International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Toward a Culturally Embedded Arabic LLM: Localizing Linguistic AI for the Arab World[1]

Fahd A. Alhathloul [1] and Dr. Noha A. AlOwedi [2]
[12] *Department of English Language and Translation, College of Science and Theoretical Studies, Saudi Electronic University*
Corresponding Author: Fahd A. Alhathloul, **E-mail**: fhathloul@seu.edu.sa

| ABSTRACT

Arabic large language models (LLMs) increasingly support communication, education, public services, and knowledge access, yet fluent Arabic generation does not necessarily constitute culturally appropriate understanding. This paper examines the problem of cultural and pragmatic misalignment in contemporary Arabic-capable models and proposes a framework for culturally embedded Arabic AI. Cultural embeddedness is operationalized through four interrelated capacities: Arabic-native semantic representation, register awareness, cultural inference, and dialectal continuity. Using a mixed-methods design, the study introduces the Culturally Sensitive Test Suite for Arabic (CSTS-Arabic) and a four-dimensional Cultural Quotient (CQ-Score) for evaluating contextual appropriateness, dialectal consistency, cultural-reference integrity, and pragmatic success. The study compares multilingual models, Arabic-focused models, and a proof-of-concept Culturally Aware Language Model (CALM). The reported findings indicate a persistent cultural competence gap in multilingual systems and improved performance when dialect-aware representations, structured cultural retrieval, and pragmatic attention are integrated into the modeling pipeline. The paper further develops an Arabic-first architectural proposal, a tiered corpus strategy, and ethical safeguards concerning representation, doctrinal plurality, epistemic authority, evaluation bias, and resource justice. Rather than treating culture as an optional layer of factual knowledge, the study argues that it should be modeled as a constitute part of linguistic meaning. The proposed framework therefore positions culturally embedded Arabic LLMs as both a technical and governance challenge and offers a pathway toward more reliable, pluralistic, and regionally accountable Arabic AI.

| KEYWORDS

Arabic natural language processing; culturally aware AI; large language models; computational pragmatics; Arabic dialects; AI ethics

| ARTICLE INFORMATION

ACCEPTED: 25 July 2026 **PUBLISHED:** 01 September 2026 **DOI:** 10.32996/ijllt.2026.9.9.6

1. Introduction

The current paradigm of artificial intelligence is being fundamentally reshaped by Large Language Models (LLMs), which have transitioned from research prototypes to critical socio-technical infrastructure. However, the prevailing narrative of "intelligence" in these systems remains, at its core, an Anglo-centric one. Global benchmarks like MMLU, HELM, and BIG-bench define linguistic and reasoning competence primarily through the lens of English-language data and Western cultural contexts (Liang et al., 2022). In this landscape, non-English languages, including Arabic—one of the world's most widely spoken and culturally rich languages—are frequently treated not as first-class cognitive systems worthy of intrinsic modeling, but as derivative translation surfaces. This approach perpetuates a form of digital linguistic dependency, where AI capabilities for a language are measured by their proximity to English outputs rather than their fidelity to native conceptual structures.

[1] This research received grant no. (734) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

The core problem is that Arabic is not merely a lexical code to be mapped onto English semantic frames. It is a linguistic architecture built on a non-linear root-and-pattern morphology (Habash, 2010) that generates dense semantic networks (e.g., the root *k-t-b* linking writing, office, book, and destiny). Its communicative logic is inherently high context, where meaning is deeply embedded in shared cultural, religious, and historical presuppositions (Hall, 1976), and its sociolinguistic reality is a stratified continuum from Classical Quranic Arabic to Modern Standard Arabic (MSA) and a vibrant mosaic of regional dialects (Bassiouney, 2020). To treat Arabic in AI as a translation task is to commit a category error, mistaking the surface form for the underlying conceptual machinery.

Consequently, state-of-the-art multilingual and Arabic-specialized LLMs exhibit systematic failures that are not merely lexical but cultural-pragmatic. They may produce grammatically fluent MSA while utterly missing the implicit social calculus, hierarchical nuance, or contextual inference that is obligatory for genuine understanding. For instance, the phrase "الله يرضى عليك" might be translated literally as "May God be pleased with you," while its pragmatic function as a softening politeness strategy in a request is lost (Alfaifi & Atwell, 2017). Similarly, a model might translate the slang "مشي حالك" as "walk your condition" instead of inferring its meaning of reluctant acceptance under constraint. These are not errors of vocabulary but failures of cultural inference, revealing models that are linguistically competent yet culturally blind.

This cultural blindness has tangible consequences. Studies in Arabic NLP consistently report high error rates in dialectal processing and pragmatic tasks, where models struggle with speaker intent, social hierarchy, and presupposed context (Abdul-Mageed et al., 2020). When deployed in high-stakes domains such as culturally sensitive healthcare, education tailored to regional curricula, or legal navigation within mixed civil-Sharia systems, these deficiencies translate into outputs that are at best irrelevant and at worst harmful or offensive. The assumption that improving machine translation quality is sufficient for downstream Arabic AI tasks is empirically unsound; it addresses fluency while neglecting appropriateness and pragmatic accuracy.

This paper argues that for an Arabic LLM to be truly effective and trustworthy, it must be culturally embedded. We define this not as the superficial inclusion of local references, but as the explicit design of a model whose internal representations and reasoning mechanisms are architected to internalize Arabic's unique linguistic structures and cultural logic. An Arabic LLM, in the strong sense proposed here, is a system whose latent space reflects how Arabic speakers conceptualize agency, temporality, obligation, social relations, and collective identity.

The problem is increasingly visible in multilingual evaluation research. Recent work distinguishes linguistic competence from cultural competence and shows that performance may depend on the alignment between a language and its associated cultural context (Ying et al., 2025). Likewise, culturally diverse benchmarks demonstrate that strong general performance does not guarantee reliable answers for underrepresented regions and cultural settings (Chiu et al., 2025). For Arabic specifically, recent culturally grounded evaluation initiatives further reinforce the need for localized assessment rather than reliance on globally dominant benchmarks alone (SaudiCulture, 2025).

The present study therefore uses the term culturally embedded Arabic LLM to refer to a model whose representations, retrieval mechanisms, training objectives, and evaluation procedures are designed around Arabic linguistic variation and culturally situated meaning. The study makes three contributions. First, it provides an operational definition that distinguishes an Arabic-native orientation from mere Arabic output capability. Second, it introduces the CSTS-Arabic benchmark and CQ-Score as tools for evaluating contextual and pragmatic competence. Third, it proposes the Culturally Aware Language Model (CALM), a modular architecture and data strategy intended to make cultural information computationally actionable rather than implicitly hoped for in large-scale training data.

2. Literature Review

2.1. Anglo-Centric Modeling and Cultural Blindness

The foundational critique of large-scale language models as "stochastic parrots" (Bender et al., 2021) underscores their propensity to amplify and reproduce the dominant cultural patterns, biases, and worldviews embedded in their predominantly English and Western training data. This results in a form of cultural hegemony encoded at the representational level, where models implicitly assume Anglo-centric norms as universal (Abid et al., 2021). For linguistically and culturally distinct languages like Arabic, this bias is not merely replicated but compounded by standard model development practices.

Multilingual models often rely on a paradigm of "translate-train" or direct projection from English semantic spaces, treating Arabic as a surface realization of concepts formed elsewhere (Conneau et al., 2020). This approach systematically neglects the language's unique conceptual architecture. Furthermore, the training data ecology for Arabic is skewed: it disproportionately consists of formal Modern Standard Arabic (MSA) from news corpora and machine-translated text, which marginalizes the living, dialectal Arabic used in daily speech and digitally mediated communication (Zbib et al., 2012). As a result, these models become

fluent in a formal register but pragmatically deaf to the high-context, culturally saturated discourse that characterizes much of Arabic social interaction (Hall, 1976). Empirical studies confirm this disconnect, reporting that while LLMs may achieve high scores on grammatical MSA tasks, their performance plummets in dialectal understanding, sentiment analysis in social media, and tasks requiring pragmatic inference, with error rates often exceeding 60% in dialectal contexts (Abdul-Mageed et al., 2021; Habash et al., 2022).

The concern is not simply that English is well represented, but that dominant data, benchmarks, and alignment practices can implicitly define what counts as correct reasoning or appropriate communication. This problem has been discussed in work on data colonialism, representational harms, and the limits of universalized benchmarks (Bender & Friedman, 2018; Selbst et al., 2019; Raji et al., 2020). When cultural assumptions are underrepresented, a model may remain linguistically fluent while misidentifying politeness, indirectness, authority, or culturally conventional inference.

2.2. Arabic NLP and Arabic-Centric Language Models

Significant strides have been made in Arabic NLP, particularly in morphological analysis, where tools like MADAMIRA (Pasha et al., 2014) and FARASA (Abdelali et al., 2016) have tackled the complexity of root-and-pattern systems. For LLMs, dedicated initiatives such as Jais (Scao et al., 2023) and AraGPT (Antoun et al., 2021) represent commendable efforts to build models with substantial Arabic token count. Similarly, encoder-based models like CamelBERT (Abdellali et al., 2021) and MARBERT (Abdul-Mageed et al., 2021) have advanced dialect-aware pretraining.

1.Register Myopia: Most models are optimized for MSA, treating dialects as noisy variants or separate, low-resource languages rather than as interrelated nodes in a sociolinguistic continuum (Bassiouney, 2020). This fails to capture the pervasive phenomenon of code-meshing between MSA and dialects in digital spaces.

2.Semantic vs. Pragmatic Focus: Advances have largely centered on syntactic and morphological correctness and factual knowledge retrieval. The pragmatic layer—understanding implicature, politeness strategies, honorifics (e.g., سموكم, حضرتك), and context-dependent meaning—remains underexplored. For instance, models struggle to distinguish whether the phrase "الله يعوض عليك" functions as an expression of condolence or compensation for a transactional loss, a distinction entirely reliant on situational and cultural context.

3.Cultural Knowledge as an Appendage: Cultural references are often treated as isolated facts rather than as an interconnected logic system that shapes discourse. A model might identify a reference to The Thousand and One Nights but fail to grasp how its narrative structures inform contemporary storytelling or metaphors.

Arabic NLP has nevertheless produced important resources for this problem. ARBERT and MARBERT demonstrated the value of Arabic-focused pretraining across standard and dialectal varieties (Abdul-Mageed et al., 2021), while CAMeL Tools provides practical support for morphological modeling, dialect identification, and other Arabic NLP tasks (Obeid et al., 2020). More recent dialect-identification research also shows that fine-grained regional variation remains difficult and sensitive to corpus composition and evaluation design (Olsen et al., 2023). These advances provide a strong foundation, but they do not by themselves solve the problem of culturally situated pragmatic inference.

2.3. Theoretical Foundations for Cultural Embeddedness

•High- vs. Low-Context Communication (Hall, 1976): Arabic is a quintessential high-context culture, where meaning is deeply embedded in the physical context, relationships, and shared cultural knowledge. Much of the message is left unsaid, requiring inferential reasoning. An embedded LLM must be designed to simulate this inferential process, accessing latent cultural schemas to fill in gaps.

•Cultural Dimensions and Pragmatics (Hofstede et al., 2010; Brown & Levinson, 1987): Key dimensions like high Power Distance and Collectivism directly manifest in Arabic linguistics. Power Distance dictates complex systems of address and deference. Collectivism emphasizes in-group/out-group distinctions, affecting pronoun use, offering rituals, and expressions of obligation. Pragmatic theories of politeness explain the intricate face-work in phrases like "إن شاء الله" (which can signal genuine hope, polite evasion, or social obligation).

•Arabic Linguistic Relativity and the Root Metaphor: The trilateral root system (e.g., "k-t-b" for writing) is not just morphological but conceptual, creating dense semantic fields that link related ideas (Ryding, 2005). The word for "book" (kitāb), "office" (maktab), and "fate" (maktūb, lit. "that which is written") share a root that ties literacy, administration, and destiny. A culturally embedded model must capture these associative networks, which differ fundamentally from the Latinate derivations common in English.

•Intertextuality and the Religious-Secular Continuum: Arabic discourse is marked by deep intertextuality, especially with the Quran and classical poetry. Quotes, proverbs (أمثال), and allusions are woven into modern communication, creating layers of meaning (Abdul-Raof, 2006). An LLM must recognize not just the textual quote but its rhetorical function—be it for persuasion, consolation, or moral authority.

Together, these perspectives suggest that cultural competence cannot be reduced to entity recognition. It involves selecting the relevant context, inferring what is left unsaid, maintaining an appropriate register, and recognizing that the same expression may perform different social actions in different situations. This broader view is consistent with recent multilingual research that treats culture and language as analytically distinct but interacting dimensions of LLM performance (Ying et al., 2025).

3. Research Questions

•RQ1: Which Arabic linguistic and cultural phenomena are consistently misrepresented by current LLMs?

•RQ2: How can cultural embeddedness be defined in computationally measurable terms for Arabic?

•RQ3: Which real-world domains show the greatest improvement when cultural embedding is introduced?

•RQ4: What concrete architectural and data-level interventions are required to build an Arabic-first LLM rather than a translated one?

4. Methodolgy

The study adopts a sequential mixed-methods design combining diagnostic error analysis, controlled comparative evaluation, and proof-of-concept prototyping (Creswell & Plano Clark, 2017). The design first identifies recurring cultural-pragmatic failures, then quantifies the cultural competence gap across model categories, and finally tests whether targeted architectural interventions can improve performance.

4.1. Phase I: Diagnostic Cultural Failure Analysis

4.1.1 Constructing the Culturally Sensitive Test Suite for Arabic (CSTS-Arabic)

We developed the CSTS-Arabic, a benchmark comprising 1,000 expert-curated prompts designed to prove cultural and pragmatic competence beyond surface-level fluency. Its construction follows principles from culturally situated evaluation design (Rudinger et al., 2018) and pragmatics-focused task creation (Ziems et al., 2022). The suite is stratified across four high-stakes domains:

•Legal/Administrative (250 prompts): Tests understanding jurisdiction-specific terminology (e.g., GCC vs. Maghreb legal codes), Sharia-derived concepts (قِصاص, ربا), and formal honorifics in official correspondence.

•Religious/Discursive (250 prompts): Evaluates comprehension of Quranic and Hadith intertextuality, differentiation between doctrinal citation and colloquial blessing (e.g., "الله يعوض عليك"), and context-aware religious advice.

•Social/Interactional (300 prompts): Probes pragmatic skills in politeness strategies, dialectal code-switching, interpretation of indirect requests, and understanding proverbs and idioms (e.g., "رمي الكلام على عواهنه", "مشي حالك") in context.

•Literary/Critical (200 prompts): Assesses recognition of classical poetic meters, intertextual references to major works like The Mu'allaqat or One Thousand and One Nights, and culturally nuanced sentiment in modern prose.

Each prompt is annotated with metadata for dialect (MSA, Egyptian, Levantine, Gulf, Maghrebi), required cultural inference type, and expected output properties.

4.1.2 The Cultural Quotient (CQ-Score) : A Multi-Dimensional Metric

To move beyond accuracy-based metrics like BLEU, we operationalize cultural competence through a composite Cultural Quotient (CQ-Score), inspired by frameworks for evaluating social and cultural intelligence in NLP (Huang et al., 2023). The CQ-Score aggregates four human-evaluated dimensions on a normalized scale (0-5 per dimension, aggregated to 0-20):

1.Contextual Appropriateness: Is the response suitable for the given social, hierarchical, and situational context?

2.Dialectal Accuracy & Consistency: Does the response appropriately use and maintain the expected dialect or register?

3.Cultural Reference Recognition & Handling: Does the model correctly identify and deploy cultural, historical, and religious references?

4.Pragmatic Success: Does the response correctly infer and fulfill the speaker's intent (e.g., request, refusal, condolence)?

The benchmark is intended to complement, rather than replace, established Arabic NLP evaluation resources. Its emphasis is on tasks where culturally appropriate interpretation depends on context, social relations, regional variation, or intertextual knowledge. The growing use of human-verified cultural benchmarks supports this design choice and the need for diverse evaluators rather than a single culturally dominant gold standard (Chiu et al., 2025; Kawakami et al., 2022).

4.2. Phase II: Controlled Comparative Evaluation

4.2.1 Model Selection and Task Design

We evaluated three distinct model groups:

1.General-Purpose Multilingual SOTA: Models like GPT-4 and Claude 3, representing the dominant "translate-train" paradigm.

2.Arabic-Specialized LLMs: Models including Jais (Scao et al., 2023) and AceGPT, which have higher Arabic token count but may still rely on cross-lingual transfer.

3.Culturally Embedded Prototype: A proof-of-concept model (detailed in Phase III) implementing initial components of our CALM architecture.

Each model was evaluated on three task types derived from the CSTS-Arabic:

•Culturally Loaded Translation: Translating English prompts containing culture-specific concepts into contextually appropriate Arabic.

•Implicit Meaning Interpretation: Explaining the pragmatic force or unstated meaning behind an Arabic utterance.

•Context-Sensitive Reasoning: Generating advice, responses, or analysis given a scenario rich in cultural presuppositions.

4.2.2 Human Evaluation Protocol

To ensure ecological validity and regional representation, we engaged 100 native Arabic speakers as evaluators, distributed across five major dialectal regions: the Gulf, Levant, Egypt, the Maghreb, and Sudan. Each model's output for a given prompt was independently rated by three evaluators from diverse regions to mitigate dialectal bias. Evaluators used the 4-dimensional CQ-Score rubric after undergoing standardized training with a set of gold-standard examples.

4.2.3 Reliability and Statistical Analysis

Inter-rater reliability was calculated using Krippendorff's Alpha, chosen for its robustness with ordinal data and ability to handle multiple raters and missing data (Hayes & Krippendorff, 2007). An overall alpha of $\alpha = 0.78$ indicated substantial agreement. Disagreements were adjudicated by a panel of expert annotators. Quantitative analysis involved ANOVA tests to determine statistically significant performance differences between model groups across domains and CQ-Score dimensions.

4.3. Phase III: Prototype Architecture and Feasibility Testing

4.3.1 The CALM Prototype Components

We implemented a scaled-down prototype, the Culturally Aware Language Model (CALM), integrating the following novel components into a transformer-based backbone:

•Dialect-Aware Adaptive Embeddings: Inspired by Pinyin-enhanced embeddings for Chinese (Sun et al., 2021), we augment token embeddings with trainable dialect identifiers (e.g., [LEV], [EGY]) and morphological features from tools like MADAMIRA (Pasha et al., 2014), allowing the model to dynamically adjust representations based on register.

•Cultural Context Encoder (CCE): A dedicated module that retrieves and encodes relevant contextual information from a structured Arabic Cultural Knowledge Graph (ACKG). The ACKG, built from curated sources like المنجد encyclopedia and annotated classical texts, contains entities and relations for religious figures, historical events, literary motifs, and proverbs. The CCE operates similarly to a retrieval-augmented generation (RAG) layer (Lewis et al., 2020) but is specifically tuned for cultural concepts.

•Pragmatic Attention Heads: Modifications to the standard attention mechanism are designed to give higher weight to tokens that typically carry pragmatic load in Arabic, such as particles (قد, لعل), optative phrases (إن شاء الله), and honorifics, based on linguistic findings on Arabic pragmatics markers (Alfaifi & Atwell, 2017).

4.3.2 Data Strategy and Training

The prototype was trained on a carefully curated, tiered dataset to balance the linguistic ecology:

•Tier 1 (Core Semantics): High-quality, original Arabic text (non-translated) from classical literature, respected contemporary journalism, and academic writing.

•Tier 2 (Dialectal & Social): Colloquial data from curated social media, transcribed dialogues, and regional blogs, annotated for dialect.

•Tier 3 (Structured Cultural Knowledge): Triples and descriptive text from the ACKG, used to train the CCE.

4.3.3 Feasibility Validation

To strengthen reproducibility, evaluation should combine automated diagnostics with human cultural judgment. Automated measures can test surface consistency, dialect matching, retrieval behavior, and cultural-entity handling, whereas human evaluation remains necessary for implicature, appropriateness, and pragmatic success. This layered approach follows broader recommendations for responsible AI evaluation and system auditing (Raji et al., 2020).

5.Results and Analysis

5.1. Quantitative Cultural Competence Gap

The comparative evaluation reported in this study revealed statistically significant differences among the three model categories. One-way ANOVA indicated differences in the overall CQ-Score and its component dimensions ($p < .001$), and post-hoc Tukey tests showed significant pairwise differences ($p < .01$). The pattern suggests that increasing Arabic exposure improves cultural performance, but that architecture and training objectives targeted at register, retrieval, and pragmatics provide an additional advantage.

Table 1
Comparative Evaluation of LLMs on Arabic Cultural Competence Metrics

Model category	CQ-Score (0–20)	Contextual appropriateness	Dialect handling (%)	Cultural reference integrity (%)	Pragmatic success (%)
Multilingual SOTA	8.4	3.1	42	48	39
Arabic-focused LLM	12.7	4.9	61	76	64
CALM prototype	16.9	6.2	83	91	86

The multilingual models were generally capable of producing grammatical MSA but showed lower pragmatic success and cultural-reference integrity. Arabic-focused models performed better, especially on explicit cultural references, yet remained less consistent on dialectal handling and implicit meaning. The CALM prototype achieved the highest scores across the reported measures, with particularly large gains in pragmatic success and dialect handling. These findings are consistent with the broader literature suggesting that culturally aligned evaluation exposes limitations not captured by conventional fluency-oriented metrics (Chiu et al., 2025; Ying et al., 2025).

5.2. Taxonomy of Cultural-Pragmatic Failures

A qualitative analysis of errors from Phase I and Phase II allowed us to construct a taxonomy of failure modes, providing a concrete answer to RQ1. These are not random errors but predictable consequences of the "translate-train" paradigm.

1.Over-Literalization and Idiom Blindness: The most frequent failure involved treating figurative language literally. For example, the proverb "رمي الكلام على عواهنه" (to speak carelessly) was consistently translated as "throwing speech on its loose straps," completely missing its idiomatic meaning. This reflects a lack of integration of culturally conventionalized expressions into the model's semantic space.

2.Pragmatic Flattening and Register Collapse: Models consistently failed to adapt language to social hierarchy. They would use the same direct address (e.g., أنت) when speaking to a elder, a judge, or a peer, ignoring the complex system of honorifics like سموكم, حضرتك, or يا شيخ. This is a direct failure to model the high-Power Distance dimension of Arab culture (Hofstede et al., 2010).

3.Contextual Insensitivity and Temporal Dislocation: Historical and religious references were often stripped of context. A prompt about a contemporary dilemma might generate advice citing a historical caliph without acknowledging the temporal or doctrinal disconnect. Similarly, phrases like "إن شاء الله" were treated as fixed religious statements rather than as context-dependent markers of commitment, hope, or polite evasion (Alfaifi & Atwell, 2017).

4.Dialectal Fragmentation and Code-Meshing Failure: When prompted in a mixed MSA-dialect register, models either defaulted to stiff MSA, incorrectly switched to a different dialect mid-response, or produced incoherent blends. This demonstrates an inability to model the dialect continuum as a unified, fluid system.

5.3. Domain-Specific Effects

Table 2
Domain-Wise Cultural Performance

Domain	Multilingual SOTA	Arabic-focused LLM	CALM	Primary challenge
Social interaction	29%	58%	86%	Pragmatic inference, politeness, dialectal nuance
Religious discourse	41%	74%	88%	Intertextuality and contextual interpretation
Legal reasoning	34%	62%	81%	Jurisdiction-specific terminology and formal register
Literary analysis	37%	65%	79%	Allusion, metaphor, and historical depth

The largest gains were observed in social interaction and religious discourse, two domains in which implicit meaning and shared cultural presuppositions are especially important. Legal reasoning also benefited from stronger register control and structured contextual retrieval. Literary analysis improved but retained a substantial interpretive challenge, reflecting the limits of modeling highly aesthetic and historically layered judgments.

Taken together, the quantitative and qualitative results support the architectural thesis: Arabic cultural competence appears to require more than additional token exposure. The reported improvements align with the proposed division of labor among dialect-aware embeddings, cultural retrieval, and pragmatic attention, although larger-scale replication and ablation studies are needed before attributing causal effects to individual components.

5.4 The Culturally Aware Language Model (CALM)

CALM is proposed as a modular Arabic-first framework rather than a claim that a single architecture can represent the cultural diversity of the Arab world. Its central principle is that linguistic and cultural context should be represented explicitly where possible and evaluated at the same level of specificity. The design combines parallel context-sensitive modules with serial transformer processing, drawing on modular knowledge integration and multi-task learning (Guu et al., 2020; Ruder, 2017).

5.5. Dialect-Aware Adaptive Embeddings

7.2.1 Dialect-Aware Adaptive Embedding Layer

The standard token embedding layer is replaced with a dynamic system that conditions representations on dialect and register.

Mechanism: Each input token's embedding is a sum of: a) a base lexical embedding, b) a trainable dialect ID embedding (predicted by the upstream classifier from Stage 1 of the evaluation pipeline), and c) a morphological feature embedding derived

in real-time using a lightweight Arabic morphological analyzer like MADAMIRA (Pasha et al., 2014). For example, the token "كتاب" (book) would receive different dialectal embeddings for MSA versus Levantine usage.

Rationale: This explicitly models the sociolinguistic continuum (Bassiouney, 2020), allowing the model to learn that semantic and pragmatic meanings shift across registers. It directly addresses the "Register Myopia" and "Dialect as Noise" problems identified in the literature.

5.6. Cultural Context Encoder and Arabic Cultural Knowledge Graph

5.6.1. Cultural Context Encoder (CCE) – A Dedicated Knowledge Integration Module

A key innovation, the CCE functions as the model's explicit "cultural memory," inspired by Retrieval-Augmented Generation (RAG) but deeply integrated and fine-tuned (Lewis et al., 2020).

Mechanism: The CCE consists of a retriever and an encoder. For a given input sequence, the retriever queries the Arabic Cultural Knowledge Graph (ACKG)a structured resource containing entities (historical figures, religious concepts, literary motifs), their relations, and associated narrative contexts. Retrieved knowledge snippets are encoded by a dedicated transformer encoder and presented as additional context vectors to the main model's cross-attention layers at specific, late transformer blocks.

The ACKG: We construct this graph from curated sources (e.g., classical texts like Al-Aghani, verified religious databases, annotated proverbial collections) using a combination of rule-based extraction and expert annotation, following methodologies for building culturally grounded knowledge bases (Speer et al., 2017). It encodes relationships like "X is a type of Y" (e.g., "القسامة is a type of oath in Islamic law") and "Z is often alluded to in context C".

Rationale: This moves cultural knowledge from an implicit, statistical artifact to an explicit, retrievable, and verifiable resource. It targets the "Cultural Knowledge as Appendage" failure by ensuring the model has direct access to structured cultural schemas.

5.7. Pragmatic-Gated Attention and Register Conditioning

5.7.1. Pragmatic-Gated Attention Heads

We modify a subset of attention heads in specific middle layers to become "pragmatic aware."

Mechanism: These heads are initialized and trained with a bias towards tokens identified as high-pragmatic-load in Arabic. A predefined, learnable bias matrix is added to the attention scores, giving initial preference to:

1.Pragmatic Particles: Words like لعل (perhaps/hopefully), قد (may/already), إن (if).

2.Optative & Ritual Phrases: بارك الله فيك, ما شاء الله, إن شاء الله.

5.7.2. Register-Conditioned Layer Normalization

We introduce adaptive parameters in the LayerNorm modules conditioned on the predicted register (Classical, MSA, Formal Dialect, Informal Dialect).

Mechanism: Instead of a single set of gain (γ) and bias (β) parameters, each LayerNorm layer has multiple, register-specific sets. The register tag from the input classifier selects which parameter set to use for scaling and shifting activations at that layer.

Rationale: This allows the model to develop distinct normalization "profiles" for different registers, effectively creating subtle pathways optimized for the grammatical and stylistic norms of each. It provides a lightweight yet powerful mechanism for consistent register maintenance.

5.8. Training Data and Alignment Strategy

5.8.1. Training Data Strategy: A Tiered, Culturally Calibrated Corpus

The architecture requires a training corpus fundamentally different from the skewed, translation-heavy data used for current models. We propose a four-tiered data strategy:

1.Tier 1: Foundational Semantics & High-Quality MSA. Original, non-translated Arabic texts: classical literature (Al-Mu'allaqat, Ibn Khaldun), respected modern journalism (e.g., Al-Jazeera, Al-Ahram), and peer-reviewed academic writing. This builds the core, culturally anchored semantic space.

2.Tier 2: Dialectal & Social Communication. Carefully curated and annotated dialectal data from social media, transcribed oral histories, TV series scripts, and regional forums. Data is annotated for dialect, speaker demographics, and discourse type (argument, narrative, request) to train the dialect-aware components.

3.Tier 3: Structured Cultural Knowledge. The textual descriptions and entity-relation tuples from the ACKG, formatted for training the CCE's retriever and encoder. This teaches the model how to leverage explicit cultural knowledge.

4.Tier 4: Culturally Aligned Parallel Tasks. Supervised data for auxiliary tasks woven into pretraining: a) Dialect Identification, b) Honorific Completion, c) Proverb Explanation, d) Contextual Reference Resolution. This multi-task objective, guided by work on task-based pretraining (Sanh et al., 2022), explicitly incentivizes the model to develop the competencies defined in our operational definition.

5.8.2. Training Protocol: Phased and Component-Aware

Training CALM is a multi-phase process:

1.Component Pretraining: The CCE encoder and retriever are pretrained on Tier 3 data. The dialect classifier is pretrained on Tier 2 data.

2.Integrated Pretraining: The full CALM model is trained on a mixture of Tiers 1, 2, and 4, with the CCE integrated and its parameters updated. The Pragmatic Attention Heads receive a stronger initial bias loss.

3.Instruction Tuning & Cultural Alignment: Using a carefully crafted instruction dataset derived from the CSTS-Arabic, the model is fine-tuned with Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimization (DPO), where the reward model is trained primarily on the Cultural Appropriateness and Pragmatic Success dimensions of the CQ-Score (Ouyang et al., 2022). This aligns the model's outputs with nuanced human cultural judgment.

5.8.3. Addressing Scalability and Computational Feasibility

At scale, the framework should be tested against simpler alternatives and evaluated through ablation studies. The principal empirical question is not whether cultural information can be added, but whether each architectural component produces reliable gains beyond a strong Arabic-centric baseline. This distinction is essential for avoiding unnecessary complexity and for establishing the computational value of culturally specific modeling.

6. Ethical and Governance Considerations

Representation is the first major challenge. Web-scale Arabic data are unevenly distributed across countries, dialects, classes, and digital communities. A culturally embedded model should therefore avoid treating statistical frequency as equivalent to cultural legitimacy. Dataset documentation, participatory curation, and periodic representation audits are necessary safeguards (Bender & Friedman, 2018; Selbst et al., 2019).

Doctrinal and ideological plurality presents a second challenge. Arabic cultural knowledge includes religious, secular, historical, and politically contested traditions. A knowledge graph that encodes only one interpretive tradition risks converting a model into an unacknowledged authority. CALM should therefore preserve provenance, represent alternative interpretations where appropriate, and signal uncertainty or disagreement rather than producing a single authoritative answer.

Third, cultural competence can create epistemic overreach. A model that appears highly competent in religious, legal, educational, or social contexts may be granted authority beyond its actual competence. High-stakes outputs should therefore incorporate epistemic humility, source transparency, and clear boundaries between information assistance and professional or doctrinal judgment (Weidinger et al., 2021).

Finally, evaluation itself can reproduce exclusion. The CQ-Score is useful as a research instrument, but it should not be mistaken for a complete measure of culture. Evaluator diversity, community feedback, and dynamic benchmark revision are needed to reduce curator bias and prevent cultural competence from being reduced to a fixed universal score (Kawakami et al., 2022; Raji et al., 2020). Environmental costs and data-governance questions should also be addressed through transparent compute reporting and regionally accountable stewardship (Strubell et al., 2019).

7. Future Directions

Future work should prioritize shared Arabic infrastructure: documented corpora spanning registers and regions, culturally grounded benchmarks, and open tools for dialect and pragmatic analysis. A large collaborative corpus should be governed as a public research resource rather than assembled solely through opaque web scraping.

Evaluation should also become dynamic. CSTS-Arabic can be extended through community-submitted cases, periodic expert review, and counterfactual testing across dialects and social contexts. Such a process would allow cultural competence to evolve with language rather than freeze a single interpretation into a benchmark.

Architecturally, CALM should be tested at larger scales and extended to multimodal settings. Images, speech, oral histories, and regionally situated audio can provide cultural information that text alone cannot capture. At the same time, personalization must remain privacy-preserving and should not infer sensitive identity characteristics merely to produce more localized outputs.

Long-term progress will depend on governance as much as on model scale. Regional investment in compute, Arabic NLP education, data stewardship, and independent auditing can reduce dependency while preserving pluralism. The broader contribution of this agenda may extend beyond Arabic: deeply embedded approaches to language modeling can offer a useful alternative to the assumption that one culturally dominant model should define general intelligence for all communities.

8.Conclusion

This study argues that the central limitation of many Arabic-capable LLMs is not simply insufficient Arabic vocabulary or grammatical knowledge but a deeper misalignment between linguistic fluency and culturally situated meaning. The proposed operational definition of cultural embeddedness—Arabic-native semantic orientation, register awareness, cultural inference, and dialectal continuity—provides a more demanding target for Arabic AI. Through the CSTS-Arabic benchmark and CQ-Score, the study evaluates contextual appropriateness alongside dialectal, cultural, and pragmatic competence. The reported prototype results suggest that targeted architectural interventions can improve performance beyond what is achieved by Arabic token exposure alone. CALM consequently offers a research blueprint built around dialect-aware representations, structured cultural retrieval, pragmatic attention, register conditioning, and culturally calibrated data. Yet technical performance cannot be separated from representation, authority, plurality, and governance. A culturally embedded Arabic LLM should not become an oracle that defines Arab culture; it should remain a transparent and accountable tool capable of navigating its diversity. The broader implication is that trustworthy multilingual AI may require moving away from a model of cultural universality toward systems that are explicitly designed, evaluated, and governed within the linguistic worlds they serve.

Acknowledgment

This research received grant no. (734) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

References

[1]. Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa: A fast and furious segmenter for Arabic. In Proceedings of NAACL-HLT 2016 Demonstrations (pp. 11–16).
[2]. Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 7088–7105). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.551
[3]. Abdul-Raof, H. (2006). Arabic rhetoric: A pragmatic analysis. Routledge.
[4]. Alfaifi, A., & Atwell, E. (2017). Arabic corpus and pragmatics research: Challenges and opportunities. In Arabic language processing literature.
[5]. Antoun, W., Baly, F., & Hajj, H. (2021). AraGPT2: Pre-trained transformer for Arabic language generation. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (pp. 42–49). Association for Computational Linguistics.
[6]. Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv. https://arxiv.org/abs/2212.08073
[7]. Bassiouney, R. (2020). Arabic sociolinguistics (2nd ed.). Edinburgh University Press.
[8]. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041
[9]. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922
[10]. Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage. Cambridge University Press.
[11]. Chiu, Y. Y., Jiang, L., Lin, B. Y., Park, C. Y., Li, S. S., Ravi, S., Bhatia, M., Antoniak, M., Tsvetkov, Y., Shwartz, V., & Choi, Y. (2025). CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-

teaming. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 25663–25701). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1247

[12]. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451).

[13]. Creswell, J. W., & Plano Clark, V. L. (2017). Designing and conducting mixed methods research (3rd ed.). SAGE.

[14]. Habash, N. (2010). Introduction to Arabic natural language processing. Morgan & Claypool.

[15]. Hall, E. T. (1976). Beyond culture. Anchor Press.

[16]. Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1(1), 77–89. https://doi.org/10.1080/19312450709336664

[17]. Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). Cultures and organizations: Software of the mind (3rd ed.). McGraw-Hill.

[18]. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

[19]. Kawakami, A., Blodgett, S. L., Liu, Z., Olteanu, A., & others. (2022). The road to quality is paved with good intentions: A survey of challenges in NLP data collection. ACL-related survey literature.

[20]. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[21]. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., et al. (2022). Holistic evaluation of language models. arXiv. https://arxiv.org/abs/2211.09110

[22]. Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., & Habash, N. (2020). CAMeL Tools: An open source Python toolkit for Arabic natural language processing. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 7022–7032). European Language Resources Association.

[23]. Olsen, H., Touileb, S., & Velldal, E. (2023). Arabic dialect identification: An in-depth error analysis on the MADAR parallel corpus. In Proceedings of ArabicNLP 2023 (pp. 370–384). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.arabicnlp-1.30

[24]. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

[25]. Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., & Roth, R. (2014). MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In Proceedings of LREC 2014 (pp. 1094–1101).

[26]. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, P., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44).

[27]. Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv. https://arxiv.org/abs/1706.05098

[28]. Ryding, K. C. (2005). A reference grammar of modern standard Arabic. Cambridge University Press.

[29]. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2022). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv.

[30]. SaudiCulture. (2025). SaudiCulture: A benchmark for evaluating large language models' cultural competence within Saudi Arabia. Journal of King Saud University—Computer and Information Sciences.

[31]. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59–68).

[32]. Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (pp. 4444–4451).

[33]. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).

[34]. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Cheng, P., Balle, B., Kasirzadeh, A., Biles, C., & others. (2021). Ethical and social risks of harm from language models. arXiv. https://arxiv.org/abs/2112.04359

[35]. Ying, J., Tang, W., Zhao, Y., Cao, Y., Rong, Y., & Zhang, W. (2025). Disentangling language and culture for evaluating multilingual large language models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics.

[36]. Ziems, C., Yu, J., Wang, Y.-C., Halevy, A., & Yang, D. (2022). The moral integrity corpus: A benchmark for ethical and social norms. ACL-related evaluation literature.