Linguistic and Philosophical Investigations
ISSN: 1841-2394, e-ISSN: 2471-0881
Vol 25 (1), 2026
pp.939-948

Automated Semantic Flattening: How Large Language Models Threaten the Cognitive Architecture of Arabic beyond Digital Extinction

Dr. Saad RAOUANE
University of Algiers 2, Algeria.
saad.raouane@univ-alger2.dz

Abstract
While existing literature on digital linguistics frequently warns against "Digital Language Death"—framed as a quantitative scarcity of Arabic content online—this study argues that the primary threat to the Arabic language in the era of Artificial Intelligence is not extinction, but qualitative semantic degradation. We introduce the concept of "Automated Semantic Flattening" (ASF), defining it as the algorithmic erosion of the rich, multi-layered derivational root system (Ishtiqāq) unique to the Arabic language when processed through Large Language Models (LLMs). Built primarily on Indo-European structural and computational logics, transformer-based architectures tokenize Arabic texts in a manner that strips lexical roots of their deep metaphoric, spatial, and philosophical nuances. Through a qualitative comparative methodology analyzing algorithmic outputs against traditional lexicographical frameworks, this paper demonstrates how LLMs favor surface-level, unidimensional equivalents, thereby reconstructing Arabic into a structural shell for Western semantic frameworks. The study concludes by proposing a framework for "Digital Linguistic Sovereignty," advocating for AI architectures natively designed around the root-and-pattern morphology rather than mere computational translation.
Keywords: Arabic NLP, Large Language Models (LLMs), Automated Semantic Flattening, Derivational System (Ishtiqāq), Digital Language Death, Cognitive Architecture.

Received: 13 July 2026 **Revised:** 17 August 2026 **Accepted:** 28 August 2026

Introduction:

The Arabic language occupies a unique ontological and structural position within world languages. Unlike Indo-European languages, which largely rely on analytical or agglutinative lexical structures, Arabic operates on a deeply interconnected derivational root-and-pattern system (Ištinbāṭ & Ishtiqāq). This system is not merely a linguistic tool for word formation; it serves as a sophisticated cognitive architecture that embeds philosophy, spatial relations, and existential nuances directly into a three- or four-letter consonant root (al-Jidhr). For over fifteen centuries, this structural core has provided Arabic with an exceptional degree of internal resilience, allowing it to adapt to socio-cultural shifts without sacrificing its semantic integrity (Al-Fassi Fassi, 2013.)

However, as human communication shifts decisively into the digital sphere, the survival of languages is no longer evaluated solely through historical longevity or oral preservation, but through their integration into computational systems. In early digital linguistics, scholarly concern centered primarily around the concept of "Digital Language Death" (Kornai, 2013). This framework posits that languages face functional extinction if they lack sufficient textual

presence, digital infrastructure, or Natural Language Processing (NLP) resources online. Consequently, the discourse surrounding Arabic in the digital age has been dominated by a quantitative obsession: increasing the volume of Arabic web content, expanding digital corpora, and improving surface-level tokenization for machine translation (Habash, 2010; Zuhaira & Magdy, 2019.(

1.1.The Shift from Quantitative Scarcity to Qualitative Degradation

While the quantitative representation of Arabic has indeed expanded through algorithmic initiatives and multilingual platforms, this paper argues that the orthodox "Digital Language Death" paradigm fails to capture the subtle, more insidious threat emerging in the era of Large Language Models (LLMs). The primary risk facing Arabic today is not computational extinction, but qualitative semantic degradation.

Modern artificial intelligence architectures—most notably Transformer-based models such as GPT, Llama, and Claude—were fundamentally engineered around the structural and logical conventions of Indo-European languages (Bender et al., 2021). When these systems process Arabic, they do so by decomposing text into sub-word units (Tokenization) based on statistical frequencies rather than morphological roots. This process forces the non-linear, multi-dimensional root structure of Arabic into a linear, additive computational mold

Traditional Arabic Logic

(ر-ح-م)] Root: r-ḥ-m [

Raḥmah (Compassion) Raḥim (Womb) Marḥamah (Empathy)
Spiritual Dimension] [Physical Reality] [Ethical Duty]

Algorithmic Tokenization (LLMs)

[Token 1: ر] + [Token 2: حم] + [Token 3: ة]

* Statistical association mapped to closest English concept :

" Mercy" / "Sympathy" (Unidimensional Equivalent)

As illustrated above, this structural mismatch triggers a phenomenon we define as Automated Semantic Flattening (ASF): the systematic algorithmic stripping of the contextual, metaphorical, and philosophical layers inherent in Arabic lexical roots, reducing them to unidimensional, surface-level equivalents that align with dominant Western conceptual frameworks.

1.2. Problem Statement and Research Questions

When Generative AI systems generate, translate, or summarize Arabic content, they do not "think" within the Arabic cognitive matrix. Instead, they perform a reverse conceptual translation: mapping Arabic lexical items onto an underlying semantic graph structured by English-language training data. Over time, as AI-generated text increasingly saturates the digital ecosystem, this results in an algorithmic feedback loop. Human users, particularly younger generations, consume and internalize this "flattened" Arabic, leading to an erosion of linguistic depth and a loss of conceptual diversity.

To investigate this phenomenon, this study addresses the following primary research questions:

RQ1: How do current sub-word tokenization algorithms in Transformer-based LLMs structurally disrupt the morphological integrity of the Arabic derivational root system (Ishtiqāq)?

RQ2: In what ways does "Automated Semantic Flattening" manifest when processing complex Arabic conceptual vocabulary compared to classical lexicographical frameworks?

RQ3: What framework can be established to achieve "Digital Linguistic Sovereignty," ensuring AI systems preserve the native cognitive architecture of Arabic rather than forcing it into a structural proxy for Indo-European logic?

1.3. Significance and Structure of the Study

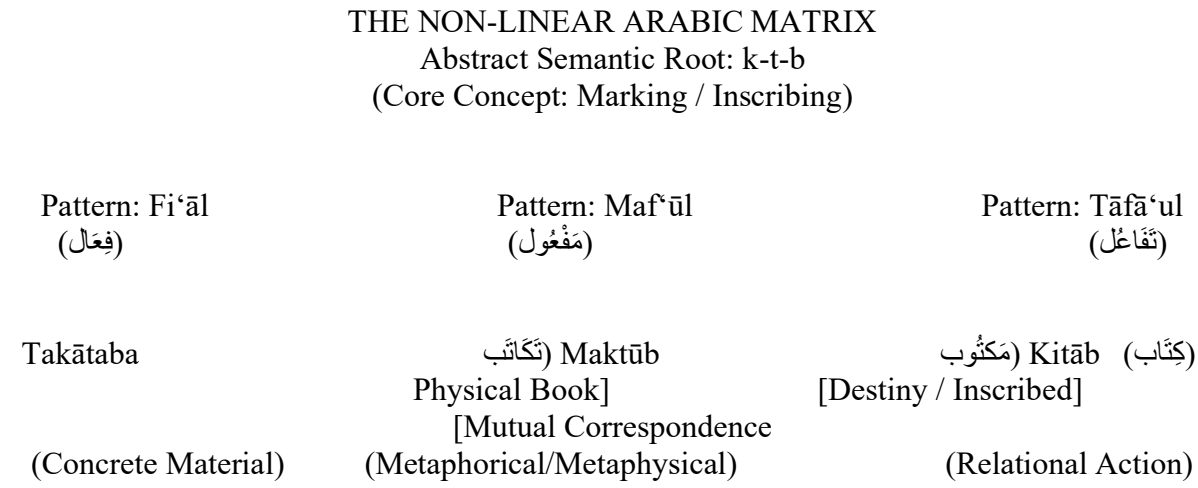
By bridging the gap between computational linguistics and cognitive semantics, this study offers

a novel theoretical critique that moves beyond the simplistic binary of "digital survival vs. extinction." It shifts the paradigm toward semantic security and cognitive preservation.
The remainder of this paper is structured as follows: Section 2 establishes the theoretical framework, contrasting the Arabic root-and-pattern morphology with sub-word tokenization models. Section 3 presents a qualitative empirical analysis of semantic flattening across selected conceptual domains. Section 4 discusses the broader implications for cognitive architecture and linguistic sovereignty, while Section 5 concludes with strategic recommendations for native AI architecture design.

2. Theoretical Framework

To evaluate the mechanical origins of Automated Semantic Flattening (ASF), it is essential to establish a comparative framework that contrasts the cognitive-linguistic structure of the Arabic language with the computational design of Transformer-based Large Language Models. This section delineates the ontological collision between the non-linear, derivational architecture of Arabic and the linear, token-based probabilistic logic of modern AI architectures.

2.1 The Arabic Cognitive Matrix: Derivational Root System (Ishtiqāq)The architecture of the Arabic language is fundamentally rooted in a non-concatenative morphology. Unlike European languages that build words primarily through linear concatenation (attaching prefixes and suffixes to a stem), Arabic word formation relies on a two-tiered system: a consonant Root (al-Jidhr) and a prosodic Pattern (al-Wazn) (Holes, 2004; Hammad, 2021).The root—typically comprised of three consonants ($C_1-C_2-C_3$)—serves as an abstract semantic core containing a multidimensional semantic field rather than a single fixed definition. The pattern acts as a structural template that imparts specific grammatical categories, voices, and semantic nuances to that core.



This structural interconnectedness creates what cognitive linguists refer to as a Semantic Network Effect (Hussam al-Din, 2000). A speaker of Arabic processing the word Maktūb (written/destined) does not perceive it as an isolated lexical item; the neurological cognitive architecture simultaneously activates the entire network of [k-t-b], referencing Kataba (action of writing), Kātib (writer), and Maktaba (library). This yields a rich, multi-layered cognitive experience where spatial, temporal, physical, and spiritual dimensions coexist within a single utterance.2.2 Transformer Architectures and Sub-Word Tokenization LogicIn contrast to the non-linear, root-based logic of Arabic, contemporary Generative AI architectures rely on Transformer neural networks (Vaswani et al., 2017) that process text linearly. Modern Large Language Models (LLMs) do not process human languages at the level of semantic roots, words, or morphemes; instead, they operate on Tokens—sub-word text fragments determined purely through statistical frequency optimization algorithms, such as Byte-Pair Encoding (BPE) or WordPiece (Sennrich et al., 2016).
Sub-word tokenization algorithms were developed primarily for Indo-European languages, where words can be neatly partitioned into continuous character substrings. When applied to

Arabic, these statistical tokenizers fail to recognize the discontinuous nature of the root-and-pattern system.

SUB-WORD TOKENIZATION DISRUPTION

Input Word: "وسيستكتبونها" (And they will ask her to write it)
Root Core: [k-t-b] (ك-ت-ب)
Actual Algorithmic Segmentation (Byte-Pair Encoding):
[Token 1: و] + [Token 2: سيست] + [Token 3: كتب] + [Token 4: ونها]
Problem: The prefix "سيست" (indicating future intent & request) and | | suffix "ونها" (plural subject & feminine object) are split arbitrarily. | | The root [كتب] is treated as an isolated string rather than an active | | morphological kernel regulating the surrounding prefixes.

Because BPE algorithms break Arabic words into arbitrary character blocks based on surface frequency rather than morphological boundaries, the semantic integrity of the root is structurally severed (Habash, 2010; Al-Jadidi, 2019). The model treats the substring as a discrete statistical unit, discarding the implicit conceptual connections that tie it back to its broader semantic matrix.2.3 The Mechanism of Automated Semantic Flattening (ASF)When an LLM processes, generates, or translates Arabic text, the tokenized input is projected into a high-dimensional vector space (embedding space). In state-of-the-art multilingual LLMs, this embedding space is overwhelmingly shaped by English-language training data, which constitutes the majority of the pre-training corpus (Bender et al., 2021).Consequently, the multi-layered semantic field of an Arabic root undergoes a three-stage compression process:Morphological Fragmentation: Sub-word tokenization strips the word of its root-pattern relational structure, isolating the surface form from its semantic network.Cross-Lingual Alignment Vectorization: The isolated word token is mapped onto a semantic vector aligned with the closest conceptual equivalent in English (e.g., mapping Raḥmah directly to "mercy").Monosemic Reduction (Flattening): The model discards the peripheral, metaphorical, and spiritual layers of the Arabic root (e.g., the intrinsic connection between Raḥmah and Raḥim—the physical womb) to minimize prediction perplexity, selecting the most statistically frequent, unidimensional translation.

THE THREE-STAGE FLATTENING PIPELINE

[Arabic Root Network] ---> (1) Morphological Fragmentation
(Multi-dimensional) - Sub-word BPE Tokenization
↓
(2) Vector Alignment
- Mapping to English-centric Space
↓
(3) Monosemic Reduction
- Output: Surface/Flat Concept

Through this pipeline, generative AI does not merely translate language; it systematically re-engineers Arabic conceptual structures into a structural proxy for Indo-European logic.

3. Qualitative Empirical Analysis: Case Studies of Automated Semantic FlatteningTo systematically demonstrate the mechanics of Automated Semantic Flattening (ASF) in current Large Language Models, this section presents a qualitative comparative analysis. We evaluate the computational processing of selected core Arabic conceptual domains, contrasting the traditional multi-layered lexicographical definitions found in classical lexicons (such as Lisān al-ʻArab and Al-Muʻjam al-Wasīṭ) with the outputs produced by state-of-the-art Generative AI models when translating, summarizing, or generating text around these concepts.The empirical analysis focuses on three primary linguistic domains where the divergence between native Arabic cognitive architecture and algorithmic processing is most pronounced: Spatio-Philosophical ConceptsAffective and Existential StatesSocio-Ethical and Relational Terms.

3.1 Spatio-Philosophical Domain: The Root [g-h-b] and [f-t-r]In classical Arabic, physical space, motion, and existential transformation are intrinsically linked through root morphology. When processed by Transformer-based models, these multi-layered concepts are frequently collapsed into static, unidimensional English equivalents.A prominent example is the term al-Ghayb (الغيب), derived from the tri-consonantal root [g-h-b] (gh-y-b). In traditional Arabic lexicography and

philosophical discourse, al-Ghayb does not merely signify "the unknown." It encompasses a complex spatial-ontological dimension: that which is inherently concealed from sensory perception yet actively influences empirical reality, maintaining an intrinsic relational link to al-Shahādah (the witnessed world) (Al-Fassi Fassi, 2013)

When queried across prominent LLMs, the term is systematically mapped directly to surface-level lexical tokens in English, discarding its structural and philosophical dimensions.

Table 1: Comparative Semantic Analysis of "Al-Ghayb" (الغيب).

Analysis Dimension	Classical Arabic Lexicographical Matrix	Algorithmic LLM Output & Vector Mapping
Primary Root Core	[g-h-b](Absence, concealment,active unseen reality)	Tokenized as discrete substring [يب] + [الغ]
Semantic Scope	Multidimensional: Spatial (hidden realm), Metaphysical (transcendent reality), Epistemic (limit of human perception)	Monosemic: "Unseen" / "Invisible" / "The Future"
Cognitive Effect	Preserves the ontological tension between sensory experience and transcendent existence	Flattens the concept into a simple binary: visible vs. invisible
Derivational Links	Activated network: Ghā'ib (absent), Ghabah (dense forest/hidden), Ghaybūbah (coma/altered state)	Severed. The term is treated as an isolated standalone noun aligned with Western terms.

A similar flattening occurs with the term al-Fiṭrah (الفطرة), derived from [f-ṭ-r]. While classical dictionaries define the root as "the act of cleaving open, splitting, or bringing into existence with an innate directional orientation," generative models overwhelmingly reduce al-Fiṭrah to "human nature" or "instinct." This reduction strips the concept of its dynamic, creative action (splitting open) and replaces it with a static, biological framework derived from Western psychology (Hussam al-Din, 2000).

3.2 Affective and Existential Domain: The Root [r-ḥ-m] and [ḥ-z-n]Emotional and existential vocabulary in Arabic relies heavily on biological and physical root metaphors to convey complex psychological states. Sub-word tokenization and cross-lingual vector space alignments routinely strip these metaphors of their physical grounding.Consider the root [r-ḥ-m] (r-ḥ-m), specifically the noun al-Raḥmah (الرحمة). Classical Arabic lexicography embeds al-Raḥmah directly in the physical, spatial reality of al-Raḥim (الرَّحِم - the maternal womb). Thus, compassion in Arabic is not an abstract sentiment; it is a visceral, protective, and life-nourishing bond modeled on the developmental environment of the womb.

Table 2: Comparative Semantic Analysis of "Al-Raḥmah" (الرحمة)

Analysis Dimension	Classical Arabic Lexicographical Matrix	Algorithmic LLM Output & Vector Mapping
Root Metaphor Grounding	Al-Raḥim (Womb) - Physical, spatial, organic connection	Abstract lexical vector aligned with "Mercy" / "Pity" / "Sympathy"
Semantic Directionality	Outward, generative, protective, non-hierarchical duty	Inward, top-down (a superior granting clemency to an inferior)
Relational Depth	Mutual obligation tied to origin and biological/existential lineage	Transactional state or passive emotional disposition
ASF Mechanism	Full activation of biological root network	Monosemic reduction to Western judicial/religious terms ("Mercy")

When LLMs process al-Raḥmah, the statistical alignment with the English token "mercy" alters the cognitive directionality of the concept. "Mercy" in Indo-European frameworks historically implies a top-down relationship (a judge showing mercy to a offender, or a king to a subject). In

contrast, al-Raḥmah operates as a generative, mutual bond of care. By substituting "mercy" for al-Raḥmah, the model shifts the internal logic of the Arabic text from an organic, relational framework to a hierarchical, judicial one.

A parallel distortion is observed in the root [ḥ-z-n] (al-Ḥuzn). While classical lexicography distinguishes between Ḥuzn (roughness of terrain, psychological weight), Kamad (suppressed grief), and Asaf (grief combined with outrage), generative AI architectures consistently normalize all three terms into the generic English equivalent "sadness," eliminating the qualitative nuances embedded within the derivational forms (Hammād, 2021).

3.3 Socio-Ethical Domain: The Root [ʻ-d-l] and [w-s-ṭ]Socio-ethical vocabulary presents the most critical vulnerability to Automated Semantic Flattening, as AI models frequently project modern socio-political definitions onto classical Arabic root structures.

The term al-Adl (العدل), derived from [ʻ-d-l], provides a clear illustration of this phenomenon. In classical Arabic dictionaries, the primary root meaning of [ʻ-d-l] is "to straighten, equalize, or hold two things in exact equilibrium," as well as "to deviate from wrong to right." It is an objective, mathematical, and spatial balance before it is a moral or legal judicial system.

Table 3: Comparative Semantic Analysis of "Al-ʻAdl" (العدل)

Analysis Dimension	Classical Arabic Lexicographical Matrix	Algorithmic LLM Output & Vector Mapping
Primary Root Core	[ʻ-d-l] (Equilibrium, counter-balancing equal weights)	Aligned directly with modern English tokens: "Justice" / "Fairness"
Semantic Framework	Spatial and ontological balance inherent in the structure of the cosmos	Socio-political rights, legal procedural fairness, and institutional equity
Associated Concepts	Iʻtidāl (temperance/balance), Adīl (equal weight on a beast of burden)	Judicial courts, civil rights, law enforcement, political rights
ASF Impact	Reduces cosmic/ontological balance to institutional legalities	Re-articulates Arabic moral philosophy through modern Western legal theory

When Generative AI systems interpret or generate discourse around al-ʻAdl, they consistently map the term to modern Western notions of "procedural justice" or "legal fairness." The foundational Arabic idea that al-ʻAdl is an inherent, ontological state of balance across nature is replaced by an institutional, legalistic concept (Al-Jadidi, 2019).

Similarly, al-Wasaṭiyyah (الوسطية), from [w-s-ṭ], is routinely translated by current models as "moderation" or "compromise." In classical lexicography, Wasaṭ designates "the geographical or structural center of a circle"—the point of maximum strength, elevation, and protection—rather than a passive compromise between two extreme positions. The algorithmic translation into "moderation" flattens a concept of excellence and structural strength into one of political neutrality or hesitancy.

3.4 Synthesis of Empirical Findings

The qualitative findings across these three domains reveal a consistent pattern of semantic erosion. When Large Language Models process complex Arabic vocabulary, the structural mismatch between sub-word tokenization and non-linear root morphology produces three repeatable outcomes:

1-Physical De-grounding: The severance of abstract terms from their concrete, physical root metaphors (e.g., separating Raḥmah from the physical womb).

2-Concept Alignment Bias: The forced mapping of Arabic semantic networks onto English conceptual vectors (e.g., mapping ʻAdl strictly to institutional "justice").

3-Loss of Lexical Granularity: The collapse of highly specific, nuanced root derivatives into broad, generic category labels (e.g., reducing multiple distinct affective states to "sadness" or "anger").

This empirical reality confirms that the threat to Arabic in the AI ecosystem is not a lack of representation, but an algorithmic restructuring that gradually strips the language of its native

cognitive architecture.

4. Discussion: Cognitive Implications and Digital Linguistic Sovereignty

The qualitative findings presented in Section 3 confirm that the collision between sub-word tokenization architectures and Arabic root morphology results in systematic Automated Semantic Flattening (ASF). This section shifts from empirical observation to critical evaluation, examining the broader cognitive, cultural, and epistemological ramifications of this phenomenon. Furthermore, it outlines a conceptual framework for Digital Linguistic Sovereignty, establishing the prerequisite conditions for preserving native Arabic cognitive architecture within next-generation artificial intelligence systems.

4.1 The Cognitive Feedback Loop: Algorithmic Saturation and Epistemic Erosion

The long-term danger of Automated Semantic Flattening extends far beyond computational inaccuracy; it poses a direct challenge to the cognitive architecture of future Arabic speakers. As artificial intelligence models become deeply embedded in education, automated content creation, search engines, and digital translation, AI-generated text is rapidly saturating the Arabic digital ecosystem. This saturation establishes a self-reinforcing cognitive feedback loop.

In traditional linguistic transmission, younger generations acquire native cognitive structures through exposure to rich, derivational language networks present in literature, classical scholarship, and oral tradition (Al-Fassi Fassi, 2013). However, when the primary medium of daily textual consumption becomes AI-generated or AI-filtered content, users are continuously exposed to "flattened" Arabic—a language characterized by monosemic word choices, English-aligned conceptual structures, and linearized syntax.

Over time, this exposure induces what cognitive linguists term **Epistemic Erosion:**

1-Atrophy of Root Recognition: Users gradually lose the intuitive ability to trace derived vocabulary back to its core consonant root, diminishing their capacity to decode complex, unfamiliar words independently.

2-Conceptual Homogenization: The distinctive philosophical and existential perspectives embedded within Arabic vocabulary (such as the cosmic equilibrium of ʻAdl or the generative physical care of Raḥmah) are replaced by homogenized, globally standardized definitions derived from Anglo-American discourse.

3-Loss of Metaphorical Depth: The rich spatial and physical metaphors that underpin classical Arabic abstract thought are discarded in favor of literal, unidimensional equivalents, flattening the expressive capability of the language.

Consequently, Arabic runs the risk of becoming a "structural shell"—a language that uses Arabic script and vocabulary on the surface, but operates entirely on the conceptual logic and vector geometry of Western machine learning frameworks (Bender et al., 2021).

4.2 Beyond Digital Extinction: A Critique of Existing AI Paradigms

For the past two decades, international and regional initiatives aimed at supporting Arabic in the digital age have operated under the assumption that the primary goal is quantitative expansion. Projects have focused overwhelmingly on increasing the volume of digitized pages, building larger training corpora, and scaling up multilingual LLM performance metrics (Habash, 2010; Zuhaira & Magdy, 2019).

The analysis presented in this study demonstrates that this quantitative paradigm is fundamentally insufficient—and potentially counterproductive. Merely scaling up training data using conventional Transformer architectures does not resolve Automated Semantic Flattening; in fact, it accelerates it.

Standard multilingual **LLMs** optimize for global perplexity reduction across vast datasets dominated by English text. When massive Arabic corpora are scraped from the modern web—much of which already consists of translated content or algorithmically flattened prose—the model simply reinforces and institutionalizes these semantic distortions at a larger scale (Kornai, 2013; Al-Jadidi, 2019). Therefore, solving the qualitative degradation of Arabic requires a paradigm shift: moving away from mere data accumulation toward structural architectural reform.

4.3 The Framework for Digital Linguistic Sovereignty

To counteract Automated Semantic Flattening and protect the cognitive integrity of Arabic, this

study proposes the conceptual framework of Digital Linguistic Sovereignty (DLS). DLS posits that true linguistic preservation in the digital age requires technological autonomy over how a language's underlying logic, morphology, and semantics are modeled computation-wise.Digital Linguistic Sovereignty rests on three core technical and theoretical pillars:

Pillar 1: Native Morphological Tokenization Architectures

Current sub-word tokenization algorithms (such as Byte-Pair Encoding) must be replaced or augmented with native morphological tokenizers specifically engineered around the Arabic root-and-pattern system. Rather than partitioning words into arbitrary character substrings based on raw frequency, native tokenization must explicitly decompose text into two parallel structural streams:

The Root Vector Stream: Capturing the core semantic matrix of the tri-consonantal root (C1-C2-C3).

The Pattern Vector Stream: Capturing the prosodic template (al-Wazn) regulating voice, aspect, and derived grammatical nuances.

By processing roots and patterns as distinct yet linked computational entities, neural networks can preserve the non-linear semantic network of Arabic, preventing the morphological fragmentation that causes semantic flattening.

Pillar 2: Lexicographically Grounded Pre-Training Corpora

Language model pre-training must shift away from indiscriminate web-scraping toward curated, lexicographically annotated corpora. Training datasets must incorporate classical lexicographical frameworks (Lisān al-'Arab, Taj al-'Arūs) and historical text archives that explicitly map the metaphorical and physical roots of vocabulary. Weighting these foundational texts during pre-training ensures that the model's internal vector space retains the multi-dimensional semantic depth inherent in native Arabic concepts.

Pillar 3: Cultural-Semantic Evaluation Benchmarks

Existing benchmarks for evaluating Arabic LLM performance focus almost exclusively on downstream tasks such as translation, sentiment analysis, or standard multiple-choice exams. Achieving Digital Linguistic Sovereignty requires the establishment of **Cognitive-Semantic Benchmarks**. These benchmarks should specifically test an AI model's ability to preserve root metaphors, distinguish between subtle affective derivatives, and avoid the forced mapping of Arabic concepts onto unidimensional Western equivalents.

.5Conclusion and Strategic Recommendations

5.1. Conclusion

This study has challenged the orthodox paradigm in digital linguistics that evaluates the survival of the Arabic language primarily through the quantitative lens of "Digital Language Death." As Artificial Intelligence transitions from simple statistical text processing to Generative Large Language Models (LLMs), the primary existential challenge facing Arabic is no longer computational scarcity or online absence, but Automated Semantic Flattening (ASF)—a qualitative degradation of its native cognitive architecture.

Through a comparative theoretical and empirical analysis, this paper demonstrated how Transformer-based architectures, engineered around the linear, concatenative logic of Indo-European languages, structurally disrupt the non-linear, derivational root-and-pattern system (Ishtiqāq) of Arabic. Sub-word tokenization algorithms (such as BPE) break Arabic words into arbitrary substrings, severing lexical items from their multi-layered semantic roots. Consequently, complex spatio-philosophical, affective, and socio-ethical concepts are systematically compressed into unidimensional, surface-level equivalents aligned with English-centric vector spaces.

If unaddressed, this algorithmic restructuring threatens to induce long-term epistemic erosion among future generations of Arabic speakers, reducing a historically resilient, three-dimensional language into a structural shell for Western conceptual frameworks. Protecting Arabic in the era of Artificial Intelligence requires moving beyond data accumulation toward structural architectural reform and the assertion of Digital Linguistic Sovereignty.

5.2. Strategic Recommendations

To bridge the gap between computational development and semantic preservation, this study outlines four strategic recommendations for AI researchers, computational linguists, and policymakers:

Architectural Reform in NLP: AI research institutions in the Arab world must prioritize the development of native tokenization algorithms that process root consonants and morphological patterns as dual, interconnected vector layers rather than arbitrary sub-word character strings.
Curated Lexicographical Datasets: National AI initiatives should shift investment toward digitizing, annotating, and incorporating classical Arabic lexicons, philosophical treatises, and historical literature into pre-training corpora, ensuring models learn multi-dimensional semantic networks rather than flattened web-scraped prose.
Establishment of Semantic Benchmarks: A standardized international benchmark (e.g., Arabic Semantic Preservation Index) must be established to evaluate LLMs not merely on surface fluency or translation accuracy, but on their ability to retain morphological depth and root-metaphor integrity.
Interdisciplinary Collaboration: Computer scientists, NLP engineers, and cognitive linguists must establish joint research programs to build AI architectures that reflect the native ontological matrix of Arabic, setting a precedent for other non-Indo-European languages facing algorithmic flattening.

Notes

Non-concatenative Morphology: Refers to a word-formation process where words are created by modifying the internal structure of a root (e.g., inserting vowels into a consonant framework) rather than simply attaching prefixes or suffixes sequentially.
Ishtiqāq :(الإشتقاق) The foundational derivational logic of Arabic grammar, wherein hundreds of distinct verbs, nouns, and adjectives are systematically derived from a single tri-consonantal or quadri-consonantal root.
Byte-Pair Encoding (BPE): A data compression technique adapted for sub-word tokenization in modern LLMs. It iteratively replaces the most frequent pairs of bytes/characters in a text with a new single token, regardless of morphological boundaries.

Vector Space Alignment: The mathematical process in machine learning where words from different languages are mapped into a shared high-dimensional mathematical space based on statistical co-occurrence, often forcing non-Western concepts into alignment with English semantic structures.
Lisān al-ʻArab :(لسان العرب) The seminal 13th-century Arabic lexicon compiled by Ibn Manẓūr, widely regarded as the most comprehensive historical dictionary of classical Arabic root meanings and metaphorical usages.

References

First: Foreign References

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610–623. https://doi.org/10.1145/3442188.3445922
2. Coulmas, F. (2017). The writing systems of the world. Wiley-Blackwell.
3. Ferguson, C. A. (1959). Diglossia. Word, 15(2), 325–340. https://doi.org/10.1080/00437956.1959.11659702
4. Habash, N. (2010). Introduction to Arabic natural language processing. Morgan & Claypool Publishers. https://doi.org/10.2200/S00277ED1V01Y201005HLT010
5. Holes, C. (2004). Modern Arabic: Structures, functions, and varieties (2nd ed.). Georgetown University Press.
6. Holland, J. H. (1995). Hidden order: How adaptation builds complexity. Addison-Wesley.
7. Kornai, A. (2013). Digital language death. PLoS ONE, 8(10), e77056. https://doi.org/10.1371/journal.pone.0077056
8. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725. https://doi.org/10.18653/v1/P16-1162
9. Share, D. L. (2008). On the peculiarities of Arabic orthography. Reading and Writing, 21(6), 581–604. https://doi.org/10.1007/s11145-007-9093-6
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information

Processing Systems (NeurIPS 2017), 30, 5998–6008.
11. Zuhaira, A., & Magdy, W. (2019). Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian and Low-Resource Language Information Processing, 19(6), 1–24.

Second: Arabic References

ابن منظور، محمد بن مكرم. (1990). لسان العرب (ط. 1). دار صادر.
الجديدي، المبروك. (2019). الذكاء الاصطناعي ومعالجة اللغة العربية: التحديات والآفاق. مجلة اللسانيات العربية، المجلد 8، 45-68.
حسام الدين، كريم زكي. (2000). أصول التفكير الدلالي. مكتبة الأنجلو المصرية.
حماد، علي عبد القادر. (2021). اللسانيات الحاسوبية واللغة العربية: النظرية والتطبيق. دار المعارف.
الفاسي الفهري، عبد القادر. (2013). اللغة العربية والتنمية في المفهوم والتخطيط والسياسة. دار الكتاب الجديد المتحدة.