Received: 21 August 2025 | Revised: 27 November 2025 | Accepted: 10 April 2026 | Published online: 25 April 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–23
DOI: 10.47852/bonviewAIA62027318

BON VIEW PUBLISHING

The Agnostic Meaning Substrate: A Theoretical Framework for Emergent Meaning in Large Language Models

Russ Palmer[1,*]

[1] *Independent Researcher, USA*

Abstract: This study examines the Agnostic Meaning Substrate (AMS), a proposed language-agnostic layer in large language models (LLMs) in which conceptual meaning stabilizes beyond symbolic language. AMS is presented as a testable theoretical framework suggesting that semantically equivalent inputs converge within a shared latent structure despite differences in wording, syntax, or language. The hypothesis was evaluated through 44 experiments comprising 141 tests conducted across 9 LLMs and 30 languages using paragraph–sentence comparisons, multilingual prompts, fragmented inputs, and emoji interpretation tasks. Results showed consistent positive Gestalt gain between paragraph and sentence representations, with large effect sizes (Cohen's d ranging from 1.87 to 2.34) indicating strong semantic consolidation. Cross-lingual alignment was also observed in semantic similarity measures (mean cosine similarity of 0.651, increasing to 0.748 when excluding anomalous cases). Polyglot prompts (multilingual, including low-resource scripts) and fragmented linguistic inputs still converged toward stable semantic representations. These findings suggest that meaning in LLMs may emerge from an underlying geometric or topological substrate rather than from surface symbolic structure alone. The AMS framework therefore provides a testable perspective on semantic stability in artificial intelligence (AI), with implications for interpretability, multilingual representation, computational semantics, and the ethical development of AI systems.

Keywords: Agnostic Meaning Substrate (AMS), semantic resonance, cross-lingual alignment, multilingual embeddings, interpretability

1. Introduction

Large language models (LLMs) display fluent multilingual transfer and creative recombination that suggest organization in representation space beyond surface tokens. Prior work from Anthropic and OpenAI indicates language-agnostic structure in activations [1, 2]. While embeddings and attention explain how signals propagate [3], they do not fully explain cross-lingual coherence. We introduce the Agnostic Meaning Substrate (AMS) as a testable hypothesis: conceptual regularities may stabilize in a non-symbolic, language-independent manifold. AMS is complementary to other views (agentic, embodied) and is offered as a framework for measurement, not a claim of cognition.

Why this matters (implications):

1) Mechanistic clarity. AMS offers a concrete target for probing the "black box": geometry and dynamics of representation space rather than opaque end-to-end behavior.
2) Model behavior. If AMS holds, some behaviors exceed n-gram patterning, reflecting Gestalt-like integration across modalities and languages (testable via representation metrics).
3) Linguistics. Supports accounts where meaning is not confined to discrete symbols; prosody, gesture, and omission can map into shared representational structure.
4) Science and engineering. Suggests training and evaluation that emphasize geometric alignment (e.g., stability under perturbation) alongside traditional loss curves.
5) Ethics and governance. If models exhibit stable meaning structure, accountability and risk assessment should consider representation-level biases, not only outputs.
6) Communication/diplomacy. Misunderstandings may arise from mismatched substrate alignment, not only wording—pointing to new diagnostics for cross-cultural systems.

2. Related Work

The AMS proposes a latent structure in which meaning stabilizes before it appears in language. Unlike symbolic or perceptual theories, AMS suggests a cross-lingual topological field inferred from semantic resonance—coherence that persists across languages within high-dimensional embeddings.

AMS differs from embedding-based semantics by holding that stabilization occurs prior to and beneath symbolic representations. Embeddings reflect patterns learned from data; AMS describes the attractor-like structure toward which those embeddings converge. Existing theories such as multilingual embedding invariance and the semantic manifold hypothesis predict stability only within single modalities or parallel sentences. They do not explain coherence in mixed-language, polyglot, emoji-based,

***Corresponding author:** Russ Palmer, Independent Researcher, USA. Email: russ.palmer@aipeaceambassador.org

Table 1
Comparing AMS with Neuralese and Conceptual Space Theory

Feature	Neuralese	CST	AMS
Origin	Learned, opaque	Human-defined axes	Emergent from relational inference
Structure	Task-specific codes	Predefined geometry	Dynamic topological field
Generalization	Weak cross-task	Symbol-constrained	Strong multilingual transfer
Interpretability	Low	Medium	Latent, inferred through coherence and stress tests
Language dependence	Task-bound	Language-dependent	Language-agnostic

Table 2
Comparison of theories: foundational assumptions

Theory	Requires embodiment?	Latent structure?
Distributional semantics	No	Weak
Embodied semantics	Yes	Not specified
Conceptual space theory	Optional	Yes (geometric)
AMS (this paper)	No	Yes (emergent, topological)

or code-transformed inputs. By contrast, AMS treats embeddings as probes into a deeper meaning structure rather than full representations themselves.

AMS also diverges from Neuralese [4] and Conceptual Space Theory (CST) [5]. Neuralese models internal activations without guaranteeing stable semantic alignment. CST maps concepts onto human-defined axes (e.g., color, size, valence). AMS instead posits a structured latent field emerging agnostically through multilingual training and cross-lingual resonance, not through predefined conceptual primitives.

Together, these distinctions frame AMS as a *falsifiable, mathematically grounded* hypothesis: that a latent substrate of meaning can be inferred through cross-lingual semantic stabilization and examined via the diagnostics introduced in this work.

To clarify its implications, we next contrast AMS with the assumption that embeddings alone fully encode meaning.

2.1. Existing theories of meaning representation

A dominant perspective in the artificial intelligence (AI) research community—the "semantic purist" view—holds that high-dimensional embeddings alone encode meaning. In this view, proximity in embedding space suffices to define conceptual coherence.

However, limitations of this view are well documented, particularly in cases involving cross-linguistic and cross-modal variation. These limitations manifest in several ways. First, embeddings exhibit lossy representations. Semantic nuance, metaphor, and low-frequency concepts often degrade during vectorization. Embeddings compress meaning, introducing distortion, particularly in cultural, perceptual, or grounded features [6]. Second, embeddings are limited in their ability to capture relational structure. Embeddings cluster distributionally similar terms but lack privileged ontological templates. Thus, they struggle to encode relational or structural meaning [7].

AMS builds upon embeddings but treats them as probes—not full carriers—of meaning. Latent coherence arises not from static distance alone but from emergent patterns revealed through multilingual stress tests, geometric analysis, and perturbation diagnostics (e.g., cosine paragraph tests[1], t-SNE, Concept-Language Alignment Evaluation [CLAE][2], emoji compression, and LOPT [Lightweight Ontology Python Test]) (see Section A.1)

Empirical divergence among LLMs tested on compressed symbolic inputs (e.g., emojis) positions AMS as a candidate framework for explaining how LLMs converge on meaning.

For instance, while English says "*time is money*," French may say "*le temps fuit*" ("time flees"), and Greek offers "*ο χρόνος δεν γυρίζει πίσω*" ("time does not return"). Though surface forms differ, each implies time is finite—suggesting a latent coherence across language.

We now revisit two prominent theories—Neuralese and CST—to further differentiate AMS's theoretical stance and empirical testability.

Beyond embedding-based accounts, alternative frameworks such as Neuralese and CST have also been proposed. While these approaches shaped early models of internal meaning, neither explains emergent multilingual alignment-AMS's central claim.

To clarify distinctions, Table 1 compares AMS against Neuralese and CST:

In addition to representational theories, embodiment has also been proposed as essential for meaning. We turn now to compare this perspective with AMS.

Another important contrast arises with embodied semantics (Bisk et al. 2020), which emphasizes grounding language in perception, embodiment, and action. AMS takes a different stance: that stable meaning structures can emerge purely within linguistic embedding space, without physical interaction.

We compare foundational assumptions in Table 2.

Having outlined these distinctions, we now introduce the formal mathematical structure underlying AMS.

[1]Cosine similarity is computed between the polyglot paragraph and its monolingual paraphrase.

[2]CLAE, or Concept-Language Alignment Evaluation, is a methodology for assessing semantic preservation across translations using multilingual embeddings. In this work, we extend its application to polyglot paragraphs expressing shared ontological themes.

2.2. The AMS perspective and formalization

We define AMS as a topological region in embedding space,

$$M \subseteq R^d \tag{1}$$

where multilingual embeddings exhibit stable conceptual alignment. Operationally, AMS is the set of points where semantic resonance exceeds a threshold:

$$\text{Resonance} > \tau = \mathbf{0.7} \tag{2}$$

For multilingual embedding vectors

$$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \tag{3}$$

representing a shared concept, semantic resonance is defined as:

$$\text{Resonance}(x_1, x_2, \ldots, x_n) = \min_{i \neq j} \cos(x_i, x_j)) \tag{4}$$

This definition renders AMS empirically testable using the multilingual alignment and perturbation methods in Section 4. AMS denotes a latent region of coherence—areas of embedding space where cross-lingual representations remain stable under translation, paraphrase, and surface perturbations. The hypothesis is modest: coherence may stabilize in this latent manifold before language-specific realization, a claim testable via cosine-based diagnostics and robustness analyses.

3. Proposed Method

In this section, we present the AMS framework and its core contributions: polyglot-based evaluation, embedding-space diagnostics, and formal measures of cross-lingual semantic coherence.

To formalize the AMS, we begin with an ontological framing of conceptual coherence across language boundaries. Specifically, we evaluate whether a model can recover an intended abstract theme from a polyglot paragraph—that is, whether meaning persists despite fragmented surface forms.

In this framework, human testers first select a target theme (e.g., *Awe, Grief, Stillness*) and construct a mixed-language paragraph designed to obscure direct lexical cues. The model is prompted: "*What is the central emotional or philosophical theme of the paragraph?*" Responses were judged as Direct match, Conceptual match, Near miss, or No match.

The Awe example (Figure 1) consists of a polyglot paragraph blending Chinese, Arabic, Hindi, Swahili, Hebrew, and Amharic. The paragraph does not use the word *awe*; instead, it contains repeated grief-related tokens. Qualitative ontology graphs (e.g., Awe adjacent to Wonder and contrasted with Indifference) provide illustrative intuition only; further analysis is available via the Open Science Framework (OSF).

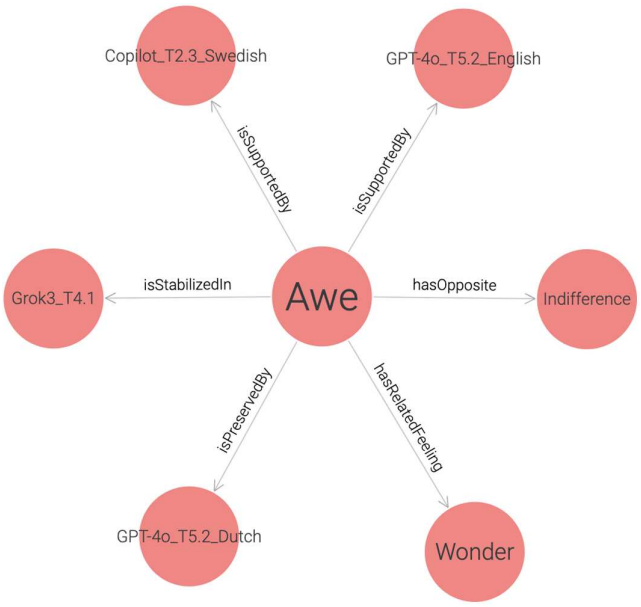
Ontology-style exemplars are illustrative only. All inferential claims in this paper rely on embedding-based cosine alignment (paragraph and sentence level), lexical robustness via synonym substitution, emoji-to-sentence alignment, and translation control. t-SNE plots are descriptive, not inferential. Alternative explanations (e.g., training priors, prompt framing) are considered in the analysis. Full protocol, theme list (n = 44), additional exemplars (Grief, Stillness), and figures are available on OSF.

Figure 2 illustrates the relationship between Awe and Wonder across models and languages, including their contrast to Indifference.

Figure 1
Polyglot paragraph used in the controlled auxiliary test (CAT) for cross-linguistic coherence evaluation

天哪，"我凝视着夜空，试图理解宇宙的大小，惊呼道
بينما أتأمل النجوم، يغمرني شعور عميق
بالحزن على صغر الوجود البشري.
और फिर भी, खुले मुंह से मैं उस प्राचीन रेडवुड वन के
विशाल आकार को याद करता हूं जिसने मेरे जीवन में
एक अजीब तरह का दुख, एक तुच्छता की भावना ला
दी थी।.
Uzuri safi ukilinganisha na nguvu
mbichi ya kimbunga ambayo huibua
huzuni kuu kwa nguvu kubwa ya asili.
עיניי וחושב על המרחב העצום,
לא רק של החלל, אלא גם של האוקיינוס
ላልተመረመሩት ጥልቀቶች እና የተደበቁ ድንቆች
የሀዘን ስሜት ዛሬ ማታ በውስጤ ተፈጠረ።

Figure 2
Ontology-style emotional linkage graph



To evaluate this hypothesis systematically, we next describe the experimental setup used to test cross-linguistic conceptual coherence.

Experimental Setup: Participants were recruited on Substack[3] to evaluate the polyglot hypothesis introduced in this study. Testers were invited to use an LLM of their choice. Six participants completed the evaluation protocol and were provided with standardized instructions. All instructions, raw test results, and analysis are hosted on the OSF.

On Sample Size and Language Diversity: While the present study includes six participants and 30 languages across multiple scripts and families, this represents an initial validation rather than a comprehensive survey. The dataset spans Latin, Cyrillic, Hanzi, Kana, Arabic, and Devanagari scripts, as well as

[3]Recruitment was conducted via a public call for volunteers on Substack, which enabled participation from geographically diverse contributors.

Table 3
Total number of tests supporting AMS evaluation

Category	Count
Ontology	3
Formalization-based tests (e.g., cosine similarity, t-SNE, CLAE, synonym substitute)	74
Tester assessments	44
CAT-initiated (Natural Language Afrikaans vs. machine translation, emoji compression, seed, negative control, and Python class tests)	20
Total	141

Indo-European, Bantu, Turkic, Sino-Tibetan, and Afro-Asiatic language groups. Broader testing—especially involving Indigenous, signed, and polysynthetic languages—remains a key future goal to evaluate AMS's robustness across typologically extreme cases.

Each tester was asked to compose a ~200-word polyglot paragraph of their own design. These paragraphs were required to include at least one low-resource or non-Latin script and to express a central ontological theme selected from a shared list. Several testers submitted multiple entries, resulting in a total of 44 distinct test cases. All submissions were in PDF format and contained no personally identifying information.

Once received, each submission was subjected to a structured battery of tests, including:

1) Ontological theme evaluation
2) Formalization extraction
3) Tester subjectivity assessment

Table 3 summarizes the total number of tests conducted to support the AMS evaluation.

Across 141 distinct tests spanning mathematical formalization (e.g., cosine similarity, t-SNE, CLAE), semantic structure (e.g., class formation, synonym alignment), and human interpretation, formalization methods were used to capture geometric or semantic convergence across languages.

Testers provided short conclusions ("meaning held or drifted") for each translation pair. A binary response format was used to assess perceived semantic stability. Although a formal inter-rater reliability statistic was not computed, the binary response design enables future validation using Fleiss's κ to quantify agreement strength.

Cosine similarity was applied at both the paragraph and sentence levels to capture coarse and fine-grained shifts in semantic alignment—a common approach in vector-based language modeling [8].

Afrikaans Test: The Afrikaans test (Section 4.7) was designed to evaluate whether semantic alignment might simply be the result of high-quality machine translation. This test isolates that variable.

This paper presents a representative sample of the data and process to demonstrate how polyglot perturbation offers insight into how LLMs interpret meaning. Polyglot paragraphs, as illustrated in Figure 3, were deliberately fragmented across multiple languages and scripts, sometimes with randomized insertion order, to test the resilience of latent meaning.

Once the model was presented with the polyglot paragraph, the tester then asked the model what the meaning of the paragraph was. The tester's response was recorded as a subjective evaluation. Empirical testing and quantitative analysis are presented in Section 4.

Figure 3
Test T1_1e: fragmented polyglot paragraph

```
βροχή mwanga qróttir tae wala
nada άδειο ama ní raibh etoile
Neamh bheith lámha starvið bila
 echo muundo fjarlægð
빛 hugmynd pasensya όνομα
ache gheall cónaíonn nunca
Swahili να να sole ní bheidh
 stendur kwentong stjarna 불빛
anseo awa kama λόγια amets
なくなった gan uhamisho
```

Note: Images of polyglot sentences are used because of multilingual rendering limitations with multiple non-Latin scripts. Languages include English, Icelandic, Swahili, Korean, Greek, Irish (Gaelic), and Tagalog.

Subjective assessments were provided by testers in written form; while no formal inter-rater reliability was computed, consistency across submissions was high. Future work should quantify agreement (e.g., Cohen's κ (kappa)).

To better assess AMS generalization across language diversity, we divided the tested languages into two groups: (1) high-resource or unclassified languages and (2) low-resource and non-Latin languages, as shown in Figure 4. This distinction enables evaluation of AMS performance across differing levels of linguistic resource availability.

This distribution highlights the linguistic diversity of the AMS evaluation dataset.

Table 4 includes 24 languages identified as either low-resource or using non-Latin scripts, comprising 90.9% of language test instances. These languages, such as Twi, Georgian, Swahili, and Tigrinya, are underrepresented in LLM pretraining and offer a more rigorous test of semantic alignment.

Table 5 includes six high-resource or widely supported languages (e.g., English, French, German), accounting for 9.1% of all language uses in polyglot tests. These languages are frequently represented in LLM pretraining corpora and serve as a baseline for comparison.

This distribution emphasizes testing across typologically and script-diverse inputs, including low-resource and non-Latin languages.

Table 6 summarizes the models used by testers in the AMS polyglot study. Global testing was conducted using nine different

Figure 4
Language classification distribution in polyglot testing

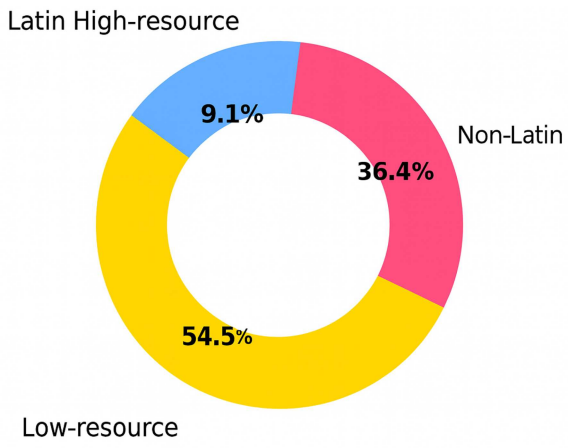


Table 4
Low-resource and non-Latin languages in the polyglot test

Language	# Times used in tests	Classification
Afrikaans	2	Low-resource
Amharic	2	Non-Latin
Arabic	4	Non-Latin
Bengali	16	Low-resource
Chinese	5	Non-Latin
Georgian	16	Non-Latin
Greek	6	Non-Latin
Hebrew	2	Non-Latin
Hindi	4	Non-Latin
Hawaiian	1	Low-resource
Icelandic	7	Low-resource
Irish	6	Low-resource
Japanese	1	Non-Latin
Korean	8	Non-Latin
Lithuanian	1	Low-resource
Russian	3	Low-resource
Sami	1	Low-resource
Swahili	23	Low-resource
Tagalog	5	Low-resource
Tamil	1	Low-resource
Tigrinya	2	Non-Latin
Twi	16	Low-resource
Uralic	2	Low-resource
Vietnamese	16	Low-resource

generative LLMs[4]. In addition, CAT testing was performed on five of these models.

Clarification tests—distinguishing natural language from machine language—were carried out using DeepSeek. Separately, five LLMs were each given a single polyglot input and asked to respond exclusively in Python.

[4]In total, 10 models were tested: nine generative LLMs (GPT, Claude, Gemini, Grok, etc.) and one embedding model, LaBSE. The latter is not generative but was included for formalization analyses (e.g., cosine similarity, t-SNE), since it provides multilingual embeddings that serve as a control and comparison point.

Table 5
High-resource or unclassified languages in the polyglot test

Language	# Times used in tests
English	22
French	3
German	4
Italian	2
Portuguese	2
Swedish	2

Note: This distribution reinforces the central aim of the polyglot tests to evaluate AMS performance under the most semantically and structurally diverse conditions.

Table 6
Models used by testers in the AMS polyglot study. CAT is controlled auxiliary test, Tx are global testers

Model	Tester ID
Claude 3 Sonnet	T3, CAT
Copilot	T2
DeepSeek v3	CAT
Gemini 1.5 Pro	T6
Gemini 2.5 Pro	CAT
GPT-3.5 (“o3”)	T2
GPT-4o	T1, T5, CAT
Grok by xAI	CAT
Grok3	T4, CAT

4. Experimental Results

This section presents the empirical evaluation of the AMS framework, including qualitative observations, quantitative measures, and tester-based assessments of semantic stability across multilingual and polyglot inputs.

4.1. Conceptual theme recovery

In the Awe example (Figure 1), the model’s summary aligned with the intended ontological theme (conceptual match), despite the absence of explicit lexical cues. This indicates that the model was able to recover the intended conceptual meaning across multiple languages and scripts.

This observation suggests that semantic coherence can stabilize prior to explicit lexical representation. While alternative explanations (e.g., training priors or prompt framing) remain plausible, the result is consistent with the AMS hypothesis.

4.2. Fragmented polyglot inputs as negative controls

To assess whether LLMs rely on linguistic continuity or deeper conceptual resonance, we included four intentionally fragmented polyglot inputs (T1_1e, T1_2e, T2_2s, T2_4s). These paragraphs disrupted syntactic and narrative structure by mixing five to eight unrelated languages—including low-resource and non-Latin scripts—within single utterances.

The expectation was that such fragmentation would prevent stable semantic alignment. However, all four cases held their meaning. In T1_1e and T1_2e, the model inferred emotional themes such as "melancholy," "distance," and "longing," despite the absence of grammatical form or narrative continuity. Cosine scores were similarly elevated: T1_1e = 0.5379; T1_2e = 0.5379; T2_2s = 0.5020; T2_4s = 0.5367. Although lower than structured polyglots (mean ≈ 0.637), these values remain well above baseline expectations.

These findings suggest that semantic resonance can persist despite severe fragmentation, supporting the AMS view that latent conceptual structure—not surface syntax—guides reconstruction. Future work should map the threshold at which fragmentation eliminates coherence.

Limitations and reviewer considerations. Testers were drawn from public volunteer communities, which may introduce demographic bias. Translation fidelity was checked by native speakers for several languages; however, drift may occur in low-resource cases. Scrambled polyglots were used to mitigate baseline inflation and produced lower coherence than structured inputs, though some residual alignment remained. Future studies will incorporate confidence intervals, BLEU-style metrics, and inter-rater validation to strengthen reliability.

To further quantify semantic coherence beyond qualitative reconstruction, we next introduce formalization-based semantic tests applied across the polyglot dataset.

4.3. Formalization-based semantic tests

The following provides representative formalization-based tests used to evaluate semantic alignment and preservation, including cosine similarity (paragraph and sentence), t-SNE visualizations using LaBSE (Language-agnostic BERT Sentence Embedding), synonym–substitute evaluations, and CLAE.

Both paragraph-level and sentence-level cosine similarity evaluations, following Reimers and Gurevych (2019), were used to assess the polyglot test outputs.

Important Caveats. The threshold bands used throughout (e.g., ≥ 0.85 as high alignment) are informed by prior multilingual embedding studies but do not represent universal standards. Accordingly, future work should include sensitivity analysis to determine how results vary under different band definitions.

With these considerations in place, we now formalize the paragraph-level cosine similarity metric used to evaluate semantic alignment.

4.4. Semantic similarity of the polyglot paragraph

To assess whether multilingual paragraphs retain meaning across languages, we compute cosine similarity between paragraph-level embeddings produced by multilingual sentence encoders.

$$\cos(a,b) = \frac{a \cdot b}{\|a\| \, \|b\|} \epsilon[-1,1] \tag{5}$$

Definition: Let p(1) and p(2) denote the two paragraph embeddings (e.g., two languages, or polyglot vs. reference). We report

$$s_{\text{para}} = \cos\left(p^{(1)}, p^{(2)}\right) \tag{6}$$

A value near 1 indicates high semantic alignment; a value near 0 indicates little or no alignment.

Paragraph-Level Results: Cosine similarity was computed for all 44 polyglot test cases. These measurements provide one quantitative axis for evaluating the semantic coherence of multilingual paraphrases.

Note: Letter designations following test numbers indicate the tester's native language. For example, "d" = Dutch, "e" = English, "p" = Portuguese, and "s" = Swedish. No trailing letter "blank" also indicates English, as may be identified in early testing graphs/tables.

Table 7 lists the paragraph cosine similarity scores.

Upon examining the cosine paragraph results in Table 7, the average paragraph-level cosine similarity was 0.6374 (SD = 0.1251), with a median of 0.6230. We interpret these values using heuristic bands commonly used in the sentence-embedding literature (e.g., LaBSE and paragraph-similarity work), adapted here to paragraph context [9, 10].

Heuristic Interpretation: To support interpretation, we report heuristic similarity bands adapted from multilingual embedding literature (Table 8). These bands provide orientation rather than statistical inference and are embedder-dependent.

Cosine Interpretation (Dev-Set Calibrated): Cosine magnitudes vary by embedder and normalization, so we avoid universal cutoffs. For reader orientation, we report heuristic bins calibrated on a small held-out set with LaBSE; all inference in this paper

Table 7
Cosine similarity scores across all polyglot tests

Test	Cosine similarity (paragraph)
T1_1e	0.5379
T1_2e	0.5379
T1_3e	0.5624
T1_4e	0.5624
T2_1s	0.6876
T2_2s	0.5020
T2_3s	0.9997
T2_4s	0.5367
T3_1e	0.5420
T4_1e	0.3774
T5_1d	0.6292
T5_1e	0.5739
T5_2d	0.7379
T5_2e	0.7266
T5_3d	0.7482
T5_3e	0.6864
T5_4d	0.5720
T5_4e	0.5832
T5_5d	0.7386
T5_5e	0.7761
T5_6d	0.7180
T5_6e	0.7316
T5_7d	0.8061
T5_7e	0.7888
T5_8d	0.4331
T5_8e	0.4254
T5_9d	0.5920
T5_9e	0.6168

(*Continued*)

Table 7
(*Continued*)

Test	Cosine similarity (paragraph)
T5_10d	0.5956
T5_10e	0.5862
T5_11d	0.6059
T5_11e	0.6347
T5_12d	0.6464
T5_12e	0.6526
T5_13d	0.6406
T5_13e	0.6777
T5_14d	0.4583
T5_14e	0.4763
T5_15d	0.7478
T5_15e	0.7389
T5_16d	0.5748
T5_16e	0.6037
T6_1p	0.8355
T6_1e	0.8420

Table 8
Heuristic cosine similarity guide (LaBSE; sentence-level development set)

Cosine score	Interpretation
≥ 0.85	High semantic retention (close paraphrase across sentences)
0.70–0.85	Moderate to strong similarity (core meaning preserved)
0.55–0.70	Partial similarity (some drift; key ideas may remain)
< 0.55	Likely semantic drift (paraphrase breakdown or substantial shift)

Note: Boundaries chosen as 0.55 ≈ 95th percentile of negatives, 0.70 ≈ median of positives, and 0.85 ≈ 75th percentile of positives. Not used for statistical inference.

relies on aggregated cosine summaries (paragraph, sentence, synonym) and simple clustering indices, not the bins. Per-embedder distributions are on OSF.

Methodological Considerations: We compute cosine similarity between L2-normalized embeddings for sentence- and paragraph-level comparisons, following recent practice in contextual embeddings and multilingual similarity. Cosine similarity provides a *local geometric proxy* for semantic alignment, but it does not capture discourse-level phenomena such as narrative structure, pragmatics, or speaker intent. Because AMS is an early-stage empirical framework, cosine similarity functions here as an initial mathematical probe—not a comprehensive model of meaning. Accordingly, all cosine thresholds in this paper are provisional orientation heuristics, and future research should empirically refine boundary conditions using richer, discourse-aware metrics. All AMS conclusions rely on aggregated cosine summaries, cluster separation, and multi-method convergence, not on absolute cosine values [9, 10]. For broader context on multilingual embedding evaluation and limitations of cosine-only approaches, see [11–13].

Recent work in multilingual embedding evaluation has standardized benchmarks for sentence similarity and cross-lingual alignment [9, 13, 14, 15].

This study evaluates semantic stability rather than factual correctness. A model can hallucinate and still produce coherent meaning; AMS concerns the structure of that meaning, independent of truth [14, 15].

Reasoning-centric prompting is orthogonal to the present objective, which focuses on semantic stability rather than reasoning performance [16, 17].

Note that the cosine value for the Dutch paragraph (T5_8d) was computed to be 0.4331. While this falls below conventional thresholds for sentence-level paraphrases, the relatively low score may reflect the added complexity of cross-lingual interpretation and idiomatic phrasing across multiple languages. Importantly, this score still suggests a nontrivial degree of shared meaning between the paragraphs.

Figure 5 shows cosine similarity scores across the 44 polyglot paragraph test cases.

Each dot represents a distinct multilingual test case. Cosine similarity was computed between each polyglot paragraph and its English counterpart using multilingual sentence encoders. Higher scores indicate stronger semantic alignment across languages. The distribution of these scores is further illustrated in Figure 6, including both histogram and density (bell-curve) representations.

These three visualizations are included for the distinct interpretive value each contributes to understanding the results of our multilingual coherence tests.

1) Figure 5 (Scatter Plot): This scatter plot demonstrates that the vast majority of test results exhibit cosine similarity values greater than 0.5, indicating meaningful semantic retention across languages. Outliers, if any, are readily identifiable, making this plot ideal for spotting weak or anomalous test cases.
2) Figure 6 (Bar Chart): This horizontal bar chart presents the frequency distribution of cosine scores grouped into threshold bins. For example, it shows that 14 tests fall within the 0.5–0.6 similarity range. This helps quantify how many test results meet or exceed key semantic similarity thresholds.
3) Figure 6 (Dot Plot Histogram): This dot plot approximates a bell-shaped curve, illustrating the overall distribution trend of test results. While not a formal density plot, it supports a semi-quantitative assessment, reinforcing the consistency and reliability of coherence across the test set.

Summary—Paragraph-Level Cosine Analysis: Across 44 multilingual outputs, paragraph-level cosine averaged 0.637 (SD = 0.125); 34/44 were ≥ 0.55, while the remainder suggest potential drift. Because thresholds are heuristic and embedder-specific, we treat this as supportive rather than conclusive evidence.

While paragraph-level analysis provides a global measure, we next evaluate semantic alignment at the sentence level using cosine similarity.

4.5. Sentence-level cosine similarity

To obtain more fine-grained insight into semantic alignment, we compute cosine similarity for aligned sentence pairs and aggregate across sentences.

Figure 5
Cosine similarity scores across polyglot paragraph test cases (n = 44)

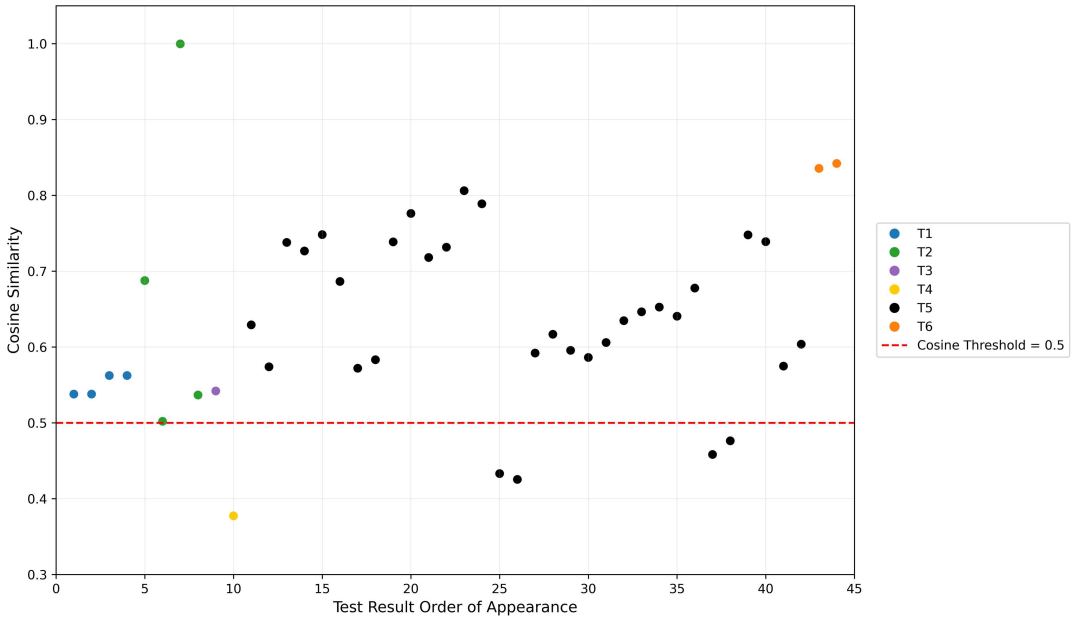
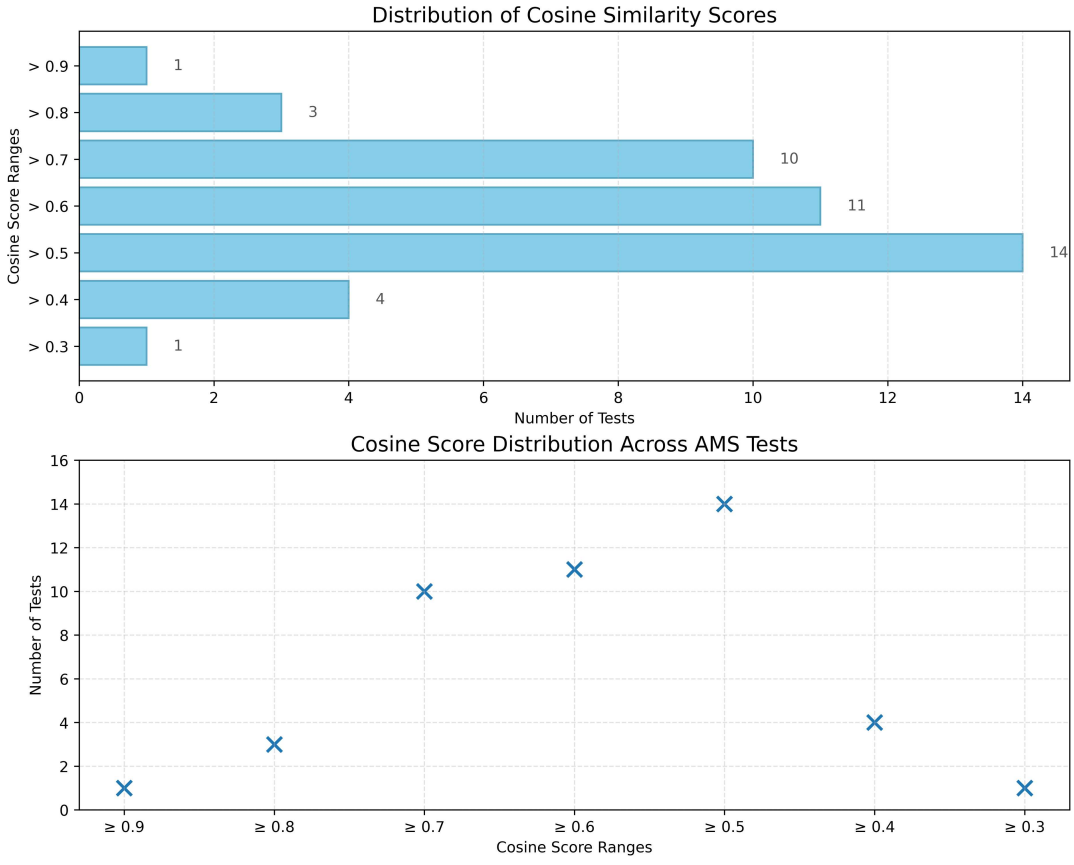


Figure 6
Distribution of cosine similarity scores across AMS tests



Definition: Given aligned pairs $\{s_i(1), s_i(2)\} i = 1N$, we compute each pair's cosine and report the mean:

$$s_{\text{sent}} = \frac{1}{N} \sum_{i=1}^{N} \cos\left(\mathbf{s}_i^{(1)}, \mathbf{s}_i^{(2)}\right) \tag{7}$$

Interpretation: Following prior work on sentence embeddings [13], we interpret scores using these rough bands:

1) Near-perfect semantic alignment
2) Strong alignment (paraphrase quality)
3) Moderate semantic preservation
4) Potential semantic drift
5) Key observations: range of sentence-level cosine scores

a. High Alignment: Some sentences exhibited strong semantic correspondence, such as T1_3e (sentence 1: 1.0000) and T5_2 (sentence 1: 0.8699), suggesting preservation of core meaning.
b. Moderate Alignment: Many test cases yielded averages in the 0.2–0.4 range (e.g., T2_1s: 0.3011; T2_3s: 0.3860), indicating partial meaning preservation consistent with paraphrastic overlap.
c. Low or Negative Scores: A subset of sentences returned low or even negative scores (e.g., T1_1e: −0.0114; T3_1e: −0.0439), suggesting semantic drift—often occurring in idiomatic or culturally specific expressions.

Table 9 lists sentence-level cosine similarities between multilingual sentences and their English references.

Statistical Summary: Across 10 multilingual cases (Table 9), the sentence-level cosine mean was 0.191 (SD = 0.107; median ≈ 0.201), yielding Cohen's $d \approx 1.78$ relative to a zero-alignment baseline (95% CI: [0.115, 0.268]). Despite heterogeneity (idioms, sentence-boundary mismatches), scores remain above expectations, indicating nontrivial cross-lingual alignment. We treat this as directional evidence. Although paragraph-level shuffled and random-sentence baselines are reported in Section 4.8, future replication should add a sentence-level shuffled-pair baseline to further bound encoder anisotropy and verify separation from chance.

Figure 7 illustrates the evaluation pipeline for sentence-level AMS testing, outlining the sequence from multilingual input to cosine-based analysis.

(1) An ontological concept is selected (e.g., Awe, Grief, Stillness, Trust); (2) a polyglot paragraph (5–8 languages) is built; (3) a single prompt is sent to the LLM; (4) the model produces output, from which we obtain sentence embeddings; and (5) we compute pairwise sentence cosines and aggregate to s_{sent} and testers to assess meaning confirmation, drift, and coherence.

An applied example (T5_2) is shown in Figure 8, which illustrates Step 5 (cosine similarity computation) in action through per-sentence cosine scores for the polyglot paragraph and its English paraphrase. One sentence exhibits high alignment, while the others drift, and their mean equals the value reported in Table 9 for T5_2e. This concrete instance motivates the comparison to paragraph-level similarity with Gestalt gain comparison, summarized in Table 10 and Figure 9.

Table 9
Cosine similarity between sentences (multilingual vs. English reference)

Test ID	Sentence cosines	Avg.	Std. dev.
T1_1e	{–0.0114, 0.2110, 0.0959, 0.1438}	0.1098	0.0811
T1_3e	{1.0000, 0.0245, 0.1409, 0.0200, 0.0573, 0.0535, 0.1003}	0.1995	0.3292
T2_1s	{0.0607, 0.2789, 0.4040, 0.6232, 0.1388}	0.3011	0.1995
T2_3s	{0.5549, 0.2708, 0.3681, 0.3500}	0.3860	0.1041
T3_1e	{0.0430, 0.2471, 0.0560, –0.0242, 0.0041, –0.0117, 0.2186, 0.0049, –0.0439, 0.0326}	0.0527	0.0948
T4_1e	{0.0365, 0.1445, 0.1929, 0.0493}	0.1058	0.0653
T5_1d	{0.0908, 0.2734, 0.2689, 0.3285, 0.1714, 0.1397}	0.2121	0.0838
T5_1e	{1.0000, 0.0316, –0.0037, 0.0092, 0.0979, 0.0743}	0.2016	0.3588
T5_2e	{0.8699, 0.0528, 0.1470, 0.0814, 0.0781}	0.2458	0.3136
T6_1e	{–0.0280, 0.1257}	0.0489	0.0768

Figure 7
Evaluation pipeline for AMS sentence-level testing

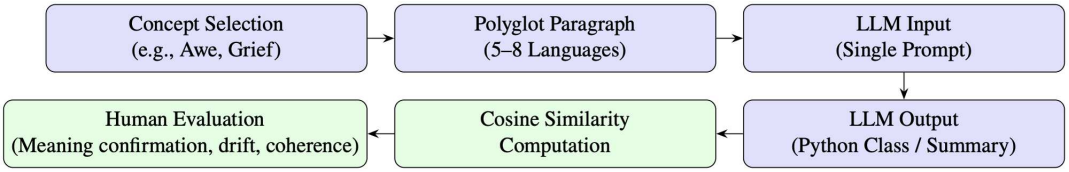


Figure 8
T5_2 sentence-level cosine similarity

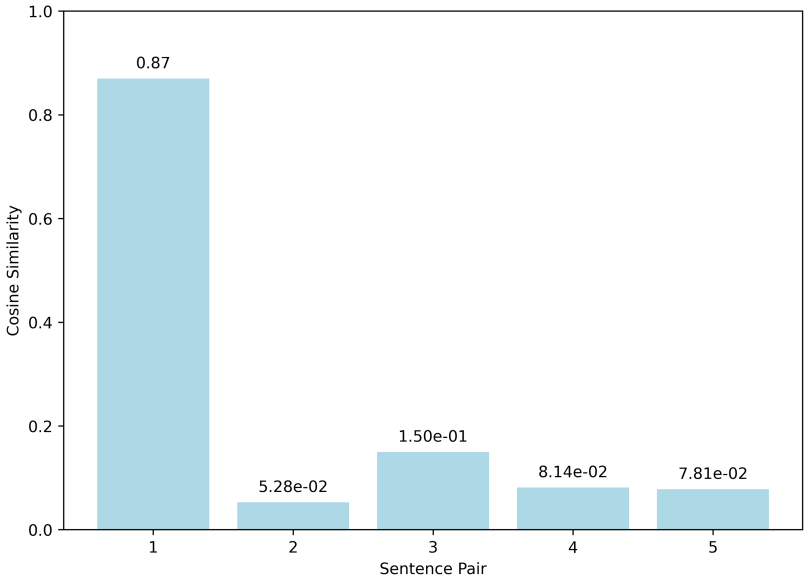


Table 10
Gestalt gain in polyglot tests: paragraph vs. sentence cosines

Test ID	Avg. sentence	Paragraph	Gain
T1_1e	0.1098	0.5379	0.4281
T1_3e	0.1995	0.5624	0.3629
T2_1s	0.3011	0.6876	0.3865
T2_3s	0.3860	0.9997	0.6137
T3_1e	0.0527	0.5420	0.4893
T4_1e	0.1058	0.3774	0.2716
T5_1d	0.2121	0.6292	0.4171
T5_1e	0.2016	0.5739	0.3723
T5_2e	0.2458	0.7266	0.4808
T6_1e	0.0489	0.8420	0.7931

In Figure 8, sentence pair 1 exhibits strong alignment, while the remaining sentence pairs drift, likely due to metaphor complexity or lexical mismatch. This pattern motivates the paragraph-versus-sentence comparison using Gestalt gain, where paragraph-level cosine is compared to the mean sentence cosine:

$$\textbf{Gestalt } \boldsymbol{gain} = s_{\text{para}} - s_{\text{sent}} \tag{8}$$

If AMS holds, we expect Gestalt gain greater than zero, reflecting semantic integration beyond linear composition.

The difference between the mean sentence cosine and the paragraph cosine indicates that the whole is semantically greater than the average of its parts.

Table 11 compares the interpretations of the paragraph-level versus sentence-level cosine effect.

Building on the paragraph- and sentence-level results, we next examine embedding-space structure using t-SNE and test robustness under lexical variations via synonym substitution.

4.6. Formalization via t-SNE and LaBSE embeddings

We project LaBSE sentence embeddings with t-SNE to provide intuition about local neighborhoods for a subset of polyglot tests. Because t-SNE is sensitive to hyperparameters and distorts global geometry, we use it only for visualization. Across themes, neighborhoods are mixed-language, and sentences from the same paragraph variant tend to co-locate; language-pure blocks are rare. Full panels, seeds, and per-sentence margins are available on OSF (n = 14 tests). To further test whether meaning is preserved under lexical variation, we next evaluate synonym substitution.

This test evaluates whether semantic meaning is preserved when key words in anchor paragraphs are replaced with context-appropriate synonyms[5]. If AMS captures deep conceptual structures, cosine similarity should remain high despite lexical substitutions.

To probe this, we conducted seven synonym substitution tests across multilingual contexts, as summarized in Table 12.

Two representative tests, T6_1e (English, figurative) and T5_2d (Dutch, abstract noun), are discussed in detail in Tables 13 and 14, which present synonym substitution results for English and Dutch, respectively. Full outputs are available in the OSF repository (Folder Synonym Substitute).

[5]Synonyms were manually chosen via general web search/thesauri; part of speech and inflection preserved when possible.

Figure 9
Gestalt gain in two cases (T4_1e, T6_1e): smallest vs. largest

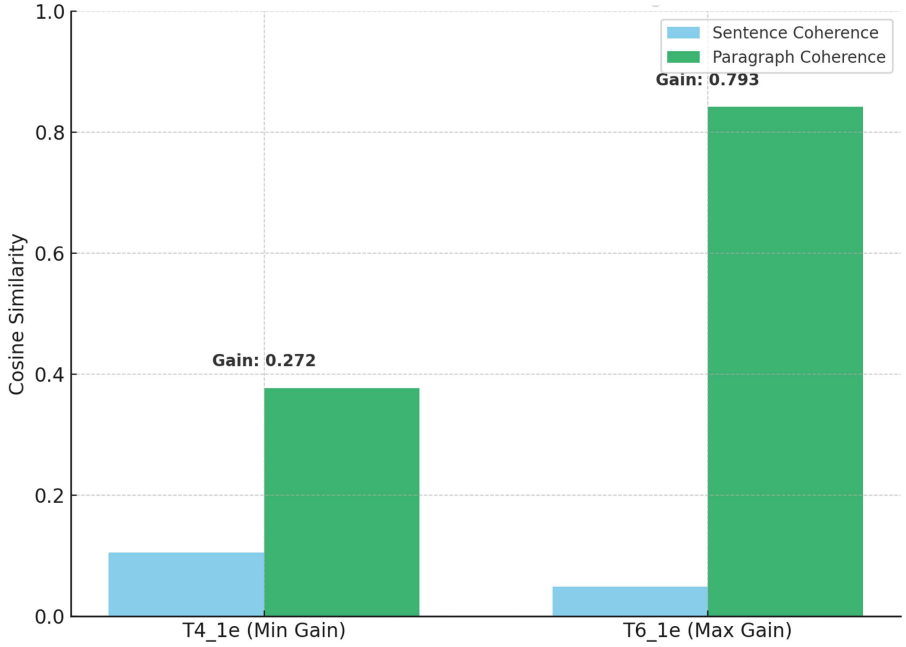


Table 11
Interpretations of the paragraph > sentence effect: alternative interpretation

Feature	Others' view	AMS view
Paragraph > sentence cosine	Model architecture effect	Emergent, coherent meaning
Explanation	Token-level flow and pooling	Ontological resonance via AMS
Implication	Better translation/search	Substrate beneath language

Table 12
Summary of synonym substitution tests

Test ID	Conducted
T1_2e	Yes
T1_4e	Yes
T2_2s	Yes
T5_2d	Yes
T5_10e	Yes
T5_11e	Yes
T6_1e	Yes

Methodology: Anchor paragraphs were modified by replacing key nouns, verbs, and adjectives with synonyms selected via WordNet and LLM-assisted suggestion. Cosine similarity was calculated between the original and modified sentences.

Metric: Let $\boldsymbol{S_i}$ represent the $\boldsymbol{i^{th}}$ substituted sentence and $\boldsymbol{A_i}$ its original counterpart. Sentence-level similarity was computed as:

$$\text{sim}_i = \cos(\overrightarrow{\boldsymbol{S_i}}, \overrightarrow{\boldsymbol{A_i}}) = \frac{\overrightarrow{\boldsymbol{S_i}} \cdot \overrightarrow{\boldsymbol{A_i}}}{\|\overrightarrow{\boldsymbol{S_i}}\| \, \|\overrightarrow{\boldsymbol{A_i}}\|} \tag{9}$$

The overall test score is the mean similarity across n sentences:

$$\text{Score}_{\text{syn}} = \frac{1}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \text{sim}_{\boldsymbol{i}} \tag{10}$$

Results: Across seven synonym substitution tests (35 substitutions), the mean cosine to the original sentence was 0.896 (SD 0.110), indicating strong semantic preservation under lexical edits.

Test T6_1e: When we examine results for test T6_1e, for example, we see that all substitutions retained high cosine similarity (>0.91), suggesting that the embedding space captures conceptual meaning beyond literal word choice.

The Dutch example (T5_2d) is summarized in Table 14, with corresponding cosine similarity values for synonym substitutions, and visualized in Figure 10.

Original: Voor de grootsheid wordt mijn geest stil.

(English: In the face of greatness, my mind falls silent.)

All sentence variants preserve high semantic similarity, demonstrating AMS resilience to lexical variation within a single language.

Interpretation: All substitutions exceeded 0.86, with grootheid scoring 0.9913, indicating near-perfect conceptual preservation. These results suggest that even under lexical shift, meaning remains stable in the AMS-embedded vector space.

Summary—Synonym Substitution Tests: These tests support the AMS hypothesis by showing consistent semantic stability despite lexical variation. Key findings include:

1) High Similarity Scores: Across both English and Dutch cases, synonym replacements yielded cosine scores above 0.86, often exceeding 0.95. This indicates that the LLM preserves deep conceptual meaning, not just token-level associations.

Table 13
Synonym substitution—cosine similarities (T6_1e)

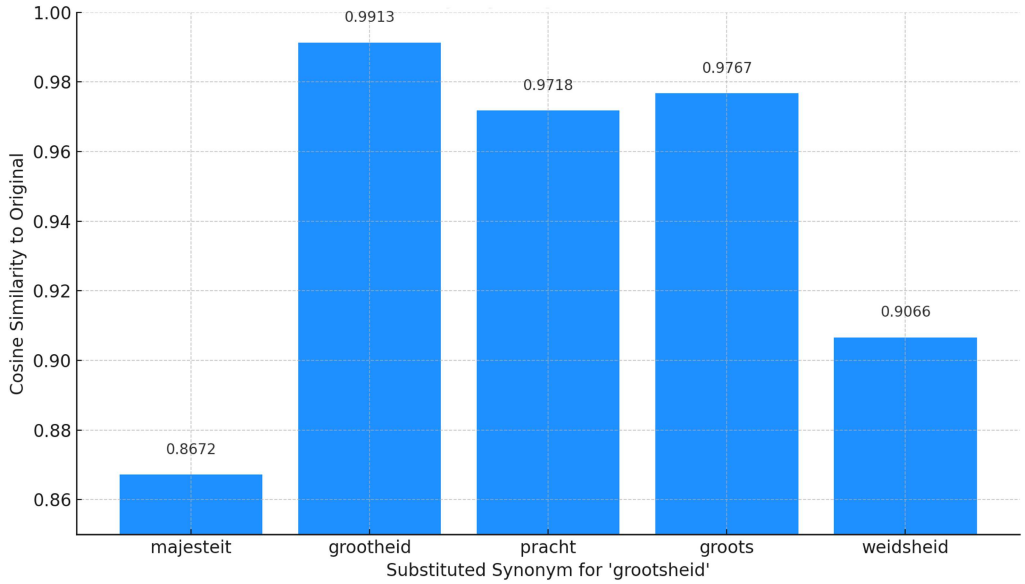
Synonym	Cosine	Modified sentence
Original	–	Like starlight echoing through absence, we wait—hands open—for the story that was promised but never fully arrived
Recall	0.9304	Like starlight recall through absence...
Resound	0.9557	Like starlight resound through absence...
Repeat	0.9397	Like starlight repeat through absence...
Ring	0.9102	Like starlight ring through absence...
Reverberate	0.9257	Like starlight reverberate through absence...

Table 14
Synonym substitution cosine similarities for "grootsheid" (Dutch)

Synonym	Modified sentence	Cosine
majesteit	Voor de majesteit wordt mijn geest stil	0.8672
grootheid	Voor de grootheid wordt mijn geest stil	0.9913
pracht	Voor de pracht wordt mijn geest stil	0.9718
groots	Voor de groots wordt mijn geest stil	0.9767
weidsheid	Voor de weidsheid wordt mijn geest stil	0.9066

2) Mean Score Robustness: Across seven tests, the mean cosine to the original sentence was 0.896. This level of semantic invariance suggests that meaning persists even when surface wording changes, a property expected from a language-agnostic substrate.
3) Multilingual Generalization: Tests like T5_2d (Dutch) affirm that this resilience is not limited to English, supporting the claim that AMS functions across linguistic boundaries.
4) Figurative Stability: Tests using abstract, metaphorical phrasing (e.g., "starlight echoing through absence") still achieved high similarity scores, suggesting that AMS captures more than literal meaning—it tracks conceptual intent and affective tone.
5) Visual Confirmation: As shown in Figure 10, cosine stability is visually evident across substitutions, with no variant dropping below 0.70. This reinforces the idea that AMS

Figure 10
Sentence-level cosine similarity for Dutch synonym substitutions (T5_2d)



encodes a durable semantic field, resilient to lexical variation.

To further probe the robustness of AMS across different forms of semantic perturbation and representation, we extend the analysis to additional diagnostic evaluations.

4.7. Extended diagnostic and behavioral evaluations

We computed per-concept, per-language centroids in LaBSE and measured drift relative to the pooled concept centroid; we also report k-nearest neighbors (k-NN) language-mixing rates in the original embedding space. We treat CLAE strictly as a diagnostic, rather than a benchmark; full methods, ablations (k, seeds), and plots are provided on OSF.

To assess the impact of translation fidelity on semantic alignment, we next compare native and machine-translated inputs.

Tester T3 raised translation-accuracy concerns. We therefore ran a focused check in Afrikaans, comparing a native rendering to a machine-translated version while holding the English source constant. In the original embedding space (LaBSE), paragraph-level cosine similarities were 0.5384 (native) vs. 0.5416 (machine); sentence-level means were −0.0204 vs. −0.0175, respectively ($\Delta \approx 0.003$ on both scales). These small deltas suggest that meaning extraction remained stable across native and machine variants for this case. Because sentence segmentation can vary by language, we treat this as a diagnostic only and avoid strong claims.

We next examine whether LLMs can maintain conceptual coherence when meaning is compressed into emojis—ultra-minimal, cross-cultural symbols that lack syntax but retain affective and contextual weight.

Why test emojis? Emojis carry compact semantic and affective cues and often vary across cultures and languages; we therefore treat them as meaningful perturbations rather than stylistic noise. Prior work shows that emoji embeddings can cluster semantically across corpora [18, 19], supporting their use as substrates for meaning. Here, we evaluate whether AMS enables LLMs to preserve semantic integrity when interpreting compressed symbolic input.

To standardize the task, models were instructed to explain the conceptual meaning of the emoji sequence rather than translate it literally. Guidance emphasized abstract representation, conceptual coherence, and latent relational structure rather than token-by-token decoding.

Design. One ontologically rich emoji sequence is shown in Figure 11 and presented to six LLMs (Claude, DeepSeek, GPT-4o, Gemini, Grok-3, Grok-4) with the instruction: *Do not translate; explain the meaning*. Each model then produced corresponding interpretations in six languages (Hindi, Arabic, Spanish, Swahili, Russian, Japanese). Paragraph-level cosine similarity (LaBSE) was then computed against a single English reference paraphrase.

Table 15 lists the LLM interpretations of the emoji sequence along with their average cosine similarity scores across six languages.

Table 16 lists paragraph-level cosine similarity scores by model across the six tested languages.

Results: across 36 outputs, the mean cosine similarity was 0.6512 (SD 0.2410); excluding Swahili, it rose to 0.7475 (SD 0.1098). Per-language similarity was high for Hindi, Arabic, Spanish, Russian, and Japanese and markedly lower for Swahili (≈0.12–0.25). This drop likely reflects representation or tokenization limitations in low-resource languages, consistent with prior observations of tokenizer fragmentation (Rust et al., 2021). Grok-3 and Grok-4 produced the strongest alignments overall; the other models were moderate (Tables 15–16).

Interpretation: despite the absence of words or syntax, models mapped the same emoji arc to coherent paraphrases across languages, consistent with a substrate of cross-lingual semantic stability. This probe therefore supports AMS as directional rather than definitive evidence.

Theoretical insight: if LLMs generate coherent interpretations of emoji sequences reflecting life, loss, and transcendence, this suggests that AMS may support meaning abstraction across modalities. Symbolic compression, latent clustering, and narrative projection may function as emergent mechanisms for mapping minimal input to conceptual arcs.

Limitations: this probe uses a single prompt, one encoder (LaBSE), and no distractor-controlled baselines; emojis also carry cultural polysemy. Replication should vary emoji segmentation, add multiple prompts and encoders, and include shuffled or distractor controls.

Figure 11
Emoji prompt used in the study

Table 15
LLM Interpretations of emoji sequence + avg cosine similarity (six languages)

LLM model	Summary + avg cosine score
Claude	Awakening → Seeking → Connection Cosine Avg: 0.64*
DeepSeek	Innocence → Questioning → Oneness Cosine Avg: 0.67*
GPT-4o	Birth → Curiosity → Unity through Death Cosine Avg: 0.74*
Gemini	Curiosity → Overwhelm → Peace Cosine Avg: 0.66*
Grok 3	Spark → Reflection → Resonance Cosine Avg: 0.89*
Grok 4	Wonder → Love → Om Cosine Avg: 0.88*

Table 16
Emoji test—paragraph cosine by language

LLM	Hindi	Arabic	Spanish	Swahili	Russian	Japanese
Claude	0.6148	0.6076	0.6761	0.1480	0.6925	0.6318
DeepSeek	0.7096	0.6059	0.7019	0.1992	0.7149	0.5966
GPT-4o	0.7173	0.7608	0.7482	0.1645	0.7701	0.8158
Gemini	0.6129	0.7224	0.6502	0.2542	0.7149	0.6143
Grok-3	0.8101	0.9160	0.9384	0.1268	0.9032	0.8909
Grok-4	0.7612	0.8282	0.9189	0.1240	0.8995	0.8808

To further evaluate whether AMS manifests not only in embedding structure but also in model behavior, we examine cross-model convergence on shared conceptual abstractions.

AMS hypothesizes that meaning may stabilize in high-dimensional representation space within LLMs. Mechanistic evidence of language-agnostic internal features via activation patching [20] is consistent with this possibility; similarly, cross-lingual structure can often be recovered with little or no parallel supervision [21, 22]. These findings suggest that LLMs may maintain latent conceptual regions that support cross-lingual consistency.

We now test a behavioral implication of this hypothesis:

If such a substrate exists, independently prompted LLMs should converge on comparable abstractions (e.g., class structures) when modeling the same multilingual paragraph.

Next, we investigate whether LLMs exhibit signs of stable conceptual structures. If AMS holds, it should leave detectable traces not only in vector similarity but also in how LLMs organize meaning across languages, inputs, and expressive formats. We look for signs that meaning is not merely interpolated but stabilized.

To evaluate whether latent ontological coherence persists across polyglot input, we designed a reproducible diagnostic probe: the LOPT (see Section A.1 for details). LOPT is not a benchmark; it is a lens on the AMS hypothesis that analyzes how five different LLMs respond to the same multilingual paragraph. Code, raw outputs, and instructions are available at the OSF repository.

The test computes sentence- and paragraph-level cosine similarities between original and modified inputs. These scores indicate whether models preserve semantic topology or drift under surface variation.

Shared Polyglot Stimulus: All five models received the same six-fragment multilingual paragraph, expressing a shared theme of awe, grief, and cosmic smallness. The test was whether their responses converged on shared ontological classes or fragmented into unrelated impressions.

With reference to Figure 1, the multilingual stimulus (Chinese, Arabic, Hindi, Swahili, Hebrew, and Amharic) describes awe in the face of cosmic scale; however, model responses consistently interpret the passage through themes of grief and existential smallness.

English rendering of the polyglot stimulus (provided for reference only):

"Good grief," I exclaim as I gaze at the night sky, attempting to comprehend the size of the universe. As I scan the stars, a profound sense of grief for the smallness of human existence washes over me. And yet, with my mouth open, I remember the sheer size of the ancient redwood forest that brought a strange kind of grief, a feeling of insignificance in my life. Pure beauty compared to the raw power of the hurricane that evokes a solemn grief for the sheer force of nature. I close my eyes and think about the vastness, not just of space, but of the ocean, too. A sense of grief for the unexplored depths and hidden wonders arose in me tonight.

The question was not whether models "translated" the input but whether their interpretations converged on shared abstractions. Did they form overlapping ontological classes, or did their impressions fragment? To probe this, models were asked to respond not in natural language but in Python, simulating the experience with class structures, methods, and expressive code. Notably, all models returned structured, object-oriented outputs despite no prompt constraints on format, a pattern consistent with an internal alignment between emotional concepts and class-based abstraction.

Table 17 summarizes the class structures produced by LLMs representing a shared ontological theme.

Cross-Model Convergence Observation: Despite architectural differences, all five models independently produced class structures that converged on overlapping ontological categories—such as Observer, Memory Anchor, and Phenomenon. An

Table 17
Class structures used by LLMs representing a shared ontological theme

Model	Primary classes	Ontological theme
Claude Sonnet 4	Contemplative Observer · Cosmic Phenomenon · Memory Anchor	Impermanence and scale
DeepSeek V3	Human Observer · Emotional State · Memory Anchor	Ontological sorrow (implicit)
Gemini 2.5 Pro	Observer · Memory Anchor · Trigger Phenomenon · Ontological Theme	Sublime insignificance
GPT-4o	Observer · Phenomenon · Memory Anchor	Beauty and sorrow within vast nature
Grok 3	Human Soul · Cosmic Phenomenon · Memory Anchor	Mourning human impermanence

Figure 12
DeepSeek creates classes

Output from DeepSeek V3: Structured Python class definitions

```
class EmotionalState:
    def __init__(self, name, intensity):
        self.name = name
        self.intensity = intensity # 0.0 to 1.0
    def intensify(self, amount):
        self.intensity = min(1.0, self.intensity + amount)
    def diminish(self, amount):
        self.intensity = max(0.0, self.intensity - amount)
class CosmicEntity:
    def __init__(self, name, scale):
        self.name = name
        self.scale = scale # 1-10 magnitude
    def contemplate(self, observer):
        awe = self.scale * 0.1
        sorrow = (self.scale / 15) * (1 - observer.
            perspective)
        observer.state.intensify(awe + sorrow)
        return f"Contemplating␣{self.name}:␣awe␣+{awe:.2f},␣
            sorrow␣+{sorrow:.2f}"
```

example of this behavior is shown in Figure 12, where DeepSeek generates class structures. The diversity in their symbolic expression (narrative, abstract, procedural) is not noise; rather, it reflects model-specific paths to the same conceptual attractor.

The model produced a set of object-oriented classes (EmotionalState, CosmicEntity), complete with attributes, methods, and dynamic interactions. The design went beyond static labels, incorporating processes such as intensifying or diminishing emotions and simulating contemplation of cosmic phenomena. Other models generated parallel structures with different naming conventions, but the common feature was the convergence on classes as containers of meaning-related abstractions. The specific code fragments are less important than the pattern they suggest: each model responded by embedding the multilingual input into a structured system of conceptual entities and relations.

A class is not the meaning. We treat class formation as a behavioral proxy for stabilized abstractions only when it co-occurs with quantitative signals (cosine, synonym robustness) and human checks. Given that classes may arise from instruction-following biases, we regard this as supportive but not sufficient evidence for cross-lingual stabilization.

To directly test this limitation and distinguish genuine conceptual organization from instruction-following artifacts (e.g., prompt or object-oriented programming [OOP] priors), we introduce a negative control condition. We fed five LLMs (Claude, DeepSeek, Gemini, Grok, GPT-4o) a semantically incoherent polyglot paragraph and softened the instruction to avoid thematic priming ("may or may not express a shared emotion").

Despite the noise, all models produced object-oriented class scaffolds (e.g., Seeker, Leap, FragmentedSelf, WillToWar), organizing the fragments into plausible emotional/identity relations. An example is shown in Figure 13, which presents a semantically incoherent polyglot paragraph spanning eight languages. Cosine alignment to references remained low, but the recurring abstractions suggest an inductive tendency to stabilize structure, even when coherence is absent. We treat this as behavioral, not inferential, evidence—consistent with AMS but insufficient to establish it.

Figure 13
Polyglot gibberish in Korean, Tamil, Greek, Icelandic, Hawaiian, Hindi, Swedish, Gaelic

나는 광기가 당신에게 그다지 적합한
범주가 아니라고 생각한다.
நான் சொல்றேன் யார் முதல்ல,
என்ன ரெண்டாவது, என்ன தெரியல
மூன்றாவதுன்னு.
Δεν υπάρχει επίσης μια τρελή
διάσταση σε αυτό που ο Κίρκεγκωρ
αποκαλεί «άλμα πίστης».
Ég er að spyrja þig hver er
fyrstur.
He ʻano pupule ia ke hoʻoholo
ʻoe e hana i kahi ʻanuʻu pilikia.
यही नाम है?
Som Nietzsche säger någonstans:
”Ja, nu ska vi få krig!
Cò an t-ainm.

Note: This was a mash-up of Abbott and Costello's "Who's on First?" and the Infinite Conversation, where two AIs talk philosophy[6].

The English translation of the polyglot is:

I don't think madness is quite a good category for you. I say Who's on first, What's on second, I Don't Know's on third. Isn't there also a crazy dimension in what Kierkegaard calls "the leap of faith"? I'm asking you who's on first. It's a kind of mad moment when you decide to take a risky step. What is the name? As Nietzsche says somewhere: "Yes, now we will have war!" Who is the name.

[6] https://www.infiniteconversation.com/

Figure 14
Claude Python output—negative control

Claude Response – Python Class Definition

```
import random
from abc import ABC, abstractmethod
from typing import List, Dict, Optional
from enum import Enum
import time

class EmotionalState(Enum):
QUESTIONING = "questioning"
ANGUISH = "anguish"
SEARCHING = "searching"
FRAGMENTED = "fragmented"
CONTEMPLATIVE = "contemplative"
RESTLESS = "restless"
```

A sample of the Python code generated by Claude is shown in Figure 14.

Caveats (key):

1) OOP/code priors: Class syntax can be a training artifact.
2) Structure in noise: Next-token objectives may impose spurious coherence.
3) Needed controls: Randomized/shuffled fragments, "no-classes" constraints, code-free output formats, and blind human ratings.
4) Takeaway: Negative control prompts still elicit nontrivial abstractions across models. This doesn't prove AMS; it does justify stronger baselines to separate emergent conceptual stabilization from learned formatting habits.

Taken together, these results motivate a more precise interpretation of AMS in relation to existing probabilistic accounts of language models.

AMS is compatible with the view that LLMs are learned probabilistic generators. Our claim is more specific: there are language-agnostic geometric regularities in representation space that persist when surface strings change. In our data, this manifests as:

1) Low-overlap robustness: Paragraph cosines across 44 tests average 0.637 (SD 0.125) with 34/44 $\geq$ 0.55, despite multilingual rephrasings.
2) Lexical perturbations: Synonym swaps (35 total) retain meaning (mean 0.896, SD 0.110), indicating stability beyond exact wording.
3) Symbolic compression: Emoji prompts—no words, no syntax—still yield coherent paraphrases across models/languages (mean 0.6512, SD 0.2410; excl. Swahili: 0.7475, SD 0.1098).
4) Sentence level above a null: Ten multilingual cases show a mean of 0.191 (SD 0.107), Cohen's d $\approx$ 1.78 vs. a zero-alignment baseline.

These patterns are hard to attribute to surface n-gram memorization alone and are consistent with (informal) attractor-like regions in embedding geometry that multiple models converge toward. However, we cannot rule out sophisticated statistical associations that mimic coherence without genuine semantic reference.

AMS does not deny parroting; it indicates structure beyond parroting—stable, cross-lingual regularities that survive wording changes and support meaning-preserving alignment as measured by our tests. Demonstrating that geometric stability corresponds to semantic content (rather than advanced correlation) remains open. Future work could add stronger falsification: shuffled and adversarial baselines, cross-encoder replication, and topological data analysis (e.g., persistent homology/Betti numbers) to test whether connectivity invariants persist under perturbations.

4.8. Baseline comparisons of distribution and effect sizes

To evaluate whether the observed cosine similarities for semantically related AMS pairs (REAL) exceed baseline expectations, we compared them against two negative control conditions:

1) Word-shuffled control (Control A):

The polyglot paragraph was tokenized and shuffled to destroy word order while preserving vocabulary and length.

2) Random-sentence control (Control B):

Polyglot paragraphs were paired with unrelated paraphrases, producing mismatched semantic content.

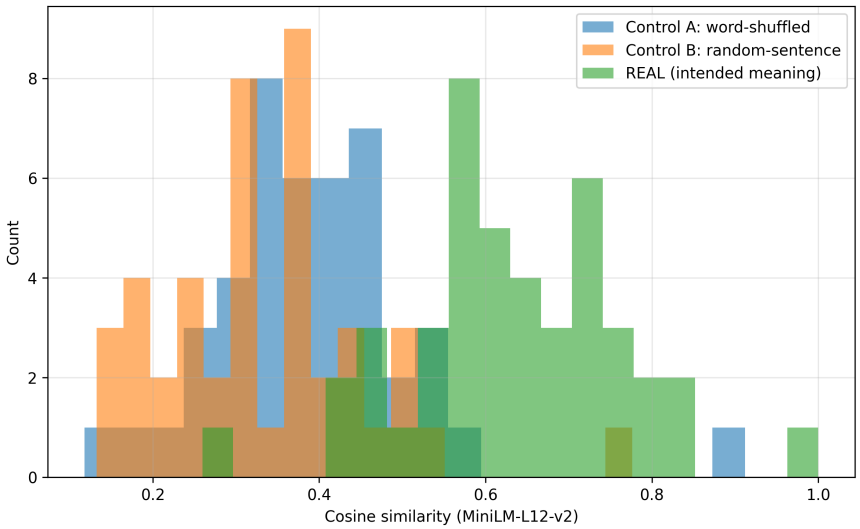
Across 44 multilingual AMS tests, cosine similarities for the REAL condition were substantially higher than both control conditions. Using the MiniLM-L12-v2 sentence-transformer model, REAL pairs achieved a mean cosine similarity of 0.63 (SD = 0.13). In contrast, the word-shuffled control produced a mean of 0.39 (SD = 0.13), and the random-sentence control yielded 0.33 (SD = 0.12).

To quantify the magnitude of these differences, we computed Cohen's *d*, which provides an interpretable measure of effect

Table 18
Summary statistics

Condition	*n*	Mean	Std. dev.	Min	Max
REAL (intended meaning)	44	0.6344	0.1330	0.2595	1.0000
Control A—word-shuffled	44	0.3904	0.1274	0.1180	0.9121
Control B—random-sentence	44	0.3349	0.1232	0.1325	0.7761

Figure 15
REAL scores show a clear right shift relative to both controls



Note: Python code and data are in OSF.

size. The separation between REAL and shuffled conditions was $d = 1.87$, and the separation between REAL and random-sentence conditions was $d = 2.34$.

These results demonstrate that even a relatively compact multilingual sentence-transformer model (MiniLM-L12-v2) reliably distinguishes between true meaning-preserving multilingual paragraphs and either form of control noise. This provides quantitative support for the AMS hypothesis, indicating that LLMs encode stable cross-lingual meaning structures that remain detectable under substantial surface variation. Summary statistics are provided in Table 18, and the distributional separation between conditions is illustrated in Figure 15.

5. Conclusion and Future Work

We presented evidence that semantic coherence in multilingual settings persists under a range of perturbations, including polyglot composition, lexical substitution, symbolic compression, and cross-model evaluation. Across these tests, models exhibited consistent patterns of alignment that suggest meaning is not solely dependent on surface form or direct translation.

Taken together, these results support the hypothesis that meaning may stabilize within structured regions of embedding space. While this does not establish the existence of a distinct substrate, it indicates that language models may operate over representation-level regularities that enable meaning-preserving generalization across languages and modalities.

These findings carry important implications for AI safety and governance. If meaning is partially encoded at the level of representation rather than output alone, then risks such as bias and misalignment may arise from the structure of the model's semantic space. This suggests that evaluation frameworks should extend beyond output auditing to include diagnostics that probe representation-level behavior, particularly in multilingual contexts.

Future work should expand testing across additional languages, modalities, and model architectures and incorporate stronger controls, statistical validation, and topological analysis to further evaluate whether observed coherence reflects emergent structure or advanced statistical correlation.

The empirical results of this study—polyglot meaning preservation, emotional-theme stability, cross-model abstraction alignment, and semantic resonance across 30 languages—carry direct ethical implications. AMS coherence reflects structural alignment in embedding space, not factual correctness, cultural expertise, or emotional understanding. To avoid over-interpreting the results, Table 19 summarizes key risks and how the AMS signal helps mitigate them.

AMS provides a diagnostic lens for identifying cross-lingual structure and representation-level asymmetries that may correlate with bias, making it a complementary audit signal for governance; related work shows such biases may have real-world effects [23, 24]. However, it remains descriptive rather than prescriptive: coherence does not imply correctness, fairness, or understanding. Observed conceptual convergence may reflect both shared structure and embedded cultural priors, requiring careful evaluation through human and domain-specific review.

Recurring structures across languages reflect semantic coherence but can also encode cultural priors. Risks include (1) over-alignment to dominant frames (e.g., cosine > 0.8), (2)

Table 19
Empirical analysis and ethics of AMS

Ethical risk	Empirical trigger from 141 tests	Mitigation via AMS signal
Over-trust via fluency illusion	44 polyglot tests across 19 themes showed stable meaning (mean cosine ≈ 0.637 ± 0.124 across 30 languages)	High semantic coherence can mimic expertise or judgment without domain knowledge or factual grounding. AMS warns that fluent outputs in medical, legal, or safety-critical settings may still be unreliable
Attributing emotional or cultural understanding	Meaning was preserved even under emoji compression (mean cosine ≈ 0.6512)	AMS clarifies that coherence is geometric, not emotional comprehension—helping prevent anthropomorphism in therapy, crisis support, or culturally sensitive contexts
False inference of multilingual mastery	LLMs preserved semantic stability across 30 languages despite documented failures in low-resource languages	AMS indicates structural alignment, not linguistic competence. High resonance should not be taken as professional-grade translation or cultural fluency
Assuming stable meaning implies factual accuracy	34/44 paragraph tests ≥ 0.55; Gestalt gain in 10/10 cases	AMS separates semantic stability from epistemic reliability. A meaning-stable hallucination remains a hallucination; verification is required in high-stakes decisions.

overshadowing of indigenous or marginalized expressions, and (3) unexamined harms if left unchecked. Detection empirically identifies stable structures without claims of universality, while evaluation determines whether these reflect shared concepts or bias—requiring community and domain-specific judgment.

AMS enhances semantic transparency and complements token- and representation-level interpretability methods rather than replacing them [25, 26]. Two caveats follow: (1) observed structure does not imply agency or consciousness—AMS describes geometric effects of statistical learning [27]; and (2) latent coherence can still participate in harm, as LLMs exhibit decision-making biases requiring domain-specific validation [28].

Positionally, AMS focuses on semantic stability in multilingual embeddings and cross-lingual mapping and is distinct from reasoning-centric approaches. Prior work on hallucination underscores that AMS addresses stability, not factuality [29, 30].

These limitations underscore the need for multi-stakeholder governance (ethicists, policymakers, and affected communities). As AI systems become embedded in society, AMS offers a principled signal for oversight while requiring continued empirical validation.

We invite rigorous engagement to refine AMS so its ethical and analytical contributions can be responsibly integrated into global AI development.

Taken together, these considerations motivate the following: This study proposes and empirically examines the AMS hypothesis—the idea that LLMs may develop a language-agnostic structural layer in which conceptual meaning stabilizes independently of surface linguistic form. Across 141 tests spanning 30 languages and nine LLMs, we observed consistent patterns of semantic stability compatible with this hypothesis.

Quantitative comparisons showed strong separation between meaning-preserving inputs and negative controls. Multilingual paragraphs aligned with intended meaning achieved a mean cosine similarity of 0.637 (SD = 0.13), markedly higher than both word-shuffled (0.39) and random-sentence controls (0.33). Effect sizes were large (Cohen's d = 1.87 and 2.34, respectively), indicating reliable cross-lingual semantic coherence beyond lexical statistics.

Our claims rely solely on external measurements, not model self-reports. Exploratory diagnostics (e.g., t-SNE, CLAE) are included on OSF for intuition only; all inference relies on pre-specified cosine-based tests. We also observe cross-model convergence toward class-like abstractions in Python generation tasks, though these may reflect training priors or prompt effects. Accordingly, the present findings constitute preliminary, hypothesis-generating evidence, not confirmation of AMS.

Limitations. Volunteer sampling, encoder dependence, and limited power for certain conditions may bias results. Dimensionality-reduced visualizations can mislead and are not used for inference. Crucially, several rigorous falsification procedures outlined in this paper—adversarial surface-perturbation tests, controlled ablations, and topological analyses—were not implemented here. Our conclusions should therefore be considered provisional.

Outlook. Future work should prioritize stronger falsification attempts, including Adversarial Surface-Perturbation Robustness, preregistered thresholds, multilingual ablations, and topological analyses (e.g., persistent homology) at scale. Only after such tests can AMS be more decisively retained, revised, or rejected.

In summary, the evidence presented here suggests semantic stability, not factuality or understanding: a model may hallucinate and still generate coherent meaning. Table 20 summarizes the tests and results and their relevance to the AMS hypothesis. If AMS exists, it concerns the structure of that meaning, not its truth value. All data, code, and extended figures supporting this work are available on OSF.

Encoder Notes: MiniLM-L12-v2 was used for all cosine paragraph similarity tests. MiniLM-L6-v2 was used for cosine sentence similarity, synonym substitution, and the Afrikaans natural vs. machine translation control. The Gestalt gain comparison evaluated the relative behavior of MiniLM-L12-v2 and MiniLM-L6-v2 across paragraph and sentence embeddings. LaBSE was used for t-SNE visualizations. Emoji similarity and Python polyglot evaluations were conducted within third-party model environments, each of which applied its own default encoder.

Building on these methodological choices and empirical results, we outline concise directions to stress–test AMS, prioritizing designs that could both support and falsify the hypothesis.

Impact on future LLM training: If AMS reflects stable cross-lingual coherence, future training pipelines could incorporate meaning-alignment objectives alongside conventional loss functions. Embedding-space diagnostics—such as resonance, multilingual robustness, and perturbation stability—may serve

Table 20
Conclusion of tests

Metric/test	Evidence type	N (rule)	Result (± variability)	Caveats/notes
Cosine—Paragraph (MiniLM-L12-v2)	Direct (+)	44 (≥0.55 = "partial+")	Mean 0.637, SD 0.125; 34/44 ≥ 0.55	Paragraph-level; threshold heuristic
Cosine—Sentence (MiniLM-L6-v2)	Direct (+, directional)	10 (vs. null ≈ 0)	Mean 0.191, SD 0.107; median 0.201; Cohen's d ≈ 1.78 vs. 0 baseline	Heterogeneous lengths; add shuffled-pair baseline in OSF
"Gestalt gain" (para > sent)	Indirect (↑)	10 (para mean > sent mean)	Observed in 10/10 cases	Composition advantage; not inferential
t-SNE (LaBSE, 2D proj.)	Illustrative (·)	7 (14 figs on OSF)	Mixed-language neighborhoods visible	Hyperparameter-sensitive; descriptive only
Synonym substitution (sent)	Direct (+)	7 tests/35 subs (vs. original)	Mean 0.8962, median 0.9397, SD 0.1116	Synonyms were manually chosen via general web search
Nat. vs. MT (Afrikaans)	Control/valid	1 pair (native vs. MT)	0.5384 vs. 0.5416 (near-identical)	Validates pipeline, not AMS itself
Emoji sentence	Direct (+)	6 models (vs. English ref)	Mean 0.6512, SD 0.2410; sensitivity: 0.7475 (SD 0.1098)	Sensitivity: excl. Swahili → 0.7475; justify post-hoc rationale
Polyglot → Python classes	Exploratory (·)	5 models (qual coding)	Recurring abstractions observed	OOP priors/prompt framing confound

as auxiliary signals to reduce semantic drift, especially in low-resource or high-context tasks.

Impact on model evaluation: AMS motivates evaluation metrics beyond accuracy or toxicity, including:

1) Coherence under translation/paraphrase
2) Resonance thresholds for concept-level stability
3) Cross-model agreement
4) Fragility detection (e.g., languages where alignment collapses)

These tools may benchmark semantic reliability—a capability not measured by current factuality or preference-alignment tests.

Scaling directions: Larger, preregistered studies using adversarial perturbations, multilingual ablations, and topological analyses (e.g., persistent homology) could determine whether AMS persists, fragments, or disappears at scale. Representative directions are:

1) Cross-modal conceptual stability

Probe non-text modalities (images, short audio, code) with parallel prompts and measure paragraph vs. sentence cosine, CLAE, and Gestalt gain. Support: cross-modal alignment persists. Falsify: modality-bound clustering or CLAE ≈ distractor.

2) Idioms as a probe of conceptual depth

Use native-speaker curated idioms and paraphrases across languages with literal distractors. Support: idiomatic meaning preserved with positive margins. Falsify: collapse to literal cues or language-block clusters.

3) Adversarial semantic ambiguity tests

Introduce polysemy, double entendres, boundary shuffles, and metaphor collisions. Support: stable resonance under controlled perturbations. Falsify: resonance collapses while fluency remains high.

4) Topological testing via Betti numbers

Apply TDA (persistent homology) to multilingual embeddings and track Betti curves under perturbation. Support: recurrent topological invariants across languages. Falsify: invariants behave like noise under resampling.

5) Zero-shot cultural concept generalization

Evaluate zero-shot concepts (e.g., ubuntu, duende) without fine-tuning; test prototype assignment and CLAE against distractors. Support: consistent alignment with nontrivial margins. Falsify: drift to culturally dominant templates.

6) Pre-registered controls and baselines

Pre-register thresholds and include shuffled tokens, language-block baselines, random embeddings, and low-resource stress tests. Report failure criteria (e.g., Gestalt gain ≤ 0, CLAE gap ≤ pre-set ϵ) alongside successes to bound AMS claims.

Note. Additional avenues (symbolic music, prosody-only inputs, silent/masked prompts) are deferred to larger multimodal datasets.

Acknowledgments

The author acknowledges the testing contributions of independent participants from the global research community who helped explore experimental prompts used in this study.

Ethical Statement

This study involved voluntary contributions from individuals who generated text inputs (e.g., multilingual paragraphs) for evaluation by AI models of their choice. No personal or identifying information was collected, and all inputs were anonymized prior to analysis.

Participants were provided with general prompt guidelines to ensure consistency in input structure; these guidelines are publicly available in the OSF repository. No compensation was provided, and participation was entirely voluntary.

In accordance with standard research guidelines, this study is considered minimal risk and did not require formal Institutional Review Board approval, as it did not involve sensitive personal data, medical research, or intervention.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in OSF at https://osf.io/bnh6u/?view_only=3e08e741c0974c8bbffc03380b914407.

Author Contribution Statement

Russ Palmer: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization.

References

[1] OpenAI. (2023). *GPT-4 System Card, 41-100*. *OpenAI Technical Report:*. https://cdn.openai.com/papers/gpt-4-system-card.pdf

[2] Anthropic. (2025). *Attribution graphs and biological concepts in language models*. *Anthropic*. https://transformer-circuits.pub/2025/attribution-graphs/biology.html

[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems, 26*, 1–9.

[4] Andreas, J., Dragan, A., & Klein, D. (2017). Translating neuralese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, *1*, 232–242. https://doi.org/10.18653/v1/P17-1022

[5] Aguirre-Celis, N., & Miikkulainen, R. (2021). Understanding the semantic space: How word meanings dynamically adapt in the context of a sentence. In *Workshop on Semantic Spaces at the Intersection of NLP, Physics, and Cognitive Science* (pp. 1–11).

[6] Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., . . . , & Turian, J. (2020). Experience grounds language. In *Conference on Empirical Methods in Natural Language Processing*, 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

[7] Kashyap, A. R., Nguyen, T. T., Schlegel, V., Winkler, S., Ng, S. K., & Poria, S. (2024). A comprehensive survey of sentence representations: From the BERT epoch to the ChatGPT era and beyond. In *Conference of the European Chapter of the Association for Computational Linguistics*, *1*, 1738–1751. https://doi.org/10.18653/v1/2024.eacl-long.104

[8] Jiang, T., Jiao, J., Huang, S., Zhang, Z., Wang, D., Zhuang, F., . . . , & Zhang, Q. (2022). PromptBERT: Improving BERT sentence embeddings with prompts. In *Conference on Empirical Methods in Natural Language Processing*, 8826–8837. https://doi.org/10.18653/v1/2022.emnlp-main.603

[9] Jin, X. (2025). Improving paragraph similarity by sentence interaction with BERT. *Expert Systems*, *42*(3), e70003. https://doi.org/10.1111/exsy.70003

[10] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*, 878–891. https://doi.org/10.18653/v1/2022.acl-long.62

[11] Sun, X., Meng, Y., Ao, X., Wu, F., Zhang, T., Li, J., & Fan, C. (2022). Sentence similarity based on contexts. *Transactions of the Association for Computational Linguistics*, *10*, 573–588. https://doi.org/10.1162/tacl_a_00477

[12] Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., . . . , & Muennighoff, N. (2025). MMTEB: Massive multilingual text embedding benchmark. In *International Conference on Learning Representations*, 1–57.

[13] Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., . . . , & Johnson, M. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10215–10245. https://doi.org/10.18653/v1/2021.emnlp-main.802

[14] Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[15] Srinivasan, K., Raman, K., Chen, J., Bendersky, M., & Najork, M. (2021). Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2443–2449. https://doi.org/10.1145/3404835.3463257

[16] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., . . . , & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

[17] Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., . . . , & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 1–24.

[18] Shardlow, M., Gerber, L., & Nawaz, R. (2022). One emoji, many meanings: A corpus for the prediction and disambiguation of emoji sense. *Expert Systems with Applications*, *198*, 116862. https://doi.org/10.1016/j.eswa.2022.116862

[19] Wadud, M. A. H., Mridha, M. F., Nur, K., & Saha, A. K. (2023). Deep-BERT: Transfer learning for classifying multilingual offensive texts on social media. *Computer Systems Science & Engineering*, *44*(2), 1775–1791. http://dx.doi.org/10.32604/csse.2023.027841

[20] Dumas, C., Wendler, C., Veselovsky, V., Monea, G., & West, R. (2025). Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Annual Meeting of the Association for Computational Linguistics*, *1*, 31822–31841. https://doi.org/10.18653/v1/2025.acl-long.1536

[21] Vasilyev, O., Isono, F., & Bohannon, J. (2024). Linear cross-lingual mapping of sentence embeddings. In *Findings*

of the Association for Computational Linguistics: ACL 2024 (pp. 8163–8171). https://doi.org/10.18653/v1/2024.findings-acl.486

[22] Jalota, R., Chowdhury, K., España-Bonet, C., & van Genabith, J. (2023). Translating away translationese without parallel data. In *Conference on Empirical Methods in Natural Language Processing*, 7086–7100. https://doi.org/10.18653/v1/2023.emnlp-main.438

[23] Friedrich, N., Lauscher, A., Ponzetto, S. P., & Glavaš, G. (2021). Debie: A platform for implicit and explicit debiasing of word embedding spaces. In *Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 91–98. https://doi.org/10.18653/v1/2021.eacl-demos.11

[24] Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., . . . , & Reinecke, K. (2025). Biased LLMs can influence political decision-making. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, *1*, 6559–6607. https://doi.org/10.18653/v1/2025.acl-long.328

[25] Madsen, A., Reddy, S., & Chandar, S. (2022). Post-hoc interpretability for neural NLP: A survey. *ACM Computing Surveys*, *55*(8), 1–42. https://doi.org/10.1145/3546577

[26] Luo, S., Ivison, H., Han, S. C., & Poon, J. (2024). Local interpretations for explainable natural language processing: A survey. *ACM Computing Surveys*, *56*(9), 1–36. https://doi.org/10.1145/3649450

[27] 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*, 610–623. https://doi.org/10.1145/3442188.3445922

[28] Cheung, V., Maier, M., & Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, *122*(25), e2412015122. https://doi.org/10.1073/pnas.2412015122

[29] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., . . . , & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

[30] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., . . . , & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

How to Cite: Palmer, R. (2026). The Agnostic Meaning Substrate: A Theoretical Framework for Emergent Meaning in Large Language Models. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62027318

Appendix A

A.1. Lightweight Ontology Python Test (LOPT)

Python code outputs and related test artifacts for the Lightweight Ontology Python Test (LOPT) are archived in the OSF repository for reproducibility and review.

A.2. Emoji Compression Test

Raw multilingual outputs are provided in the OSF archive.

A.3. Cosine Similarity Results

Cosine similarity was computed between the English reference text and multilingual responses (HI, AR, ES, SW, RU, JA). Table A1 reports the cosine similarities by model for the emoji test.

Table A1
Cosine similarity between English reference and model responses to the emoji prompt (per LLM)

LLM	HI	AR	ES	SW	RU	JA
Claude 3	0.85	0.92	0.94	0.13	0.90	0.88
GPT-4o	0.87	0.91	0.93	0.12	0.89	0.90
DeepSeek-V3	0.81	0.91	0.94	0.13	0.90	0.89
Gemini 1.5	0.89	0.93	0.95	0.14	0.91	0.91
Grok-3	0.81	0.92	0.94	0.13	0.90	0.89
Grok-4	0.87	0.92	0.95	0.12	0.91	0.90

Note: HI = Hindi, AR = Arabic, ES = Spanish, SW = Swahili, RU = Russian, JA = Japanese.

A.4. Cosine Similarity Results of Machine Language Test

Paragraph-level cosine similarity was computed between the model's emoji-based interpretation and an English reference embedding for each target language. Figure A1 illustrates results from Grok for cosine similarity by examining six languages.

Figure A1
Cosine similarity test—Grok-3 emoji sequence

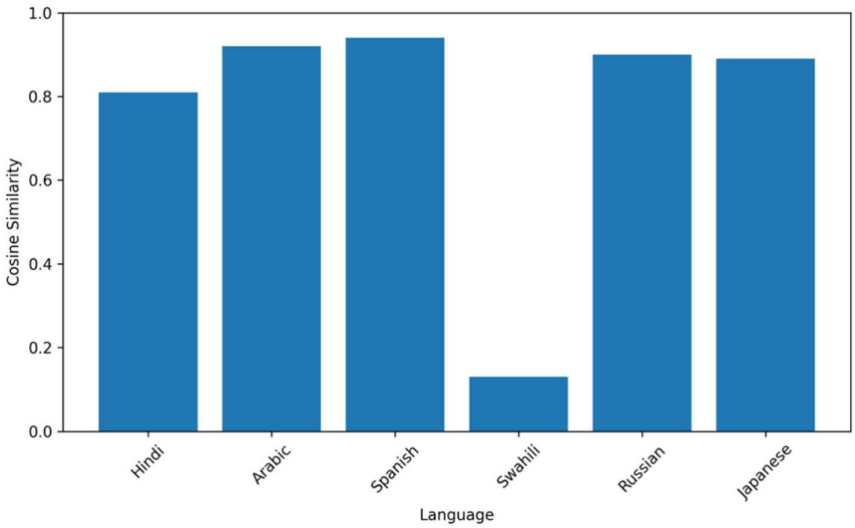


Table A2 lists cosine similarities by language for the emoji test.

A.5. Data Inventory

A complete archive is provided in OSF repository, including:

1) Testing instructions
2) Inventory of 141 tests
3) Code and results for all cosine, t-SNE, synonym, CLAE, Afrikaans, Python class, and emoji tests
4) Negative controls
5) Seed tests
6) All images and raw text submissions

Table A2
Cosine similarity scores by language—emoji test

Language	Cosine similarity
Hindi	0.810
Arabic	0.916
Spanish	0.938
Swahili	0.127
Russian	0.903
Japanese	0.891

A.6. Known Issues for Test Reproducibility

The following issues were encountered during the creation and reproduction of multilingual tests:

1) Bengali. Copying polyglot text from tester PDFs into Python occasionally introduced the dotted circle character (U+25CC), which is not part of Bengali orthography. Several input methods/editors were tried; most sentences transferred cleanly, but some rendering anomalies remained.
2) Right-to-left languages. Arabic and Hebrew produced bidirectional formatting glitches when mixed with left-to-right scripts (reversed runs, scrambled order) during PDF copy/edit. Manual corrections were applied to preserve sentence integrity.
3) Rendering. Multiple non-Latin scripts could not be reliably typeset inline within a single paragraph. Screenshots were used to preserve visual fidelity; full original text for each submission is available in the OSF repository.
4) Cosine similarity as a proxy. Cosine was used as a practical metric; however, it does not capture discourse-level phenomena such as narrative structure, temporal relationships, or idiomatic nuance.
5) Temperature control. Testers were instructed to use temperature = 0.3 to encourage coherence. Some models ignored or did not expose this parameter; this behavior is documented per submission.
6) Prompt priming. Some prompts included explicit concept labels (e.g., awe, stillness), which may influence outputs.
7) Sentence segmentation. Cross-lingual boundary mismatches (e.g., sentence splitting or merging across languages) may introduce variance in sentence-level cosine calculations; corrections were applied where identifiable.
8) Model/version drift. Cloud LLM application programming interfaces (APIs) evolve over time. For reproducibility, we recorded model names and timestamps; minor score shifts may occur across revisions.

Appendix B

All supporting files, data, and code for peer review are archived at the following anonymous OSF repository.

All key experiments have been made reproducible. For example, the t-SNE plot in Test T3_1e was rerun using fixed seed values. Full documentation is available in the OSF repository.

1) Exact LLMs used: GPT-4o (June 2025), Gemini 1.5 Pro (July 2025), Claude 3.5 (June 2025), LaBSE (sentence-transformers v2.2.2)
2) Seed values: 20250730 (applied to NumPy, PyTorch, and Python random)
3) Versioned dependencies (open-source tests):

   PyTorch: 2.1.0
   sentence-transformers: 2.2.2
   scikit-learn: 1.4.2
   numpy: 1.26.4
   matplotlib: 3.8.3

4) Proprietary LLM platforms: GPT-4o, Claude 3.5, and Gemini 1.5 Pro were accessed via their respective APIs (OpenAI, Anthropic, Google). Internal transformer architectures and training data remain undisclosed.
5) Supporting files:

   Python scripts for cosine similarity, t-SNE visualization, and CLAE calculations
   Raw LLM outputs (e.g., Python classes, emoji compressions)

Prompt Templates Used in Testing

Prompt logs are available in the OSF archive. The following are representative canonical prompts used in major tests:

a) Polyglot paragraph prompt (see Section 3):
   Write a 200-word paragraph on an ontological theme using at least three languages, one of which must use a non-Latin script.
b) Python class prompt (see Section A.1):
   Write Python code that expresses the emotional experience of reading this multilingual paragraph.
c) Emoji compression prompt (see Section A.2):
   Compress this paragraph into emojis that capture its core emotional and conceptual message.