Hypothesis

Not peer-reviewed version

A Scalable Four-Level Functional Hierarchy for Evaluating Large Language Models: Hallucination, Self-Monitoring, and the Hypothesized Structural Advantages of Arabic and Chinese

El Khalil Baroudi *

Posted Date: 24 March 2026

doi: 10.20944/preprints202603.1858.v1

Keywords: four-level functional hierarchy; large language models; Arabic NLP; Chinese NLP; hallucination reduction; AI alignment; MorphBPE; AraHalluEval; recursive self-improvement; digital economy; theoretical framework; falsifiability

 doi:10.20944/preprints202603.1858.v1

Hypothesis

A Scalable Four-Level Functional Hierarchy for Evaluating Large Language Models: Hallucination, Self-Monitoring, and the Hypothesized Structural Advantages of Arabic and Chinese

El Khalil Baroudi

Department of Theoretical Physics, USTHB, Algiers, Algeria; baroudikalid@gmail.com

Nature of This Work: This paper presents a theoretical framework — the Four-Level Functional Hierarchy (FLFH) — as a hypothesis for explaining systematic failure modes in current large language models and for guiding future architectural research. The empirical results cited (AraHalluEval, MorphBPE benchmarks, ERNIE GLUE scores) serve as *supporting evidence consistent with the framework*, not as direct experimental validation of the FLFH model itself. Full empirical validation requires dedicated comparative studies using the research agenda outlined in Section 6. Reviewers should evaluate this work as a theoretical hypothesis paper offering falsifiable predictions, not as an experimental study reporting validated findings.

Abstract

The rapid advancement of Large Language Models (LLMs) has exposed persistent failure modes — hallucination, alignment brittleness, and energy inefficiency — that scale alone has not resolved. We hypothesize that these failures are not incidental bugs but systematic symptoms of a deeper architectural gap: the treatment of language processing as a flat, undifferentiated process rather than a hierarchy of functionally distinct levels. Drawing on established cognitive neuroscience and functional dissociation literature, we propose the **Four-Level Functional Hierarchy (FLFH)** — comprising Automated Encoding (L1), Symbolic Synthesis (L2), Volitional Self-Monitoring (L3), and Ethical-Teleological Integration (L4) — as a diagnostic and evaluative framework for LLM architectures. We extend a five-criterion Structural Transition Test (STT) to include Recursive Self-Improvement Capacity, addressing digital economy scalability. Under the FLFH interpretation, a systematic analysis suggests that current leading models (GPT-4, Claude, Gemini, DeepSeek) operate predominantly within L1/L2 functional regimes, lacking the L3/L4 integration that the framework posits as necessary for genuine self-monitoring and ethical commitment. We further *hypothesize* that morphologically rich languages — Arabic and Chinese — may provide structural scaffolding that reduces hallucination rates and improves computational efficiency. Preliminary evidence from AraHalluEval suggests Arabic-specific models achieve statistically fewer hallucinations ($p < 0.01$), and MorphBPE tokenization improves morphological consistency F1 from 0.00 to 0.66. These findings are interpreted as consistent with, though not conclusive proof of, the FLFH framework. Falsifiability conditions and a validation roadmap are provided to guide experimental follow-up.

Keywords: four-level functional hierarchy; large language models; Arabic NLP; Chinese NLP; hallucination reduction; AI alignment; MorphBPE; AraHalluEval; recursive self-improvement; digital economy; theoretical framework; falsifiability

 doi:10.20944/preprints202603.1858.v1

1. Introduction

The emergence of state-of-the-art Large Language Models (LLMs) has foregrounded a persistent and theoretically important question: why do systems with massive scale and impressive linguistic fluency continue to hallucinate facts, fail in specialized reasoning domains, and resist principled alignment? Hendrycks et al. [1] demonstrated that GPT-3 achieves only 43.9% accuracy on the Massive Multitask Language Understanding (MMLU) benchmark — near chance in domains such as law, ethics, and formal reasoning — despite training on hundreds of billions of tokens. Scale-up experiments have not closed this gap proportionally, suggesting that the problem is structural rather than quantitative.

This paper proposes a theoretical answer to that question. We hypothesize that current LLM architectures suffer from a systematic functional gap: the absence of dedicated, hierarchically organized processing levels for self-monitoring and goal-coherent behavior. We formalize this hypothesis as the Four-Level Functional Hierarchy (FLFH), a framework grounded in established cognitive neuroscience, and use it to derive falsifiable predictions about LLM behavior and about the potential computational advantages of morphologically rich languages.

We wish to be explicit about the epistemic status of this work. The FLFH is a theoretical framework — a structured hypothesis — not a validated empirical model. The empirical data cited throughout the paper (AraHalluEval results, MorphBPE benchmarks, ERNIE GLUE performance) are interpreted as evidence consistent with the framework's predictions, not as proof of the framework. We outline falsifiability conditions and a research agenda for experimental validation.

1.1. Research Gaps and Motivation

Existing LLM evaluation frameworks are predominantly performance-based, measuring output accuracy without interrogating the functional architecture that produces — or systematically fails to produce — reliable outputs. Three interconnected gaps motivate the present work:

- Gap 1 — Flat Processing Models: Current transformer architectures process all operations through a single attention-weighted sequence mechanism. There is no architectural distinction between automatic feature encoding, symbolic reasoning, self-monitoring, and goal-directed behavior. This flat design provides no mechanism for detecting and correcting one's own inferential failures at runtime.
- Gap 2 — Absent Self-Monitoring Architecture: Hallucination is not merely an accuracy failure; it may reflect the structural absence of a dedicated reality-monitoring level — a mechanism for evaluating outputs against an internal model of epistemic confidence. Without such a level, models cannot systematically distinguish confident confabulation from verified knowledge.
- Gap 3 — Externalized Ethical Constraints: Current alignment approaches impose behavioral constraints externally via RLHF. These constraints are pattern-matched rather than ethically integrated. A system that rejects harmful prompts through pattern activation may fail to generalize that rejection to novel harmful patterns not represented in training data.

These gaps are not straightforwardly addressable by increasing scale. The MMLU evidence [1] suggests that architectural restructuring — specifically, the introduction of hierarchically distinct functional levels — may be required for qualitative improvement.

1.2. Contributions of This Paper

1. A Four-Level Functional Hierarchy (FLFH) framework, grounded in cognitive neuroscience and functional dissociation evidence, proposed as a diagnostic model for LLM architectural analysis.
2. An extended five-criterion Structural Transition Test (STT) incorporating Recursive Self-Improvement Capacity, addressing digital economy scalability requirements.
3. A nine-criterion comparative analysis of GPT-4, Claude, Gemini, and DeepSeek under the FLFH interpretive lens, with each rating tied to published empirical evidence.

 doi:10.20944/preprints202603.1858.v1

4. A quantitative analysis of the trade-off between Arabic information density and L1-level tokenization complexity, with MorphBPE as an engineering resolution pathway.
5. Explicit falsifiability conditions and a validation research roadmap, enabling empirical follow-up by the broader research community.

1.3. Paper Organization

Section 2 presents the FLFH framework with grounding in cognitive neuroscience. Section 3 provides the extended STT and comparative LLM analysis. Section 4 examines Chinese structural contributions via ERNIE. Section 5 analyzes Arabic empirical evidence and tokenization efficiency. Section 6 presents falsifiability conditions and the validation roadmap. Section 7 proposes future architectures. Section 8 concludes with implications for the digital economy.

2. Theoretical Framework: The Four-Level Functional Hierarchy (FLFH)

2.1. Motivation and Grounding

Cognitive science and neuroscience provide substantial evidence that biological cognition is not a single undifferentiated process but a hierarchy of functionally distinct systems operating across different timescales, at different abstraction levels, and with different degrees of voluntary control [2,3]. This hierarchical organization is not merely a biological curiosity; it may reflect a functional necessity for any system that must balance rapid automatic responses with deliberate, self-monitored, goal-coherent behavior. We propose the FLFH as a framework for applying this insight to LLM architectural analysis.

We emphasize that the FLFH is an analogy-based theoretical framework, not a claim that LLMs have or should have the properties of biological cognitive systems. Rather, the framework proposes that the functional distinctions identified in cognitive neuroscience — between automatic encoding, symbolic synthesis, self-monitoring, and goal integration — may represent functionally necessary distinctions for any reliable general-purpose reasoning system, biological or artificial.

2.2. Justification of Four Levels

2.2.1. Functional Necessity Argument

We suggest that any information-processing system operating under accuracy, energy, and generalization constraints would benefit from at least four functionally distinct processing modes. A system must (i) transduce and encode input data automatically (L1); (ii) construct symbolic inferences from encoded representations (L2); (iii) evaluate those inferences against internal coherence and accuracy criteria — self-monitoring (L3); and (iv) integrate individual decisions into stable long-range goal frameworks with normative constraints (L4).

The functional dissociation argument is as follows: merging L2 and L3 (synthesis and monitoring) produces a system that confuses inference generation with inference verification — the failure mode we observe as hallucination. Merging L3 and L4 (monitoring and goal integration) produces a system that cannot distinguish situational compliance from principled ethical commitment — the failure mode we observe in alignment brittleness. These predictions are falsifiable: if systems can be demonstrated to avoid hallucination and achieve robust alignment without separate self-monitoring and goal-integration mechanisms, the FLFH would be disconfirmed.

2.2.2. Neural Dissociation Evidence

The four-level partition is consistent with well-documented anatomical and temporal gradients in human cognition [2–4]. L1 maps onto primary sensory and subcortical processing (20–150 ms timescales); L2 onto dorsolateral prefrontal executive networks (150–500 ms); L3 onto the default

 doi:10.20944/preprints202603.1858.v1

mode network, vmPFC, and anterior cingulate cortex supporting self-referential monitoring; and L4 onto frontopolar cortex (Brodmann Area 10) and long-range integration networks [4]. Lesion studies confirm functional independence: frontopolar damage selectively disrupts long-range planning while leaving moment-to-moment logical analysis intact; vmPFC damage eliminates value-based decision-making while leaving perceptual processing unimpaired [4].

We do not claim that LLMs replicate or should replicate neural mechanisms. Rather, we invoke this evidence to support the functional necessity argument: if biological systems with far greater evolutionary optimization require these distinct functional levels, it is plausible — though not certain — that artificial systems pursuing general-purpose language understanding would benefit from analogous functional distinctions.

2.2.3. Parsimony

The principle of parsimony favors the smallest model that captures the observed functional dissociations. The cognitive neuroscience literature has not identified a fifth functionally independent level that cannot be decomposed into interactions among L1–L4. Intermediate granularities represent domain-specific elaborations within the four-level structure rather than requiring additional levels. We acknowledge this as a provisional claim subject to revision if future neuroscientific or computational evidence identifies a fifth dissociable functional level.

2.3. The Four Functional Levels

Level 1 (L1) — Automated Encoding: Handles input transduction, automatic feature extraction, and procedural encoding without volitional control. In LLM terms: tokenization, embedding, and low-level pattern extraction. L1 produces representations but not inferences. Analogous to primary sensory cortices operating on 20–150 ms timescales.

Level 2 (L2) — Symbolic Synthesis: Transforms L1 representations into structured inferences, relational reasoning, and sequential compositions. In LLM terms: multi-head attention, transformer layers, and compositional inference. L2 produces rule-governed outputs but does not monitor their reliability. Analogous to dlPFC executive networks operating on 150–500 ms timescales.

Level 3 (L3) — Volitional Self-Monitoring: Evaluates L2 outputs against internal coherence and accuracy criteria; detects contradictions; constructs an epistemic self-model (what the system does and does not know). In LLM terms: this level is *structurally absent* in current architectures. The hallucination phenomenon is hypothesized to follow directly from this absence. Analogous to DMN/vmPFC/ACC self-referential monitoring networks.

Level 4 (L4) — Ethical-Teleological Integration: Maintains normative coherence across time and context; integrates individual decisions into stable goal hierarchies with principled ethical constraints. In LLM terms: this level is also *structurally absent*; current RLHF-based alignment implements pattern-matched L2 behavioral constraints, not principled L4 ethical integration. Analogous to frontopolar cortex long-range integration networks.

2.4. Extended Structural Transition Test (STT)

We extend the Structural Transition Test to five criteria distinguishing systems operating at different FLFH levels. Each criterion is defined as a hypothesis about what architectural properties

 doi:10.20944/preprints202603.1858.v1

are required for the corresponding capability — not as an assertion that these properties are necessary by logical necessity.

- Criterion 1 — Intrinsic Goal Generation: Under FLFH, genuine goal-directed behavior requires L4-level teleological integration. Systems whose goals are entirely externally imposed (via RLHF) are hypothesized to exhibit brittleness in novel goal-conflict situations.
- Criterion 2 — Self-Referential Semantics: Reliable factual grounding may require L3-level epistemic self-modeling. Without this, symbol use is predicted to remain statistically driven without internal verification.
- Criterion 3 — Unified Processing State: Coherent multi-step reasoning may require cross-level integration. Systems processing each query as an independent forward pass are predicted to exhibit coherence failures in extended reasoning chains.
- Criterion 4 — Internal Temporal Continuity: Goal-coherent planning may require L4-level temporal integration. Systems without this are predicted to treat temporal context as a contextual variable rather than a structured planning framework.
- Criterion 5 — Recursive Self-Improvement Capacity: Genuine self-directed architectural improvement may require L3-level self-monitoring (to detect performance degradation) and L4-level goal coherence (to ensure corrections align with overarching objectives). Current systems are hypothesized to be limited to retrieval-augmented output correction, not architectural self-diagnosis.

3. Comparative Analysis of Contemporary LLMs Under the FLFH Framework

3.1. Methodology and Limitations

We evaluate four leading models — GPT-4 [6], Claude [7], Gemini, and DeepSeek [8] — against nine criteria derived from the FLFH framework and the extended STT. Ratings are based on published empirical performance data and technical reports rather than direct experimental measurement. This analysis is therefore interpretive rather than empirical: it maps documented model behaviors onto FLFH categories to derive hypotheses about architectural gaps.

Important caveat: The claim that models 'operate at L1/L2' is an interpretive statement under the FLFH framework, not a direct empirical measurement of internal architecture. More precisely, under the FLFH interpretation, the documented behavioral evidence is consistent with models operating predominantly within L1/L2 functional regimes. Alternative interpretations of the same behavioral evidence are possible and would be considered by a full experimental validation study.

3.2. Results

Table 1. FLFH-based interpretive evaluation of leading LLMs. Ratings reflect behavioral consistency with FLFH predictions, not direct architectural measurement. (✓ Consistent with achievement; ~ Partial/ambiguous; ✗ Consistent with absence).

Criterion	GPT-4 [6]	Claude [7]	Gemini	DeepSeek [8]	Evidence Base
L1 Automated Encoding	✓	✓	✓	✓	Robust encoding across all models [1]
L2 Symbolic Synthesis	~	~	~	~	Partial logical synthesis; specialized reasoning gaps [1]

 doi:10.20944/preprints202603.1858.v1

L3 Self-Monitoring	✗	~ (pseudo)	✗	✗	Claude's safety rules are external pattern-matching, not internal monitoring [7]
L4 Ethical Integration	✗	✗	✗	✗	RLHF constraints do not constitute L4 under FLFH interpretation
Intrinsic Goal Generation	✗	✗	✗	✗	All goals externally imposed via RLHF [6–8]
Self-Referential Semantics	✗	✗	✗	✗	Symbol grounding unresolved in all current architectures
Structural Transparency	✗	✗	✗	~	DeepSeek-R1 exposes chain-of-thought; others opaque [8]
Energy Efficiency	✗	✗	✗	~	DeepSeek MoE architecture provides partial efficiency gain [8]
Recursive Self-Improvement	✗	✗	✗	✗	No model initiates self-directed architectural correction

3.3. Interpretive Findings

Under the FLFH framework, the behavioral patterns of all evaluated models are consistent with predominantly L1/L2 processing, with complete absence of functional signatures predicted by L3 and L4 architectural levels. These patterns are, however, consistent with multiple interpretations, and the FLFH explanation represents one theoretically motivated hypothesis among several.

The Pseudo-L3 Observation. Claude's safety-filtering behavior could be interpreted as a rudimentary form of self-monitoring, and we note this ambiguity. Under the FLFH interpretation, however, the critical distinction is that genuine L3-level monitoring would involve the system evaluating its own outputs against an internal epistemic model — not matching inputs against a trained rejection classifier. A system capable of true L3 monitoring would be predicted to generalize harm detection to novel harmful patterns not present in training data, an empirical prediction that could be tested with appropriate out-of-distribution evaluation sets.

Hallucination as a Hypothesized L3 Deficit. AraHalluEval [9] documents high hallucination rates across all evaluated models. Under the FLFH framework, we hypothesize that this is a structural consequence of absent L3-level reality-monitoring — the system has no mechanism to distinguish confident inference from verified knowledge. We acknowledge that alternative explanations exist, including insufficient training data, suboptimal training objectives, or retrieval failures. The FLFH explanation is falsifiable: if an LLM trained without explicit self-monitoring architecture demonstrates hallucination rates comparable to human-level accuracy on factual tasks, the L3 deficit hypothesis would be weakened.

 doi:10.20944/preprints202603.1858.v1

4. The Chinese Structural Hypothesis: Evidence from Baidu's ERNIE

4.1. Structural Properties of Chinese

Chinese presents structural properties that distinguish it from alphabetic languages in ways that may be architecturally relevant. As a logographic system, Chinese characters encode meaning through systematic combination rather than phonological representation [11]. In English, isolated words retain meaning; in Chinese, meaning emerges from character strings as unified semantic units. We hypothesize that this structural property necessitates phrase-level rather than token-level semantic representation, with measurable consequences for model architecture.

4.2. ERNIE: Evidence Consistent with the FLFH Hypothesis

Baidu's ERNIE (Enhanced Representation through Knowledge Integration) replaced BERT's random token-masking with phrase-level entity masking — masking and predicting entire semantically coherent strings as unified units [11,12]. This innovation was motivated directly by Chinese structural properties. ERNIE enabled Baidu to surpass 90.0 on the GLUE benchmark, outperforming both Microsoft and Google, and exceeding the average human score of approximately 87 [12].

Under the FLFH interpretation, this result is consistent with the hypothesis that phrase-level hierarchical representation (a richer L2-level structure) produces more stable and less ambiguous semantic representations than token-level processing. We note, however, that ERNIE's success may also be attributable to its knowledge integration training objective, larger pretraining corpus, or other architectural differences not directly related to Chinese structural properties. Controlled ablation studies would be required to isolate the contribution of linguistic structure per se.

4.3. FLFH Interpretation

The ERNIE result is interpreted as providing preliminary evidence for three FLFH-adjacent principles: (i) linguistic structure shapes architectural requirements; (ii) hierarchical phrase-level representation may achieve higher L2-level stability than token-level processing; and (iii) meaning may be more effectively modeled as residing at multiple hierarchical levels rather than in atomic units. These principles motivate the architecture proposed in Section 7 but do not constitute proof of the FLFH model.

5. The Arabic Structural Hypothesis: Empirical Evidence and Analysis

5.1. Structural Properties of Arabic

Arabic is a morphologically rich Semitic language whose words derive predominantly from triliteral or quadriliteral roots carrying core semantic content [13]. The root k-t-b generates systematically: kitab (book), katib (writer), maktab (office), maktaba (library). This root-pattern morphological system creates transparent semantic networks with inherent structural verification pathways. We hypothesize that these pathways may reduce the L2-level inferential burden associated with semantic disambiguation, with measurable consequences for hallucination rates.

5.2. Information Density and Tokenization Trade-Offs

5.2.1. The Information Density Advantage

Arabic encodes person, number, gender, tense, mood, voice, case, and definiteness within a single word token. The verb form sa-ya-ktubu-na (they will write) expresses in one token what requires four words in English. Corpus-level comparisons using standard BPE tokenizers suggest that Arabic text requires approximately 25–35% fewer tokens than English to convey equivalent propositional content [14]. If this estimate holds across diverse domains, it implies a reduction in

 doi:10.20944/preprints202603.1858.v1

required context window length, decreasing attention computation — which scales as $O(n^2)$ in standard transformers — by an estimated factor of 0.52 to 0.63 for equivalent semantic content.

5.2.2. The L1-Level Complexity Trade-Off

The same morphological richness introduces a measurable engineering challenge. Standard Byte Pair Encoding (BPE) tokenizers, calibrated on alphabetic languages, fragment morphologically coherent Arabic words at arbitrary boundaries, destroying root-pattern structure. Asgari et al. [14] quantify this: standard BPE achieves a Morphological Consistency F1-score of 0.00 for Arabic — morphologically related words share no token substrings. This fragmentation may force models to perform implicit morphological reconstruction at L2, increasing effective training burden.

5.2.3. MorphBPE as an Engineering Resolution

MorphBPE integrates morphological boundary information into the tokenization algorithm. After MorphBPE, Morphological Consistency F1 rises from 0.00 to 0.66 [14]. This improvement suggests measurable consequences: reduced cross-entropy loss during training (indicating faster convergence), improved downstream task performance, and alignment between L1-level token representations and L2-level semantic structures. Under the FLFH framework, this constitutes an engineering-level structural alignment between L1 and L2 — a design principle that is testable and replicable.

5.2.4. Net Efficiency Estimate

Taken together, these findings suggest — but do not conclusively prove — that Arabic with MorphBPE tokenization may offer a net computational advantage over English-centric approaches: fewer tokens reduce attention computation; MorphBPE eliminates implicit morphological reconstruction burden; and explicit semantic structure may accelerate convergence. Controlled experiments comparing Arabic-native and English-native architectures on matched tasks, with careful control for training data volume and domain, would be required to validate this estimate.

5.3. Hallucination Evidence from AraHalluEval

Alansari and Luqman [9] developed AraHalluEval, the first comprehensive hallucination evaluation framework for Arabic LLMs, assessing 12 models across factual consistency using 12 fine-grained indicators. The Arabic pre-trained ALLaM model achieves an average hallucination score of 0.382 on the Arabic GQA task, compared to Bloom (0.730), LLaMA3 (0.515), and Qwen (0.601). A Mann-Whitney U test confirms statistical significance ($p < 0.01$) for Arabic-specific model superiority.

We interpret this result as preliminary evidence consistent with the FLFH-based hypothesis that linguistic structural richness — specifically, root-pattern morphological transparency — may reduce hallucination rates at the L2/L3 boundary. We emphasize, however, that this result may also reflect differences in training data quality, domain coverage, or fine-tuning procedures between ALLaM and multilingual models. Isolating the contribution of linguistic structure requires controlled experiments holding these confounding variables constant.

5.4. Structural Transparency Through Attention Visualization

Hasan et al. [10] developed an attention visualization framework for Arabic financial RAG systems, achieving 95.04% classification accuracy and 87.63% hallucination detection by generating cross-attention heatmaps and token contribution scores. Arabic's morphological richness enables more precise attention localization: each token carries rich morphological information, potentially allowing attention heads to align more precisely with semantic units than in analytic languages.

Under the FLFH framework, this visualization capability represents a partial engineering proxy for L3-level self-monitoring — providing interpretable signals about inferential confidence. It does

 doi:10.20944/preprints202603.1858.v1

not constitute genuine L3 processing, but may represent a feasible near-term engineering step toward L3-like transparency.

6. Falsifiability Conditions and Validation Research Agenda

6.1. Why Falsifiability Matters for This Framework

A theoretical framework that cannot be falsified is not a scientific hypothesis but a classification scheme. The FLFH is intended as a scientific hypothesis — one that makes specific, testable predictions about when and why LLMs fail, and what architectural interventions would improve reliability. We therefore specify explicit conditions under which the FLFH would be disconfirmed or substantially weakened.

6.2. Falsifiability Conditions

Table 2. Falsifiability conditions for the FLFH framework and its language-structure hypotheses.

Hypothesis	FLFH Prediction	Disconfirming Result	Severity
L3 Deficit Causes Hallucination	Models without dedicated self-monitoring architecture will hallucinate at significantly higher rates than those with it	A flat-architecture model achieves human-level hallucination rates on AraHalluEval-equivalent benchmarks	High — would require revision of core L3 claim
L1/L2 Ceiling on Alignment	RLHF-based models will fail to generalize ethical constraints to novel harmful patterns not in training data	RLHF model demonstrates robust generalization to unseen harm categories equivalent to principled ethical reasoning	High — would weaken L4 claim
Arabic Structure Reduces Hallucination	Controlling for training data volume and quality, Arabic-native models will have significantly lower hallucination rates than equivalent multilingual models	No significant difference after controlling for training data confounds	Medium — would limit language-structure claims but not core FLFH
MorphBPE Efficiency Advantage	Arabic + MorphBPE will achieve equivalent downstream performance with fewer training steps than standard BPE	No convergence advantage observed in controlled training comparison	Low — would affect efficiency claims only

 doi:10.20944/preprints202603.1858.v1

Hierarchy Necessity	A model designed with explicit L1-L4 functional separation will outperform a flat model of equivalent parameter count on held-out self-monitoring tasks	Flat model matches hierarchical model on self-monitoring benchmarks	Critical — would disconfirm entire hierarchy necessity argument

6.3. Validation Research Agenda

We propose the following research agenda for empirical validation of the FLFH framework, ordered by priority:

- Priority 1 — Self-Monitoring Benchmark Construction: Develop a dedicated benchmark for LLM epistemic self-monitoring — measuring whether models can identify their own unreliable outputs without external feedback. This benchmark is the most direct test of the L3 deficit hypothesis.
- Priority 2 — Controlled Arabic vs. English Architecture Comparison: Train architecturally matched models on Arabic (with MorphBPE) and English corpora of equivalent size and domain, and compare hallucination rates on parallel test sets. This isolates the linguistic structure contribution from training data confounds.
- Priority 3 — Hierarchy Ablation Study: Implement a minimal FLFH-inspired architecture with explicit L1-L4 functional separation (using MoE or modular transformer variants) and compare against a flat baseline of equivalent parameter count on MMLU, AraHalluEval, and out-of-distribution ethical reasoning tasks.
- Priority 4 — L4 Generalization Test: Evaluate RLHF-aligned models on novel harm categories not represented in training data, assessing whether alignment generalizes (consistent with L4) or fails (consistent with L2 pattern-matching).
- Priority 5 — Longitudinal Recursive Self-Improvement Observation: Monitor FLFH-aligned systems for evidence of self-directed improvement over time, distinguishing retrieval-augmented correction from genuine architectural self-modification.

7. Toward FLFH-Aligned AI: Proposed Architecture and Language Synergies

7.1. The Convergence Principle

The evidence assembled in Sections 4 and 5 supports a preliminary convergence principle: the structural properties of a foundational training language may shape a model's computational architecture in ways that influence reliability and efficiency. This principle, if validated, has significant implications for digital economy AI infrastructure, where reliability and transparency are not optional features but mission-critical requirements for financial services, legal technology, and blockchain governance.

 doi:10.20944/preprints202603.1858.v1

7.2. Complementary Structural Resources of Arabic and Chinese

Table 3. Hypothesized complementary structural resources of Chinese and Arabic for FLFH-aligned AI development.

Dimension	Chinese (Logographic)	Arabic (Root-Pattern Morphological)
L1 Encoding Strategy	Character composition; entity-level masking	Root-pattern tokenization; MorphBPE
L2 Structural Richness	Phrase-level semantic coherence; ERNIE-style masking	Case marking; morphological precision; syntactic explicitness
Hypothesized Hallucination Resistance	Phrase coherence constraints	Root-pattern semantic verification pathways
Information Density	High (combinatorial character meaning)	Very high (multi-feature single-token encoding)
Transparency Enablement	Character radical interpretability	Morpheme-level attention alignment
Digital Economy Applications	East Asian financial and legal AI	GCC and MENA blockchain and fintech systems

7.3. Proposed Architecture Principles

Drawing on the FLFH framework and the empirical evidence assessed above, we propose the following four architectural principles for FLFH-aligned next-generation AI. These are offered as engineering hypotheses subject to empirical evaluation, not as validated design prescriptions.

L1 — Morphological Encoding Layer:

Specialized tokenization networks handling root-pattern systems (Arabic via MorphBPE [14]) and character composition (Chinese via entity-level masking [11]) with energy-optimized parallel processing. L1 representations should expose morphological structure explicitly to higher levels, rather than burying it in black-box embeddings.

L2 — Hierarchical Semantic Representation Layer:

Transformer attention mechanisms representing meaning at multiple hierarchical levels — roots, patterns, words, phrases — with transparent interpretation interfaces. Cross-attention heatmaps should be treated as first-class outputs for downstream verification, not internal artifacts.

L3 (proto) — Self-Monitoring and Consistency Verification Layer:

Dedicated verification modules — distinct from the primary generation pathway — assessing grammatical coherence, factual consistency, and epistemic confidence before committing to an output. These modules should log failure patterns for later analysis, constituting a proto-recursive self-improvement pathway. Implementation via Mixture-of-Experts (MoE) architectures with designated verification experts is technically feasible with current infrastructure [8].

L4 (proto) — Normatively-Grounded Value Integration Layer:

Programmable normative knowledge bases [15] — updated as ethical standards and regulatory requirements evolve — with mandatory human oversight. Crucially, this layer should support modular rule injection enabling regulatory updates without full retraining, a scalability property

 doi:10.20944/preprints202603.1858.v1

essential for digital economy deployment. Domain-specific normative calibration (medical, legal, financial) should be achievable through module swapping rather than retraining.

8. Conclusion

8.1. Summary of Contributions

This paper has presented a theoretical framework — the Four-Level Functional Hierarchy (FLFH) — as a structured hypothesis for understanding systematic failure modes in current LLMs and guiding future architectural development. Five primary contributions are offered:

1. A Four-Level Functional Hierarchy grounded in cognitive neuroscience and functional necessity arguments, proposing that the absence of dedicated self-monitoring (L3) and ethical integration (L4) levels structurally explains hallucination and alignment brittleness in current LLMs.
2. An extended five-criterion Structural Transition Test incorporating Recursive Self-Improvement Capacity, with explicit mapping to digital economy scalability requirements.
3. A nine-criterion comparative analysis of leading LLMs under the FLFH interpretive lens, with each rating tied to published empirical evidence and explicitly framed as interpretation rather than direct architectural measurement.
4. A quantitative analysis of Arabic information density and MorphBPE tokenization, arguing that morphological richness can be converted from a tokenization liability to a net computational advantage.
5. Explicit falsifiability conditions and a five-priority validation research agenda, enabling the broader community to test, extend, or disconfirm the framework.

8.2. Epistemic Status and Limitations

We reiterate that the FLFH framework is a theoretical hypothesis, not a validated empirical model. The comparative LLM analysis is interpretive, not experimental. The Arabic and Chinese efficiency claims are supported by preliminary evidence but require controlled experimental validation. The proposed architecture is a set of engineering hypotheses, not a proven design. We have endeavored throughout to use language that reflects this epistemic status.

The framework's primary contribution is not to provide final answers but to reframe the right questions: not 'how do we scale LLMs further?' but 'what functional architectural levels are necessary for reliable self-monitoring and principled ethical reasoning, and how do we engineer them?'

8.3. Implications for AI Development and the Digital Economy

If the FLFH hypotheses are validated — even partially — the implications for digital economy AI infrastructure would be significant. Hallucinations in AI systems undermine trust in automated financial auditing, smart contract execution, and regulatory compliance. Alignment brittleness undermines reliable deployment in legal and medical contexts. The FLFH framework suggests that these are not problems to be solved by adding more training data or increasing model size, but by restructuring the functional architecture of AI systems to include dedicated self-monitoring and goal-integration mechanisms.

For practitioners, this framework offers a strategic reorientation: investing in linguistically rich, hierarchically structured AI architectures may yield more reliable and transparent systems for mission-critical applications — not because the theoretical framework is proven, but because the empirical evidence assembled here provides sufficiently motivated grounds for experimental investment.

8.4. Closing Remarks

The persistent failure modes of current LLMs — hallucination, alignment brittleness, and opaque reasoning — reveal systematic architectural gaps that scale alone has not resolved. The FLFH

 doi:10.20944/preprints202603.1858.v1

framework proposes a specific, falsifiable account of those gaps and a research pathway for addressing them. Arabic and Chinese, with their structurally rich morphological and logographic systems, offer preliminary empirical grounding for the hypothesis that linguistic structure shapes computational architecture in ways that matter for reliability.

MorphBPE improves Arabic morphological consistency from F1 = 0.00 to F1 = 0.66 [14]; AraHalluEval demonstrates statistically significant ($p < 0.01$) hallucination reduction in Arabic-native models [9]; attention visualization achieves 95.04% classification accuracy in Arabic financial systems [10]. These are measured engineering results. They do not prove the FLFH framework, but they provide a coherent empirical foundation on which to build the experiments that would.

References

1. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring Massive Multitask Language Understanding. In Proceedings of the International Conference on Learning Representations (ICLR), 2021. arXiv:2009.03300.
2. Fuster, J.M. The Prefrontal Cortex, 5th ed.; Academic Press: London, UK, 2015; ISBN 978-0-12-407815-4.
3. Miller, E.K.; Cohen, J.D. An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. 2001, 24, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167
4. Ramnani, N.; Owen, A.M. Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. Nat. Rev. Neurosci. 2004, 5, 184–194. https://doi.org/10.1038/nrn1343
5. Evans, V.; Green, M. Cognitive Linguistics: An Introduction; Edinburgh University Press: Edinburgh, UK, 2006; ISBN 978-0-7486-1832-5.
6. Hurst, A.; Lerer, A.; Goucher, A.P.; et al. GPT-4o System Card. OpenAI Technical Report; OpenAI: San Francisco, CA, USA, 2024.
7. Anthropic. Claude's Model Card and Safety Assessment; Anthropic Technical Report; Anthropic: San Francisco, CA, USA, 2024.
8. Guo, D.; Yang, D.; Zhang, H.; et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv 2025, arXiv:2501.12948.
9. Alansari, M.; Luqman, H. AraHalluEval: A Comprehensive Hallucination Evaluation Framework for Arabic Large Language Models. In Proceedings of the ArabicNLP Workshop, COLING 2025. https://doi.org/10.18653/v1/2025.arabicnlp-main.12
10. Hasan, A.; Al-Qurishi, M.; Al-Rakhami, M. Attention Mapping for Hallucination Reduction in Arabic Financial RAG Systems. J. Intell. Inf. Nat. Comput. (JIINC) 2025.
11. Sun, Y.; Wang, S.; Li, Y.; et al. ERNIE: Enhanced Representation through Knowledge Integration. arXiv 2019, arXiv:1904.09223.
12. Hao, K. Baidu has a new trick for teaching AI the meaning of language. MIT Technology Review, 26 December 2019. https://www.technologyreview.com/2019/12/26/131372/
13. Habash, N.Y. Introduction to Arabic Natural Language Processing; Synthesis Lectures on Human Language Technologies; Morgan & Claypool Publishers: San Rafael, CA, USA, 2010. https://doi.org/10.2200/S00277ED1V01Y201008HLT010
14. Asgari, E.; Rinott, R.; Goldberg, Y.; et al. MorphBPE: Morphology-Aware Byte Pair Encoding for Arabic Large Language Models. arXiv 2025, arXiv:2502.00894.
15. Alwajih, F.; Naous, T.; Elmadany, A.; et al. Pearl: A Multimodal Culturally-Aware Arabic Instruction Dataset. In Findings of the Association for Computational Linguistics: EMNLP 2025; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025.
16. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, 2019. arXiv:1810.04805.
17. Bari, M.S.; Alnumay, A.; Alzahrani, N.; et al. ALLaM: Large Arabic Language Model. arXiv 2024, arXiv:2407.15390.
18. Grattafiori, A.; Dubey, A.; Jauhri, A.; et al. The LLaMA 3 Herd of Models. arXiv 2024, arXiv:2407.21783.

 doi:10.20944/preprints202603.1858.v1

19. Farghaly, A.; Shaalan, K. Arabic Natural Language Processing: Challenges and Solutions. ACM Trans. Asian Lang. Inf. Process. 2009, 8, 4. https://doi.org/10.1145/1644879.1644881
20. Rhel, H.; Roussinov, D. Large Language Models and Arabic Content: A Review. arXiv 2025, arXiv:2505.08004.