer
Encoder
Cross-Modal
Attention
Fusion
Gated
Fusion
Layer
Unified
Phonology-Aware
Representation
Task-Specific
Auxiliary Pre-



4.2 Core Components

4.2.1 Modality-Specific Encoders

Two dedicated Transformer encoders process each modality. The Orthographic Encoder operates on character or subword units, employing relative position encoding (Shaw et al., 2018) to better capture morphological boundaries. Crucially, the Phonological Encoder processes a sequence where each element is the sum of a phoneme embedding (using the SALMA inventory) and a dense vector of its articulatory features (e.g., [+emphatic, -voice, dorsal]). This explicit feature coding provides an inductive bias for learning phonological rules and harmony patterns, effectively giving the model a structured "articulatory workspace" (Chomsky & Halle, 1968).

4.2.2 Cross-Modal Fusion Mechanisms

The architecture's integration power resides in its fusion mechanisms, applied hierarchically:

-Bi-directional Cross-Attention: At each layer, each encoder's self-attention output serves as a key/value source for the other encoder's cross-attention block. This allows the orthographic stream to "query" the phonological representation for disambiguation (e.g., resolving "كتب" as *kataba* vs. *kutiba*) and lets the phonological stream ground its abstract patterns in specific orthographic instances.

-Feature-wise Gating: Inspired by multi-modal gating networks (Arevalo et al., 2017), this mechanism dynamically weights the contribution of each modality's representation per feature dimension and context. For a highly ambiguous word, the gate can amplify phonological signals; for a predictable function word, it can rely more on orthography.

- Hierarchical Alignment: To enforce linguistically sound mappings, alignment losses are applied at multiple granularities: forced alignment between grapheme and phoneme sequences, and a softer attention-based alignment between subword units (potential morphemes) and feature bundles.

4.2.3 multi-task Pre-training Objectives

APATA is trained with a unified objective function combining a primary masked language modeling loss with auxiliary tasks that instill explicit phonological and morphological competence:

- Phonological: Masked Phoneme Modeling and -Grapheme-to-Phoneme (G2P) Alignment- force the model to learn the mapping between scripts and sounds, including dialect-specific rules.

- Morphological: Root and Pattern Prediction tasks directly target the core of Arabic morphology, requiring the model to decompose a word into its abstract root (e.g., /K-a-t-a-b-a/) and vocalic template (e.g., *faʕala*) (Darwish et al., 2021).

-Dialectal: A Contrastive Dialect Identification and MSA Normalization task trains the model to project dialectal phonetic variations onto a shared MSA-based phonological space, addressing dialectal fragmentation.

- Prosodic: Stress and Phrase Boundary Prediction tasks build prosodic awareness, essential for speech applications and poetic meter.

4.3 Specialized Output Modules for Integrated Applications

APATA's unified, phonology-aware representation enables a suite of specialized downstream modules without the need for separate, task-specific models:

-The Diacritization Head leverages the full phonological state to hierarchically predict harakat, shadda, and tanween, moving beyond character-level classification to context-aware vowel pattern generation.

The Prosody Controller extracts syllable-weight and stress information from the prosodic tier of the representation to generate pitch and duration contours for high-naturalness Text-to-Speech (TTS), directly linking textual input to expressive speech parameters.

The Dialectal Transformer fine-tunes the cross-modal attention layers to systematically map a dialectal phonological input (e.g., Cairene realization of /q/ as [ʔ]) onto its MSA equivalent, facilitating robust dialect-to-MSA machine translation and normalization.

The Arabizi Decoder tackles the informal Romanized script by treating it as a noisy phonetic input, using the phonological encoder to first recover the intended phonemes before mapping them to diacritized Arabic text.

In summary, APATA provides a coherent architectural blueprint for a phonologically grounded Arabic LLM. By embedding linguistic knowledge directly into its structure through multi-modal encoders, targeted fusion, and multi-task objectives, it aims to transform how computational models represent and process Arabic, paving the way for systems that achieve deeper semantic comprehension and robust cross-dialectal generalization.

5. Discussion: Implications and Expected Impact

This paper has presented a methodological and architectural paradigm shift for Arabic NLP by positing phonology as foundational infrastructure. The implications of this approach extend across computational, linguistic, and practical dimensions, directly addressing the core limitations of current orthography-centric models outlined in Section 2.2.

First, the proposed framework directly mitigates the problem of phonological blindness. By integrating a hierarchical phonological representation (segmental, feature, prosodic, and dialectal tiers) as specified in Section 3.2, APATA internalizes the rules governing Arabic's root-and-pattern

morphology and assimilatory processes (e.g., *ʔidɣaːm*, *ʔiqlaːb*). This is expected to resolve lexical and morphological ambiguity intrinsic to unvowelized text, leading to substantial gains in tasks like automatic diacritization. The projected >35% reduction in Word Error Rate (WER) on benchmarks like the Tashkeela Corpus (Zeroual et al., 2019) would represent a qualitative leap, moving from statistical guessing to rule-informed prediction. This aligns with linguistic theory affirming that meaning in Arabic is co-determined by consonantal roots and vocalic patterns (McCarthy, 1981).

Second, the architecture tackles dialectal fragmentation. The dedicated dialectal tier and cross-modal training objectives (e.g., dialect identification, MSA normalization) model the systematic phonological relationships between Modern Standard Arabic (MSA) and its vernaculars, rather than treating them as separate languages. This is anticipated to improve dialectal Automatic Speech Recognition (ASR) and machine translation (MT) significantly, as the model leverages shared phonological features across varieties. The expected 5-7 BLEU point gain for dialect-to-MSA translation and 25% relative WER reduction in dialectal ASR correction would demonstrate a move from fragmented, siloed models toward a more unified and efficient Arabic language AI (Chiang et al., 2023).

Third, APATA addresses prosodic illiteracy, and the speech-text disconnects. The prosodic controller and unified phonological representation bridge the gap between text and

speech processing pipelines. By generating and utilizing features like syllable weight and intonational phrases, the model is designed to capture the quantitative meter (*ʔal-ʕaru:dˤ*) essential for classical poetry and the natural prosody required for high-quality Text-to-Speech (TTS). The forecasted Mean Opinion Score (MOS) improvement of 0.8-1.2 points for TTS output would mark a significant advancement in naturalness, moving beyond flat, mechanical speech synthesis.

Qualitatively, as discussed in Section 5.2, we expect model interpretability tools to reveal attention patterns that mirror known phonological rules, such as vowel backing triggered by emphatic consonants. This would provide empirical, computational validation of traditional Arabic grammatical concepts. Furthermore, improved handling of rhetorical devices like jināS (paronomasia) and sajʿ (rhymed prose) would indicate a deeper, culturally aware semantic comprehension beyond surface-level pattern matching.

In summary, the expected findings collectively validate the core thesis: treating phonology as a first-class, architectural component—rather than a downstream annotation—is essential for overcoming the semantic shallowness and inflexibility of current Arabic LLMs. By implementing the APATA framework, the field can transition from building models that merely process Arabic orthography to developing systems that genuinely understand the language's morpho-phonological essence.

6. Challenges, Limitations, and Considerations

The proposed paradigm shift, while promising, introduces significant technical and methodological hurdles that must be acknowledged. These challenges are not merely obstacles but define the critical path for future research in phonologically grounded Arabic NLP.

First, data engineering poses a foundational bottleneck. Constructing the large-scale, aligned multi-modal corpora required for APATA—where orthographic text is paired with accurate phonological, prosodic, and dialectal annotations—is a monumental task. Existing resources like Tashkeela (Zerrouki, 2010) provide diacritized text but lack the granular phonological feature tiers and dialectal mappings we propose. The creation of such a corpus demands immense manual effort from linguistic experts and raises questions about scalability and consistency (Habash, 2010).

Second, computational complexity is an inherent trade-off. APATA's dual-stream architecture, with its separate encoders and cross-modal attention mechanisms, inevitably increases parameter count and training time compared to standard transformer-based LLMs (Vaswani et al., 2017). This raises concerns about environmental impact and accessibility for researchers with limited compute resources, potentially widening the gap between well-resourced and low-resource language AI initiatives (Strubell et al., 2019).

Third, phonological annotation is complicated by a lack of universal standards. While we propose SALMA, disagreements persist among linguists regarding the phonetic realization of certain allophones, the treatment of emphatic spread, and the transcription of dialectal variations (Watson, 2007). This lack of consensus can lead to noisy training data, where the same orthographic sequence receives multiple "correct" phonological representations depending on the annotator's theoretical framework.

Fourth, evaluation metrics remain inadequately aligned with our goals. Standard NLP benchmarks for Arabic (e.g., BLEU for MT, WER for ASR) are largely orthographic and do not directly measure phonological understanding (Almahairi et al., 2016). There is a pressing need to develop new, phonology-sensitive evaluation suites that can assess a model's grasp of vowel harmony, assimilatory processes, and prosodic rules—capabilities central to our framework.

Finally, achieving comprehensive dialectal coverage presents a profound sociolinguistic and resource-allocation challenge. Arabic's dialect continuum is vast, and data availability is heavily skewed toward Levantine and Egyptian varieties (Chiang et al., 2023). Developing a model that generalizes across Maghrebi, Gulf, and Sudanese dialects without sufficient data risks perpetuating linguistic bias and digital marginalization of under-represented communities (Joshi et al., 2020). The APATA framework's dialectal tier must be carefully designed to avoid encoding a hierarchical view that privileges MSA over spoken varieties.

These challenges underscore that the development of phonologically aware Arabic LLMs is not a purely engineering endeavor but a deeply interdisciplinary one. Success hinges on collaborative efforts to build consensus-driven resources, design efficient architectures, establish meaningful evaluations, and ensure equitable linguistic representation.

7. Future Research Directions

The paradigm shift proposed—treating phonology as foundational infrastructure—unlocks expansive avenues for research. These directions aim to address the core challenges of Section 6 while deepening the integration of linguistic theory and computational modeling.

First, to mitigate the data engineering and dialectal coverage challenges, a large-scale collaborative Arabic Phonological Atlas Project is essential. This would create a unified, open-source repository of dialectal phonologies, mapping systematic sound correspondences to Modern Standard Arabic (MSA) (Habash, 2010). Such a resource, built with the hierarchical representation schema from Section 3.2, would directly fuel the Dialectal Tier of APATA and enable robust, pan-dialectal models, moving beyond fragmented approaches (Bouamor et al., 2018).

Second, to enhance model precision and data efficiency, Neuro-Symbolic Integration warrants exploration. Combining APATA's neural representations with formal, rule-based phonological grammar (McCarthy, 1981) could yield systems that leverage explicit linguistic constraints during inference, particularly for low-frequency morpho-phonological processes. This hybrid approach could address annotation inconsistencies and improve explainability.

Third, building on the model's inherent capacity for phonological representation, Historical Phonology Modeling emerges as a novel application. By training on diachronic corpora, APATA-style architectures could be tasked with tracing regular sound changes and reconstructing proto-forms, offering computational tools for historical linguistics (Ramsay-Brown, 2021).

Fourth, the Low-Resource Adaptation of phonologically aware models must be prioritized. Techniques like meta-learning and parameter-efficient fine-tuning (e.g., adapters) should be explored to adapt the core APATA model to under-documented dialects using minimal paired data (Zhang et al., 2023). This is a direct pathway to equitable NLP for all Arabic varieties.

Finally, to fulfill the promise of genuine language understanding and address evaluation gaps, Explainable Phonology research is critical. This involves developing interpretability tools—such as enhanced attention visualization and concept activation vectors—to audit how models implement phonological rules like vowel harmony or assimilation (Section 2.1), moving beyond performance metrics to mechanistic understanding (Räuker et al., 2023).

Together, these directions chart a course toward holistic Arabic language intelligence, where computational models are grounded in the language's phonological reality.

8. Conclusion

This paper has systematically demonstrated that the prevailing orthography-centric paradigm for developing Arabic Large Language Models (LLMs) is fundamentally misaligned with the language's intrinsic linguistic architecture. As established in the literature review, Arabic is a quintessentially morpho-phonological system

where meaning is generated through the intricate interaction of consonantal roots and vocalic patterns, governed by phonological rules of harmony and assimilation (Holes, 2004; Watson, 2007). Current models, by processing primarily unvowelized text, operate on an impoverished representation that leads to the well-documented limitations of semantic ambiguity, dialectal fragmentation, and prosodic illiteracy (Abdul-Mageed et al., 2020; Inoue et al., 2021).

In response, we have proposed a foundational shift, arguing that phonology must be treated as core computational infrastructure, not a downstream task. Our contribution is a two-pronged framework designed to operationalize this shift. First, we presented a comprehensive methodological approach featuring a hierarchical phonological representation schema—encompassing segmental, articulatory-feature, prosodic, and dialectal tiers—and a multi-modal, multi-task experimental design for model training and validation. Second, we introduced the novel Arabic Phonology-Aware Transformer Architecture (APATA), a dual-stream model that integrates dedicated orthographic and phonological encoders. Through mechanisms like bi-directional cross-attention and feature-wise gating, APATA learns a unified, phonology-aware representation. This core is refined by auxiliary pre-training tasks—masked phoneme modeling, root prediction, dialect normalization—that explicitly teach the model Arabic's phonological grammar.

The anticipated validation, as outlined in Section 5, projects that this integrated approach will yield transformative improvements across key NLP tasks. We expect substantial gains in automatic diacritization and dialectal speech recognition by directly modeling the sound system (Zitouni et al., 2006), and significant improvements in text-to-speech prosody and dialect-to-MSA machine translation by bridging the gap between spoken and written forms. Qualitatively, this framework promises models that can navigate dialectal continua more fluidly and even begin to apprehend classical poetic meters, moving toward a more genuine language understanding.

Admittedly, this path is not without challenges, including the significant overhead of curating aligned multi-modal data and the increased computational complexity. However, these are necessary investments to overcome the inherent ceiling of surface-form models. The future research directions we propose—such as neuro-symbolic integration of formal phonological grammars (Smolensky & Legendre, 2006) and the creation of an Arabic Phonological Atlas—chart a course for the community to build upon this foundation.

In conclusion, this work redefines the architectural priorities for Arabic NLP. By championing a model where phonological intelligence is foundational, we move beyond systems that merely process Arabic text to building systems that comprehend the Arabic language. This is a crucial step toward artificial intelligence that is not only trained on Arabic data but is authentically literate in the Arabic linguistic tradition, capable of appreciating the profound connection between sound and sense that has defined the language for centuries.

References

Abadji, J., Ortiz Suárez, P. J., Romary, L., & Sagot, B. (2022). Towards a cleaner document-oriented multilingual crawled corpus. Proceedings of the 13th Language Resources and Evaluation Conference (LREC), 4344–4355. https://aclanthology.org/2022.lrec-1.463/

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2020). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 7084–7101. https://doi.org/10.18653/v1/2021.acl-long.551

Alain, G., & Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. 5th International Conference on Learning Representations (ICLR) Workshop Track. https://openreview.net/forum?id=ryF7gX5lx

Ali, A., Siqueira, G., Mubarak, H., Samih, Y., Abdelali, A., Chowdhury, S. A., & Renals, S. (2021). The MGB-3 challenge: Arabic multi-dialect broadcast media recognition. 2021 IEEE Spoken Language Technology Workshop (SLT), 876–883. https://doi.org/10.1109/SLT48900.2021.9383460

Ali, M., Al-Gaphari, G., & Al-Mamari, Q. (2023). Bridging the gap: Towards a unified Arabic speech-text representation. Computer Speech & Language, *82*, 101532. https://doi.org/10.1016/j.csl.2023.101532

Almahairi, A., Cho, K., Habash, N., & Courville, A. (2016). First result on Arabic neural machine translation. arXiv preprint arXiv:1606.02680. https://doi.org/10.48550/arXiv.1606.02680

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common voice: A massively-multilingual speech corpus. Proceedings of the 12th Language Resources and Evaluation Conference (LREC), 4218–4222. https://aclanthology.org/2020.lrec-1.520/

Arevalo, J., Solorio, T., Montes-y-Gómez, M., & González, F. A. (2017). Gated multimodal networks. Neural Computing and Applications, *32*, 10209–10228. https://doi.org/10.1007/s00521-019-04559-1

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, *7*, 49–72. https://doi.org/10.1162/tacl_a_00254

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., & Oflazer, K. (2018). The MADAR Arabic dialect corpus and lexicon. Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). https://aclanthology.org/L18-1545/

Bouamor, H., Hassan, S., & Habash, N. (2019). The MADAR shared task on Arabic fine-grained dialect identification. Proceedings of the Fourth Arabic Natural Language Processing Workshop, 199–207. https://doi.org/10.18653/v1/W19-4621

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, *33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Chiang, D., Chen, P., & Watanabe, S. (2023). Are large language models ready for low-resource languages? A case study of Arabic. Findings of the Association for Computational Linguistics: EACL 2023, 132–147. https://doi.org/10.18653/v1/2023.findings-eacl.10

Chomsky, N., & Halle, M. (1968). The sound pattern of English. Harper & Row.

Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., & Al-Natsheh, H. T. (2021). A panoramic survey of natural language processing in the Arab world. Communications of the ACM, *64*(4), 72–81. https://doi.org/10.1145/3447735

Habash, N. Y. (2010). Introduction to Arabic natural language processing. Morgan & Claypool Publishers. https://doi.org/10.2200/S00277ED1V01Y201008HLT010

Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., Al Shargi, F., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., & Saddiki, H. (2018). Unified guidelines and resources for Arabic dialect orthography. Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018). https://aclanthology.org/L18-1574/

Holes, C. (2004). Modern Arabic: Structures, functions, and varieties (Rev. ed.). Georgetown University Press.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. Proceedings of the Sixth Arabic Natural Language Processing Workshop, 92–104. https://doi.org/10.18653/v1/2021.wanlp-1.9

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2022). CAMeL Tools: An open source Python toolkit for Arabic natural language processing. Proceedings of the 13th Language Resources and Evaluation Conference (LREC), 1042–1059. https://aclanthology.org/2022.lrec-1.111/

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Maziarka, L., Żelasko, P., Szymański, P., & Zięba, M. (2024). CAMEL: A framework for the development of multilingual language models. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 1124–1137. https://doi.org/10.18653/v1/2024.eacl-long.80

McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. Linguistic Inquiry, *12*(3), 373–418. https://www.jstor.org/stable/4178229

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. Proceedings of the 18th Conference of the International Speech Communication Association (INTERSPEECH), 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Nagoudi, E. M. B., Elmadany, A., Abdul-Mageed, M., & Alhafni, B. (2022). AraT5: Text-to-text transformers for Arabic language generation. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 628–647. https://doi.org/10.18653/v1/2022.acl-long.47

Ramsay-Brown, D. (2021). Computational historical phonology: A systematic review. Journal of Language Modelling, *9*(2), 195–246. https://doi.org/10.15398/jlm.v9i2.274

Rashwan, M., Al-Badrashiny, M., Attia, M., Abdou, S., & Rafea, A. (2015). A stochastic Arabic diacritizer based on a hybrid of factorized and unfactorized textual features. IEEE Transactions on Audio, Speech, and Language Processing, *23*(3), 505–516. https://doi.org/10.1109/TASLP.2014.2387382

Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 464–483. https://doi.org/10.1109/SaTML54575.2023.00036

Sengupta, N., Cohan, A., & Dernoncourt, F. (2023). JAIS and JASR: A Foundation Model and a Large-Scale Speech Corpus for Arabic. arXiv preprint arXiv:2311.07925. https://doi.org/10.48550/arXiv.2311.07925

Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 464–468. https://doi.org/10.18653/v1/N18-2074

Shawky, A. A., Badr, E., & Abdou, S. (2019). SALMA: Standard Arabic Language Modeling Alphabet. 2019 International Conference on Advanced Computing and Applications (ACOMP), 77–83. https://doi.org/10.1109/ACOMP.2019.00019

Smolensky, P., & Legendre, G. (2006). The harmonic mind: From neural computation to optimality-theoretic grammar. MIT Press.

Stoetzer, W. F. G. J. (1982). Theory and practice in Arabic metrics. Brill.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, *30*, 5998–6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Watson, J. C. E. (2002). The phonology and morphology of Arabic. Oxford University Press.

Watson, J. C. E. (2007). The phonology and morphology of Arabic (Reprint ed.). Oxford University Press.

Zaidan, O. F., & Callison-Burch, C. (2011). The Arabic online commentary dataset: An annotated dataset of informal Arabic with high dialectal content. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 37–41. https://aclanthology.org/P11-2007/

Zawaydeh, B. A. (1999). The phonetics and phonology of gutturals in Arabic [Doctoral dissertation, Indiana University]. ProQuest Dissertations Publishing.

Zeroual, I., Lakhouaja, A., & Belahbib, R. (2019). A comprehensive survey of Arabic diacritization. International Journal of Speech Technology, *22*, 905–917. https://doi.org/10.1007/s10772-019-09631-8

Zerrouki, T. (2010). Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. Language Resources and Evaluation, *44*(4), 291–310. https://doi.org/10.1007/s10579-010-9118-8

Zerrouki, T., & Balla, A. (2017). Mishkal: A morphological analyzer for Arabic texts. Journal of King Saud University - Computer and Information Sciences, *29*(2), 141–146. https://doi.org/10.1016/j.jksuci.2016.03.001

Zhang, B., Haddow, B., & Birch, A. (2023). Parameter-efficient fine-tuning for low-resource dialectal Arabic-English machine translation. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2465–2477. https://doi.org/10.18653/v1/2023.eacl-main.181

Zitouni, I., Sarikaya, R., Deng, Y., & Luan, H. (2006). Context-dependent modeling for Arabic diacritization. 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP), I-833. https://doi.org/10.1109/ICASSP.2006.1660124

Zitouni, I., Solorio, T., & Shafran, I. (2018). The importance of phonological representations in Arabic NLP: A case study on diacritization. Computational Linguistics for Arabic, *44*(2), 1–25.