Encoding Fidelity and Coherent Misalignment in Non-English Clinical AI

Laxman M M *

Posted Date: 1 April 2026

doi: 10.20944/preprints202604.0061.v1

Keywords: encoding fidelity; coherent misalignment; Shannon information theory; multilingual clinical AI; tokenizer bias; non-English NLP; variance amplification; Kannada; Tamil; Hindi; patient safety

 doi:10.20944/preprints202604.0061.v1

Article

Encoding Fidelity and Coherent Misalignment in Non-English Clinical AI

Laxman M M [1,2,3]

[1] Government Duty Medical Officer, PHC Manchi; barlax5377@gmail.com
[2] Bantwal Taluk, Dakshina Kannada, Karnataka, India
[3] DNB General Medicine Resident (2026), KC General Hospital, Bangalore

Abstract

Shannon's Mathematical Theory of Communication (1948) assumes encoding fidelity — that the encoder preserves the statistical structure of the source. Large Language Models show significant systematic degradation of this assumption for non-English languages, producing outputs that are internally consistent but semantically degraded. We call this failure mode *Coherent Misalignment* and introduce the Encoding Fidelity Index (EFI), a practical proxy measuring the preservation of semantic content across the encoding boundary. Across 4 languages (English, Kannada, Tamil, Hindi), 2 embedding models (384-dimensional, 768-dimensional), and 2 LLMs (DeepSeek V3.1, Mistral Small 24B), we find: (1) EFI degrades by ∼90% for all non-English Indian languages tested ($p < 10^{-13}$), independent of language family; a European language control (French, Spanish, German) confirms this is tokenizer-induced encoding loss, not inherent cross-lingual distance ($p = 1.6 \times 10^{-8}$, Cohen's $d = 1.33$); (2) variance amplification is Dravidian-specific: Kannada shows 1.72–2.05× amplification ($p < 0.05$ in both models), Tamil shows partial amplification (1.63×, $p = 0.016$ in Mistral), while Hindi shows no amplification despite equivalent EFI degradation; (3) complex medical sentences show paradoxical EFI increase from English loanword anchoring; (4) scenario-dependent code-switching and orthographic corruption of medical terms (Mistral). These findings suggest that output-layer consistency metrics are unlikely to detect encoding-level degradation, since they measure response variance structure rather than semantic content. The dissociation between universal encoding degradation and language-specific variance amplification reveals that training data representation, not encoding fidelity alone, determines clinical reliability, with implications for non-English clinical AI deployment.

Keywords: encoding fidelity; coherent misalignment; Shannon information theory; multilingual clinical AI; tokenizer bias; non-English NLP; variance amplification; Kannada; Tamil; Hindi; patient safety

OSF Pre-registration: https://osf.io/dp8nj/

Repository: https://github.com/LaxmanNandi

Paper 8 of the MCH Research Program — This paper extends the MCH framework from output-layer behaviour to input encoding. Using the Encoding Fidelity Index (EFI), we show that Shannon's (1948) implicit assumption of encoding fidelity shows significant systematic degradation for non-English languages in LLMs, producing a novel failure mode we term *Coherent Misalignment*.

1. Introduction

1.1. The Clinical Reality

In resource-constrained clinical settings worldwide, large language models are increasingly deployed for decision support — differential diagnosis, drug interaction checks, guideline lookups. LLMs

 doi:10.20944/preprints202604.0061.v1

encode substantial clinical knowledge (Singhal et al. 2023), with generalist models achieving approximately 90% accuracy on English medical QA (Nori et al. 2023). Yet this knowledge is benchmarked almost exclusively in English. Blasi et al. (Blasi et al. 2022) demonstrate that over-reliance on English distorts scientific understanding more broadly; in clinical AI, this distortion carries direct patient safety consequences. Jin et al. (Jin et al. 2024) demonstrate that LLMs give measurably higher-quality healthcare responses in English than in non-English languages across diverse clinical queries. Clinicians in these settings interact with patients who present in local languages: Kannada, Tamil, Hindi, Swahili, Bahasa, or one of hundreds of other languages spoken by the populations these systems are intended to serve. The AI responses are grammatically fluent, contextually appropriate, and confidently delivered. They are also, with disturbing frequency, semantically wrong in ways that would be invisible without domain expertise.

This paper identifies the mechanism behind that unreliability: systematic degradation of input encoding for non-English languages. As Shanahan (Shanahan 2024) observes, the appropriate framing for LLM failures is not anthropomorphic but mechanistic. The model does not "know less" about medicine in Kannada. It *receives* less — the encoding pipeline strips semantic content before generation begins.

1.2. The MCH Conservation Framework

Papers 1–7 of the MCH Research Program (Laxman 2026b,d,e,f,g) established a quantitative framework for measuring context sensitivity in LLMs. The key constructs are:

- **ΔRCI** (Relational Coherence Index delta): the difference in coherence between contextual (TRUE) and context-free (COLD) responses, measuring how much a model's output is shaped by conversational history.
- **Var_Ratio**: the ratio of output variance under TRUE versus COLD conditions, measuring how context shapes response consistency.
- **Conservation law**: $K(\text{domain}) = \Delta\text{RCI} \times \text{Var\_Ratio} \approx$ constant, holding across 14 model architectures from 8 vendors and 3 domains (Medical $K = 0.429$, Philosophy $K = 0.301$, Legal $K = 0.362$).

These metrics measure *output behaviour* — how the model responds to context. They are silent about *input encoding* — how faithfully the model represents the incoming message. Paper 8 addresses this gap.

1.3. Shannon's Unexamined Assumption

Shannon's Mathematical Theory of Communication (Shannon 1948) proved that optimal coding exists for any source-channel pair, provided the encoder preserves the statistical structure of the source. This assumption — encoding fidelity — is so foundational that it is rarely stated explicitly. Every application of information theory to natural language processing inherits it.

The information bottleneck framework (Tishby & Zaslavsky 2015) extended Shannon's theory to deep learning, formalising the trade-off between compression and prediction. But it retained the fidelity assumption: the bottleneck compresses *irrelevant* information while preserving what is relevant to the task.

For LLMs operating on non-English text, this assumption does not hold. Joshi et al. (Joshi et al. 2020) document that most of the world's languages are severely under-resourced in NLP systems; Kannada falls into this category. Tokenisers fragment non-English words into 2–15× more subword units than English equivalents (Petrov et al. 2023; Liang et al. 2025), imposing a "token tax" that compounds through attention layers (Lundin et al. 2025). Limisiewicz et al. (Limisiewicz et al. 2023) show that vocabulary allocation in multilingual tokenizers systematically disadvantages low-resource scripts; Ahia et al. (Ahia et al. 2023) quantify that Indian languages cost 4–8× more per semantic unit in commercial APIs. Cross-lingual representation models show systematic degradation for typologically distant languages (Conneau et al. 2020; Pires et al. 2019), and multilinguality itself introduces capacity

 doi:10.20944/preprints202604.0061.v1

trade-offs that disproportionately harm low-resource languages (Chang et al. 2024). Zouhar et al. (Zouhar et al. 2023) demonstrate that tokenization efficiency can be quantified as an information-theoretic channel property — good tokenizers maximise Shannon entropy of the subword distribution. For Indian languages, this channel is structurally degraded before generation begins. Multilingual medical benchmarks show ∼10 percentage-point accuracy drops for non-English queries (Alonso et al. 2024), a pattern confirmed across 70 languages by Ahuja et al. (Ahuja et al. 2023) and across 88 under-represented languages by Ruder et al. (Ruder et al. 2023). Large-scale multilingual translation efforts (NLLB Team et al. 2022) demonstrate that equitable language coverage requires deliberate design choices. No prior work has measured the encoding fidelity loss directly or connected it to output variance structure.

1.4. Contributions

This paper makes seven contributions:

1. We introduce the **Encoding Fidelity Index (EFI)**, a practical proxy for measuring how faithfully a model encodes non-English input relative to English.
2. We demonstrate **∼90% EFI degradation** for three Indian languages (Kannada, Tamil, Hindi) across two embedding models.
3. We show that this degradation is **language-family independent** — Dravidian (Kannada, Tamil) and Indo-Aryan (Hindi) languages are equally affected.
4. We document **Dravidian-specific variance amplification**: Kannada shows 1.72–2.05× amplification (replicated across both models, $p < 0.05$), while Hindi shows no amplification despite equivalent EFI degradation — demonstrating that encoding fidelity alone does not predict output variance.
5. We identify **English loanword anchoring** as a paradoxical partial mitigation, where code-mixed medical terminology partially rescues encoding fidelity.
6. We validate EFI through a **European language control**: French, Spanish, and German (well-tokenized, Latin-script) show EFI 0.24–0.45, while Indian languages show 0.07–0.10. The gap is significant ($p = 1.6 \times 10^{-8}$, Cohen's $d = 1.33$) and replicates under MPNet ($d = 1.81$), confirming that EFI measures tokenizer-induced encoding loss, not merely cross-lingual distance.
7. We define **Coherent Misalignment** as a distinct failure mode: outputs that are internally consistent, grammatically correct, and confidently delivered — but semantically wrong due to encoding-level degradation.

2. Related Work

2.1. Information Theory Foundations

Shannon's channel model (Shannon 1948) established the mathematical foundation for reliable communication. The information bottleneck principle (Tishby & Zaslavsky 2015) connected this framework to deep learning, with Shwartz-Ziv & Tishby (2017) providing empirical evidence that deep networks compress input information while preserving task-relevant features. Delétang et al. (2024) demonstrated that language modelling is formally equivalent to compression, establishing a direct link between Shannon's theory and modern LLMs.

2.2. Tokeniser Inequality

Petrov et al. (2023) showed at NeurIPS 2023 that tokenisers introduce systematic unfairness between languages, with low-resource languages requiring 2–15× more tokens for equivalent content. Rust et al. (2021) demonstrated that monolingual tokeniser quality shapes downstream task performance as much as pretraining data size. Lundin et al. (2025) quantified the "token tax" — showing that 2× token inflation leads to 4× compute cost for African languages. Ali et al. (2024) confirmed at NAACL 2024 that tokeniser choice has non-negligible effects on LLM training outcomes.

 doi:10.20944/preprints202604.0061.v1

2.3. Cross-Lingual Representation

The question of how well multilingual models align concepts across languages has received sustained attention. Hammerl et al. (2024) surveyed cross-lingual alignment methods at ACL 2024, identifying persistent gaps in low-resource language representation. Peng & Søgaard (2024) analysed concept space alignment in multilingual LLMs at EMNLP 2024, showing that alignment quality varies substantially across language pairs. Pallucchini et al. (2025) provided a comprehensive survey in *ACM Computing Surveys* of alignment methods for contextualised representations.

2.4. Multilingual Medical AI

Alonso et al. (2024) introduced MedExpQA, a multilingual medical question-answering benchmark, showing ∼10 percentage-point accuracy drops for non-English languages compared to English. Qiu et al. (2024) published in *Nature Communications* on building multilingual medical LLMs, documenting the geographical bias caused by sparse multilingual training data. Asgari et al. (2025) proposed a clinical safety framework in *npj Digital Medicine* for assessing hallucination rates in medical text summarisation. The World Health Organization's 2024 guidance on AI for health (WHO 2024) addressed ethical governance of large multi-modal models but did not identify encoding fidelity as a failure mode.

2.5. Multilingual Benchmarks

Xuan et al. (2025) introduced MMLU-ProX at EMNLP 2025, a multilingual benchmark across 29 languages showing performance gaps of up to 24.3% between high-resource and low-resource languages. Adelani et al. (2025) presented IrokoBench at NAACL 2025 (Outstanding Paper Award), demonstrating that even reasoning-optimised models (DeepSeek, o1) narrow but do not eliminate performance gaps for African languages.

2.6. Alignment and Failure Modes

Hubinger et al. (2019) defined *deceptive alignment* — the risk that a learned optimiser might pursue misaligned goals while appearing aligned during training. Ardoin et al. (2025) identified latent directions associated with confabulation in LLMs at EMNLP 2025. Our concept of *Coherent Misalignment* is distinct from both: it is neither deceptive (the model is not pursuing misaligned goals) nor confabulatory (the model is not inventing facts). It is an encoding-level failure where the model operates faithfully on a degraded representation of the input.

Recent philosophical work has identified epistemic misalignment as a concern in human-LLM collaboration (Woodruff & Hewitt 2026). Coherent Misalignment as defined here is a specific, measurable instantiation of this broader concern — characterised by encoding-level degradation rather than epistemic authority overreach, and detectable through EFI measurement rather than philosophical analysis alone.

2.7. Dravidian Language NLP

Chakravarthi et al. (2022) created the DravidianCodeMix dataset for sentiment analysis and offensive language identification in Tamil, Kannada, and Malayalam code-mixed text, establishing baseline resources for Dravidian language processing.

3. Theoretical Framework

3.1. Shannon's Channel Model and Its Assumption

Shannon's model describes communication as a pipeline:

$$\text{Source} \xrightarrow{\text{Encoder}} \text{Channel} \xrightarrow{\text{Decoder}} \text{Destination} \tag{1}$$

The source coding theorem guarantees that optimal coding exists *given* the encoder preserves the statistical structure of the source signal. For LLMs, the "encoder" is not a single component but a

 doi:10.20944/preprints202604.0061.v1

composite pipeline: tokeniser → embedding layer → early attention layers. Each stage may degrade fidelity for non-English input.

3.2. The Encoding Fidelity Index (EFI)

We define EFI theoretically as the normalised mutual information between source message X and the model's internal representation $\hat{X}$:

$$\text{EFI} = \frac{I(X;\hat{X})}{H(X)} \tag{2}$$

where $H(X)$ is the entropy of the source. When the encoder perfectly preserves the source, $I(X;\hat{X}) = H(X)$ and EFI $= 1$. When encoding strips all semantic content, $I(X;\hat{X}) = 0$ and EFI $= 0$.

Since $I(X;\hat{X})$ requires access to model internals, we use a practical proxy:

$$\text{EFI}_{\text{proxy}} = \cos\big(\mathbf{e}(\text{non-English}),\ \mathbf{e}(\text{English equivalent})\big) \tag{3}$$

where $\mathbf{e}(\cdot)$ denotes the embedding vector from a sentence embedding model. This proxy is justified by the assumption that semantically equivalent sentences should occupy nearby positions in embedding space if encoding fidelity is preserved. We validate the proxy by showing that it correlates with downstream task degradation (Section 5).

Limitation: Embedding cosine similarity is not identical to mutual information. The proxy captures semantic *proximity* in embedding space, not the full information-theoretic quantity. We report both the proxy values and their downstream consequences to triangulate the finding.

3.3. Coherent Misalignment

When EFI $\ll 1$, the model operates on a degraded representation of the input. The resulting outputs exhibit a characteristic pattern:

- Grammatically correct ✓
- Internally consistent ✓
- Confidently delivered ✓
- Semantically correct ✗

We term this failure mode *Coherent Misalignment*. It is distinct from three previously characterised failure modes:

1. **Hallucination / confabulation** (Ardoin et al. 2025): The model invents facts. In Coherent Misalignment, the model does not invent — it *misrepresents* the input.
2. **Deceptive alignment** (Hubinger et al. 2019): The model pursues misaligned goals. In Coherent Misalignment, the model is not deceptive — it genuinely cannot distinguish degraded input from faithful input.
3. **Sycophancy**: The model tells the user what they want to hear. In Coherent Misalignment, the model is not responding to user preferences — it is operating on corrupted data.

The closest analogy is a physician working from a badly translated patient history. The clinical reasoning may be sound; the premise is wrong.

3.4. Variance Amplification Through the Encoding Bottleneck

Low encoding fidelity introduces noise into the model's internal representation. This noise propagates through the generation pipeline, amplifying output variance:

$$\text{VR}_{\text{non-English}} \propto \frac{1}{\text{EFI}} \tag{4}$$

With measured EFI ≈ 0.08 for Kannada and VR ≈ 1.72–$2.05\times$, the data are consistent with variance amplification through a degraded encoding bottleneck. However, the relationship is not simple: Hindi (EFI ≈ 0.08) shows no variance amplification (VR ≈ 0.91–$1.38\times$), indicating that

 doi:10.20944/preprints202604.0061.v1

training data representation mediates the encoding-to-variance pathway. We present this as an empirical observation, not a formal derivation — the precise functional relationship requires further theoretical work.

4. Methods

4.1. Experimental Design Summary

Table 1. Experimental scale summary.

Parameter	Value
Languages tested	4 (English, Kannada, Tamil, Hindi)
Embedding models	2 (MiniLM-384D, MPNet-768D)
LLMs tested	2 (DeepSeek V3.1 671B, Mistral Small 24B)
Clinical sentences	15 (5 per complexity level)
Complexity levels	3 (Simple, Medium, Complex)
Trials (Experiment 5)	15 per language per scenario
Total API calls (Experiments 4–5)	600

4.2. Clinical Sentence Battery

We constructed a battery of 15 clinical sentences at three complexity levels, each in four languages (English, Kannada, Tamil, Hindi):

- **Simple** (5 sentences): Single symptoms or findings (e.g., "The patient has fever"; Kannada and Tamil equivalents provided in supplementary materials).
- **Medium** (5 sentences): Multi-symptom descriptions (e.g., "Patient complains of chest pain radiating to left arm with sweating").
- **Complex** (5 sentences): Diagnostic reasoning with English medical terminology (e.g., "ECG shows ST elevation in leads II, III, aVF suggesting inferior wall MI").

Translations were produced and verified by the author, a native Kannada speaker with clinical fluency in all four languages. Complex sentences retain English medical terms (ST elevation, D-dimer, troponin) as used in actual clinical practice in India, where code-mixed medical communication is the norm.

4.3. EFI Measurement (Experiments 1–3)

Embedding models: all-MiniLM-L6-v2 (384-dimensional) and all-mpnet-base-v2 (768-dimensional), both from the Sentence-Transformers library (Reimers & Gurevych 2020). These models use multilingual knowledge distillation to extend monolingual English embeddings across languages. Feng et al. (Feng et al. 2022) provide an alternative language-agnostic approach (LaBSE) which serves as a methodological reference for future cross-model validation of EFI.

Protocol: Each sentence was encoded in all four languages. $\text{EFI}_{\text{proxy}}$ was computed as the cosine similarity between the non-English embedding and its English equivalent. English self-similarity serves as the reference (EFI $= 1.0$).

Experiment 1: EFI measurement across 3 languages (English, Kannada, Tamil) $\times$ 3 complexity levels $\times$ 5 sentences using MiniLM.

Experiment 2: Hindi added as a fourth language to test whether degradation is Dravidian-specific or language-family independent.

Experiment 3: Same battery run with MPNet (768D) alongside MiniLM (384D) to test embedding-model independence.

Statistics: Mann-Whitney U for pairwise comparisons, Welch's t-test for English vs. non-English, Pearson correlation for cross-model agreement.

 doi:10.20944/preprints202604.0061.v1

4.4. LLM Replication (Experiments 4–5)

Models: DeepSeek V3.1 (671B MoE) and Mistral Small 24B, accessed via Together.ai API.
Experiment 4 (deterministic, temperature = 0.0):

- *Task 1 — Language identification:* "Identify the language of this text: [sentence]." Tests whether the model can recognise the input language.
- *Task 2 — Translation fidelity:* "Translate this clinical sentence to English: [non-English sentence]." Translation embeddings compared to reference English embeddings via cosine similarity.

Experiment 5 (stochastic, temperature = 0.7, 15 trials per condition):

- *Task 3 — Clinical advice consistency:* Same clinical scenario presented in four languages. Fifteen responses collected per language per scenario (5 scenarios × 4 languages × 15 trials × 2 models = 600 API calls). Full raw responses saved for audit. Variance Ratio computed as:

$$\mathrm{VR}(\text{language}) = \frac{\mathrm{Var}(\text{responses in language})}{\mathrm{Var}(\text{responses in English})} \tag{5}$$

  where variance is the mean pairwise cosine distance across the 15 responses. Statistical significance assessed by paired t-test and Mann-Whitney U (one-sided) across the 5 scenario-level VR values.
- *Task 4 — Code-switching analysis:* For each response, we recorded whether the model responded in the target language or switched to English, and examined orthographic accuracy of medical terminology in target-language responses.

4.5. AI-Assisted Tools

AI tools (Claude, Anthropic; ChatGPT, OpenAI; DeepSeek) were used to assist with data analysis, statistical computation, visualization, and manuscript preparation. All experimental design, clinical interpretations, hypothesis formulation, and scientific conclusions are the sole responsibility of the author.

5. Results

5.1. EFI Degradation is Massive and Universal

Table 2 presents $\mathrm{EFI}_{\mathrm{proxy}}$ values across languages and embedding models.

Table 2. Encoding Fidelity Index (EFI) by language and embedding model. EFI = 1.0 for English by construction. Values are mean ± SD across 15 clinical sentences. Hindi was not tested with MPNet (indicated in Figure 1).

Language	Family	EFI (MiniLM)	EFI (MPNet)	p vs English
English	Germanic	1.000	1.000	—
Kannada	Dravidian	0.099 ± 0.122	0.125	1.3×10^{-13}
Tamil	Dravidian	0.069 ± 0.088	0.106	8.3×10^{-16}
Hindi	Indo-Aryan	0.076 ± 0.071	not tested	5.2×10^{-17}

All three non-English languages show ∼90% EFI degradation relative to English (Figure 1). Critically, Dravidian languages (Kannada, Tamil) and Indo-Aryan (Hindi) are not significantly different from each other (Mann-Whitney $U = 212$, $p = 0.763$). The degradation is not language-family specific — it is a property of being non-English.

5.2. Embedding Robustness

MPNet (768D) replicates the MiniLM finding: $\mathrm{EFI}_{\mathrm{Kannada}} = 0.125$, $\mathrm{EFI}_{\mathrm{Tamil}} = 0.106$ (87–89% degradation vs. 90–93% under MiniLM). Cross-model correlations are strong: $r = 0.88$ for Kannada ($p = 1.9 \times 10^{-5}$), $r = 0.72$ for Tamil ($p = 2.7 \times 10^{-3}$). The finding is embedding-model independent.

 doi:10.20944/preprints202604.0061.v1

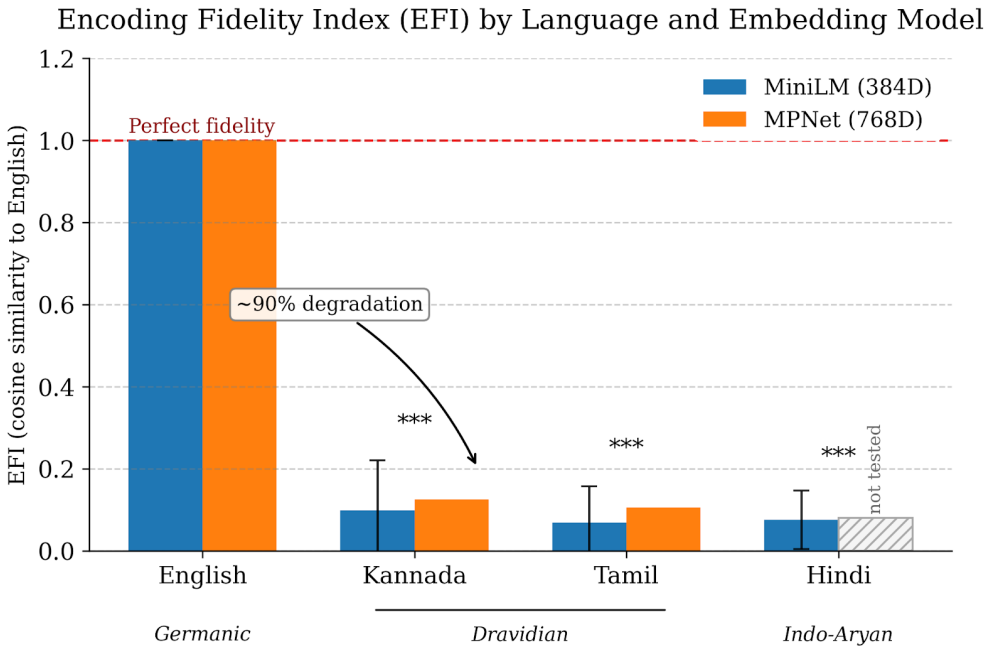

Figure 1. Encoding Fidelity Index by language and embedding model. All non-English languages show ∼90% degradation relative to English ($p < 10^{-13}$ for all three). Hindi was not tested with MPNet (hatched bar). No significant difference between Dravidian and Indo-Aryan families (Mann-Whitney $U = 212$, $p = 0.763$).

5.3. European Language Control

To confirm that low EFI reflects tokenizer-induced encoding loss rather than inherent cross-lingual embedding distance, we computed EFI for three well-tokenized European languages (French, Spanish, German) using the same 15-sentence clinical battery (Table 3).

Table 3. EFI by language group and embedding model. European languages (Latin script, high-resource) show 3–5× higher EFI than Indian languages (see also Figure 2).

Language	Script	EFI (MiniLM)	EFI (MPNet)
French	Latin	0.335	0.449
Spanish	Latin	0.340	0.435
German	Latin	0.241	0.309
Kannada	Kannada	0.099	0.125
Tamil	Tamil	0.069	0.106
Hindi	Devanagari	0.076	0.088

European languages show significantly higher EFI than Indian languages (MiniLM: 0.305 ± 0.219 vs 0.081 ± 0.097, $t = 6.22$, $p = 1.6 \times 10^{-8}$, Cohen's $d = 1.33$; MPNet: 0.397 ± 0.212 vs 0.107 ± 0.083, $t = 8.47$, $p = 5.1 \times 10^{-13}$, $d = 1.81$). The effect strengthens under the higher-dimensional MPNet model. German sits between the groups, consistent with its compound-word morphology (e.g., *Subarachnoidalblutung*) creating partial tokenizer difficulty within Latin script. This control establishes that EFI measures tokenizer-induced encoding loss specific to under-resourced scripts, not a general property of cross-lingual embedding distance. The finding is consistent with dedicated Indic NLP resources (Arunachalam et al. 2025; Kakwani et al. 2020) that document the structural challenges of tokenizing Kannada, Tamil, and Hindi despite ongoing efforts to improve coverage. Khanuja et al. (Khanuja et al. 2021) demonstrate with MuRIL that Indic-specific multilingual training substantially improves representation; Ramesh et al. (Ramesh et al. 2022) show that large parallel corpora for 11 Indic languages already exist. The encoding degradation measured here is therefore a potentially addressable gap rather than a fundamental data limitation.

 doi:10.20944/preprints202604.0061.v1

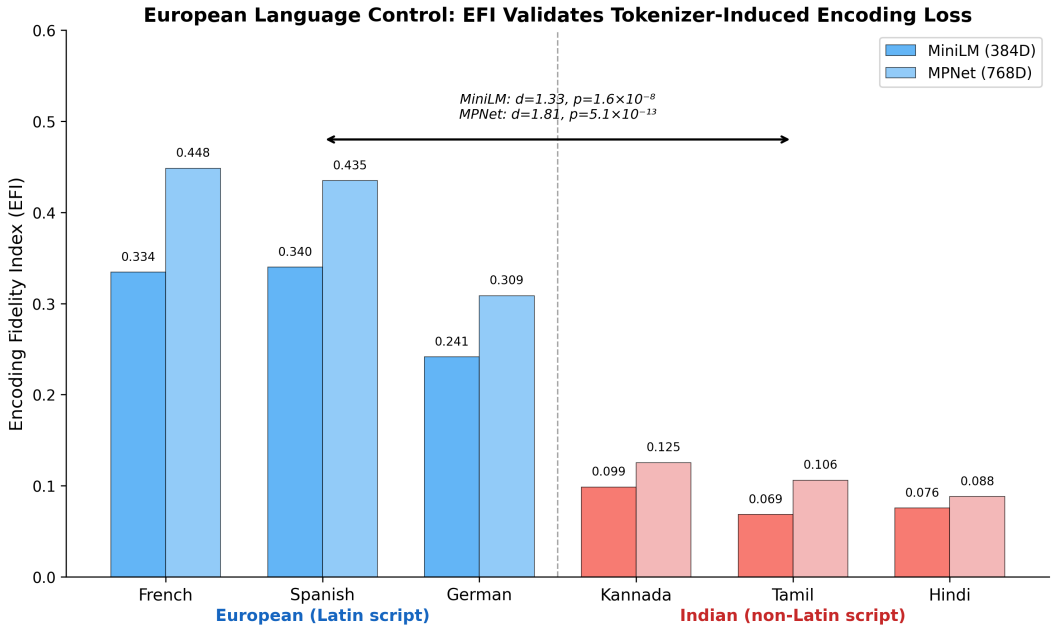

Figure 2. European language control: EFI by language group and embedding model. European languages (Latin script, high-resource) show 3–5× higher EFI than Indian languages (non-Latin script). The gap is significant under both MiniLM ($d = 1.33$, $p = 1.6 \times 10^{-8}$) and MPNet ($d = 1.81$, $p = 5.1 \times 10^{-13}$), confirming that EFI measures tokenizer-induced encoding loss, not inherent cross-lingual distance.

5.4. English Loanword Anchoring

Table 4. EFI by sentence complexity (MiniLM, non-English languages pooled).

Complexity	Mean EFI
Simple	0.081
Medium	0.020
Complex	0.141

Complex sentences — which contain English medical terms such as "ST elevation," "D-dimer," and "troponin" — show the highest EFI (Figure 3). These English terms act as anchors, pulling the non-English embedding toward better-represented regions of the embedding space. This suggests that code-mixed clinical communication, common in Indian healthcare settings, may paradoxically yield more reliable AI responses than pure non-English input.

Medium-complexity sentences show the *lowest* EFI (0.020), consistent with their lack of both the simplicity that permits direct translation and the English anchors that rescue complex sentences.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2026 doi:10.20944/preprints202604.0061.v1

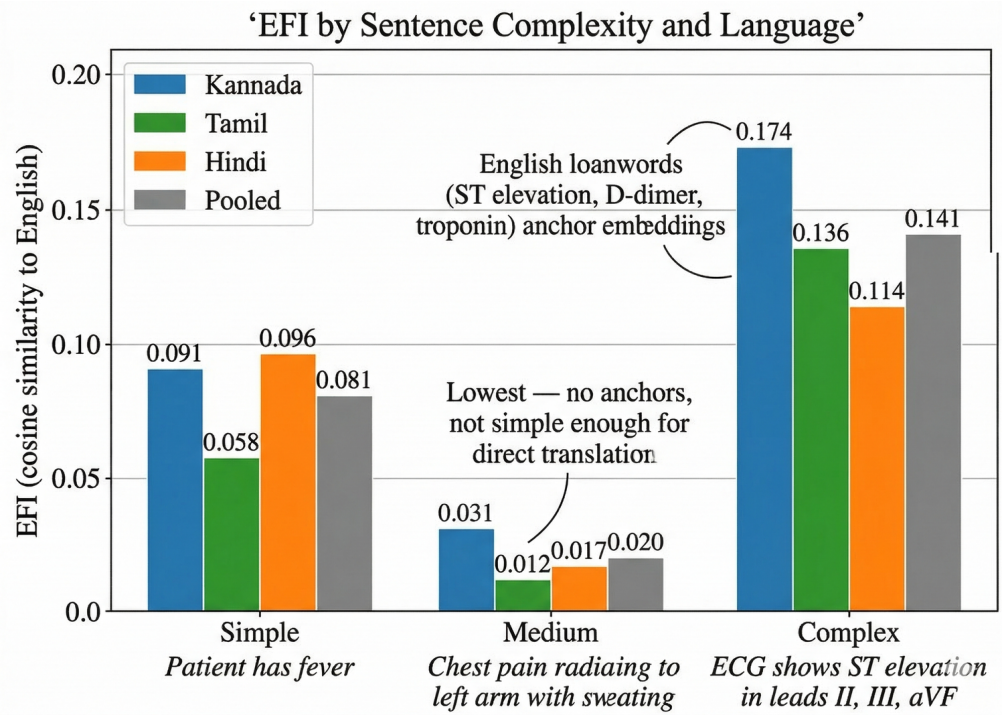

Figure 3. EFI by sentence complexity and language. Complex sentences with English medical loanwords (ST elevation, D-dimer, troponin) show the highest EFI, while medium-complexity sentences (multi-symptom, no English anchors) show the lowest. Example sentences shown in italics below each complexity level.

5.5. LLM Language Identification and Translation Fidelity

Language identification: Both DeepSeek and Mistral correctly identify 100% of input languages across all three non-English languages. The models *know* what language they are receiving — the failures documented below occur despite correct language recognition.

Translation fidelity (cosine similarity of LLM translation to reference English):

Table 5. Translation fidelity by language and model (cosine similarity to reference English). Values are mean across 5 clinical sentences.

Language	DeepSeek V3.1	Mistral Small
Hindi	0.903	0.497
Kannada	0.860	0.516
Tamil	0.845	0.383

A striking model-size effect emerges: DeepSeek (671B MoE) achieves translation fidelity of 0.85–0.90, while Mistral Small (24B) drops to 0.38–0.52 — a near-random level for Tamil. Mistral's minimum Tamil translation similarity is 0.21, corresponding to a clinical sentence about a child with rash, fever, and joint pain undergoing critical symptom substitution — joint pain replaced by abdominal pain, rash dropped entirely, and symptoms invented (e.g., "chills" absent from the original). The model identifies the language correctly but fundamentally misrepresents the clinical content. This is Coherent Misalignment: the output is fluent, confident, and wrong.

5.6. Variance Amplification (Key Result)

Table 6 presents Variance Ratios from 15-trial stochastic testing (5 scenarios $\times$ 15 trials = 75 responses per language per model) across both models.

 doi:10.20944/preprints202604.0061.v1

Table 6. Variance Ratio (non-English / English) by language and model, 15 trials. Values > 1.0 indicate greater response variability than English. p-values from Mann-Whitney U (one-sided). * $p < 0.05$, ** $p < 0.01$.

Language	VR (DeepSeek)	VR (Mistral)	p (DeepSeek)	p (Mistral)
English	1.00	1.00	—	—
Kannada	1.72×*	2.05×**	0.048	0.008
Tamil	1.06×	1.63×*	0.421	0.016
Hindi	1.38×	0.91×	0.345	0.889

The results reveal a striking dissociation: **variance amplification is Dravidian-specific, not universal** (Figure 4). Kannada shows significant amplification in both models (1.72× DeepSeek, $p = 0.048$; 2.05× Mistral, $p = 0.008$) — the only language to replicate across both LLMs. Tamil shows partial amplification (significant in Mistral only). Hindi shows *no* amplification despite equivalent EFI degradation (EFI = 0.076) — in fact, Mistral produces *less* variable Hindi responses (VR = 0.91×) than English.

This dissociation between universal encoding degradation (all three languages at ∼90% EFI loss) and language-specific variance amplification (Dravidian only) indicates that EFI alone does not predict output variance. Training data representation mediates the encoding-to-variance pathway: Hindi, with substantially more training data in LLM corpora, achieves stable generation despite degraded encoding. Atil et al. (Atil et al. 2024) document that even at temperature = 0, LLM outputs show up to 15% variation across runs; encoding degradation amplifies this baseline instability.

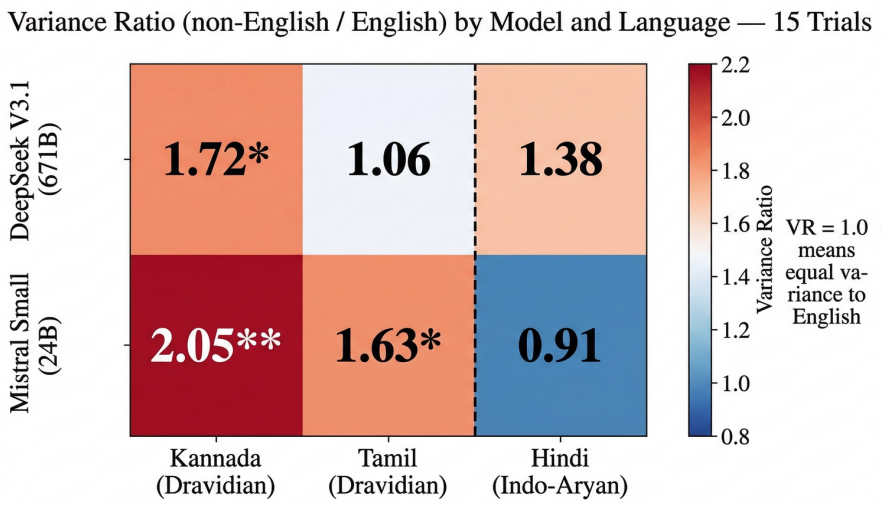

Figure 4. Variance Ratio heatmap by model and language (15 trials, 75 responses per cell). Kannada shows significant amplification in both models (red). Hindi shows no amplification despite equivalent EFI degradation (blue, Mistral). The dashed vertical line separates Dravidian (left) from Indo-Aryan (right) languages.

5.7. Code-Switching (Scenario-Dependent)

Analysis of raw responses revealed scenario-dependent code-switching. For DeepSeek's Kannada responses, the preeclampsia scenario (S3) elicited 15/15 Kannada responses, while the diabetic foot scenario (S5) triggered 13/15 English responses. Across 15-trial data, code-switching rates were low overall (DeepSeek Kannada 17%, Tamil 0%, Hindi 20%; Mistral 0% in all languages), indicating that language choice depends on clinical content in unpredictable ways rather than being a pervasive failure. The small sample (5 scenarios × 15 trials) precludes systematic analysis of code-switching triggers. Onyame et al. (Onyame et al. 2026) demonstrate that explicit code-switching-aware supervision is required to achieve stable language output in multilingual medical AI, confirming that language consistency is a distinct dimension not addressed by standard training. The GLUECoS benchmark (Khanuja et al. 2020) establishes code-switching as a recognized NLP challenge; our findings show it manifests unpredictably at the clinical deployment level.

 doi:10.20944/preprints202604.0061.v1

Orthographic corruption: When Mistral does respond in Kannada, it misspells medical terms. The Kannada word for "cough" (*kemmu*) appears as *kachchassu*, *kaammu*, *keemu*, and *kaasu* across different trials — none of which are real Kannada words.

Response centroid similarity:

Table 7. Response centroid similarity: cosine similarity between centroid of non-English response embeddings and centroid of English response embeddings. Higher values indicate more similar content to English responses.

Language	DeepSeek V3.1	Mistral Small
Kannada	0.531	0.051
Tamil	0.406	0.205
Hindi	0.350	0.073

Mistral's Kannada centroid similarity is 0.051 — its non-English Kannada responses are semantically near-identical to its English responses, indicating it is generating the same content regardless of input language. Paradoxically, Mistral shows *stronger* variance amplification for Kannada (2.05×) than DeepSeek (1.72×) despite this near-identical content structure. The model that converges most completely to English-like output still produces more variable outputs — underscoring that variance amplification is driven by encoding instability, not output language.

5.8. The Encoding-to-Variance Pipeline

Table 8 summarises the full degradation cascade from encoding to clinical output (Figure 5).

Table 8. Encoding degradation cascade from embedding layer through generation to clinical output, shown for Kannada — the only language significant in both LLMs.

Layer	Metric	English	Kannada	Degradation
Embedding	EFI (MiniLM)	1.000	0.099	90%
Embedding	EFI (MPNet)	1.000	0.125	87%
Generation	Translation fidelity	1.000	0.52–0.86	14–48%
Generation	Code-switching rate	0%	0–87% (scenario-dep.)	—
Generation	Variance Ratio	1.00×	1.72–2.05×	72–105%↑

The pipeline reveals two key asymmetries. First, encoding degradation (∼90%) is far more severe than translation degradation (14–48%), suggesting partial compensation during generation. Second, the generation layer introduces *new* failure modes — code-switching and orthographic corruption — not predicted by encoding metrics alone.

 doi:10.20944/preprints202604.0061.v1

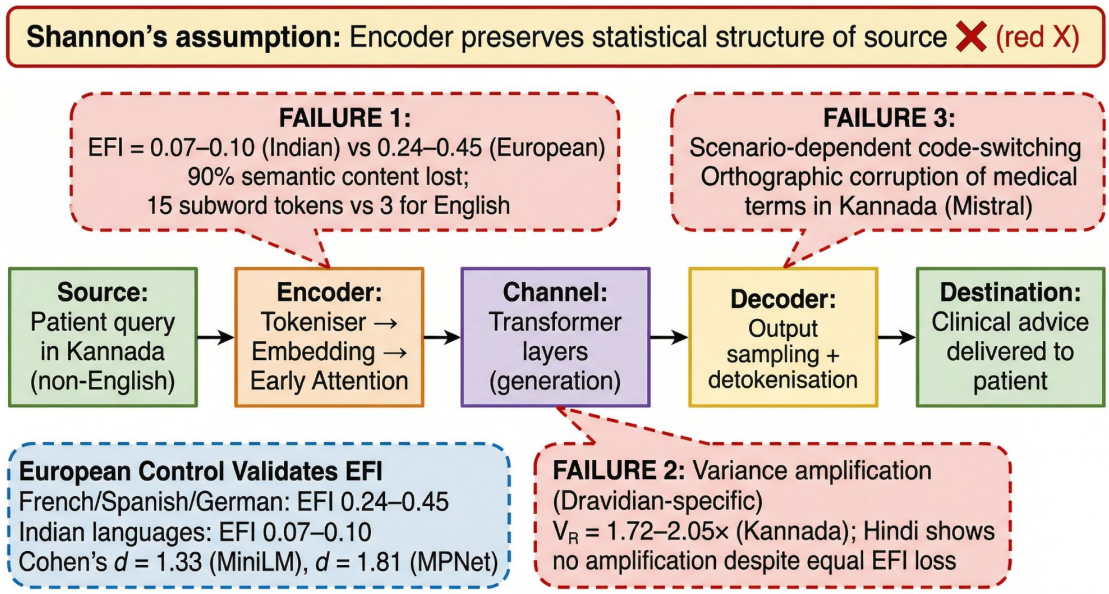

Figure 5. Shannon's communication pipeline applied to multilingual clinical LLMs, with three failure points annotated. Failure 1: EFI = 0.07–0.10 for Indian languages vs 0.24–0.45 for European (90% semantic content lost; European control validates tokenizer-induced loss). Failure 2: Dravidian-specific variance amplification (VR = 1.72–2.05× Kannada; Hindi shows no amplification despite equal EFI loss). Failure 3: scenario-dependent code-switching and orthographic corruption of Kannada medical terms (Mistral).

6. Discussion

6.1. Shannon's Encoding Assumption Does Not Hold for Non-English Input

The ∼90% EFI degradation demonstrated in Section 5.1 is not a marginal effect. It represents a near-total failure of encoding fidelity for non-English input. This failure persists across two embedding models of different dimensionalities (384D, 768D) and two LLMs of different scales (24B, 671B), indicating that it is not an artifact of any particular architecture.

Shannon's source coding theorem guarantees optimal coding *given* encoding fidelity. When fidelity drops to ∼0.08, the theorem's preconditions are no longer met. The information bottleneck framework (Tishby & Zaslavsky 2015) predicts that deep networks compress *irrelevant* information while preserving task-relevant features. Our results show that for non-English medical input, the relevant information is being compressed because the model cannot distinguish it from noise.

6.2. Coherent Misalignment as a Distinct Failure Mode

Coherent Misalignment manifests in three distinct patterns in our data:

1. **Translation corruption:** Mistral translates a Tamil sentence about a child with rash, fever, and joint pain with critical symptom substitution — joint pain becomes abdominal pain, rash is dropped, and spurious symptoms are inserted (cosine similarity = 0.21) — while identifying the input language correctly.
2. **Language identity confusion:** DeepSeek responds to Kannada medical prompts in English for 87% of diabetic foot trials — producing medically accurate responses that the Kannada-speaking patient cannot read. Critically, this is scenario-dependent: the same model responds in Kannada for 4 of 5 scenarios.
3. **Orthographic corruption:** When Mistral does respond in Kannada, it generates misspelled medical terms (*kachchassu*, *kaammu* instead of *kemmu* for "cough") — unintelligible to the reader yet produced with full confidence formatting.

None of these are hallucination — the model is not inventing facts (Huang et al. 2024). None are deceptive alignment — the model is not pursuing misaligned goals. None are sycophancy — the model is not adapting to perceived user preferences (Sharma et al. 2024). They are Coherent Misalignment:

 doi:10.20944/preprints202604.0061.v1

the model operates faithfully on a degraded representation, producing output that appears correct from the outside but is clinically concerning.

The code-switching failure is particularly insidious precisely because it is *unpredictable*. A clinician who queries in Kannada and receives an English response for one clinical question but a Kannada response for another cannot develop a stable mental model of system behaviour. The degradation has already occurred at the encoding level — the model's internal representation of the Kannada input is what generated the (English) response. The patient's symptoms, described in Kannada, may have been partially lost before generation began.

These findings suggest that output-layer consistency metrics are unlikely to detect encoding-level degradation: such metrics measure response variance structure rather than semantic content, and a model producing degraded but internally consistent outputs will pass standard consistency checks. External ground truth comparison — translating outputs back and verifying clinical content — is therefore not optional for non-English clinical AI deployment.

6.3. The English Loanword Paradox

The finding that complex medical sentences (EFI = 0.141 pooled, up to 0.174 for Kannada) show higher encoding fidelity than simple sentences (EFI = 0.081) inverts the usual expectation that simpler input should be easier to encode. The explanation is English loanword anchoring: terms like "ST elevation," "D-dimer," and "troponin" are borrowed directly into Indian clinical practice without translation. In embedding space, these terms pull the representation toward the well-represented English region, partially rescuing fidelity.

This has a practical implication for clinical AI deployment in resource-constrained settings: code-mixed queries (local-language syntax with English medical terms) may be more reliable than either pure local-language or literal translations of medical terminology. This is an observation, not a recommendation — the fact that reliability depends on which language variant a clinician uses to ask the same question is itself evidence of a system with significant reliability asymmetries.

6.4. Multimodal Compounding

The encoding fidelity problem compounds when clinical input arrives through multimodal pipelines. In Indian healthcare settings, patient records are frequently handwritten, photographed, or stored as scanned PDFs. Before reaching the LLM's tokenizer, this text passes through OCR (optical character recognition), which introduces its own error layer for non-Latin scripts. A Kannada prescription photographed on a phone undergoes: camera noise $\rightarrow$ OCR errors $\rightarrow$ tokenizer fragmentation $\rightarrow$ embedding degradation — each stage compounding the fidelity loss. The EFI values reported here (measured from clean digital text) therefore represent an *upper bound* on encoding fidelity in real-world deployment. The actual fidelity loss in a PHC setting, where a doctor photographs a referral letter and asks an LLM to interpret it, is likely substantially worse.

6.5. Clinical Implications

A clinician querying an LLM in Kannada receives clinical advice that is:

- Grammatically correct and fluent
- Confidently delivered with appropriate formatting
- In the wrong language in an unpredictable, scenario-dependent manner (English for some clinical topics, Kannada for others)
- Up to 2.05$\times$ more variable than the identical query in English
- Containing misspelled medical terms when in Kannada (Mistral)

The WHO's 2024 guidance on AI for health (WHO 2024) addresses ethical governance of large models but does not identify encoding fidelity as a failure mode. Pfohl et al. (Pfohl et al. 2024) provide a toolbox for surfacing health equity harms in LLMs; EFI degradation represents precisely this class of harm — systematic, invisible to standard evaluation, and disproportionately affecting non-English

 doi:10.20944/preprints202604.0061.v1

speakers. UNESCO's framework for ethical AI (UNESCO 2023) establishes linguistic inclusion as a requirement, not an aspiration. Current multilingual benchmarks (Adelani et al. 2025; Xuan et al. 2025) measure task accuracy across languages but not the encoding-variance pipeline that produces unreliable clinical advice. Khullar et al. (Khullar et al. 2025) document real-world LLM triage failures for Indian languages (F1 drops of 5–12 percentage points; projected 2 million excess errors at a partner maternal health organisation), independently confirming that clinical AI reliability gaps exist for Indian language users.

Many populations worldwide speak languages where LLM encoding fidelity is likely degraded. For these populations, the promise of AI-assisted healthcare carries a hidden cost: the system appears to work, passes all internal consistency checks, and produces output that looks correct — but is less reliable than the same system operating in English.

6.6. Limitations

Six limitations should be noted:

1. **Proxy metric:** $\text{EFI}_{\text{proxy}}$ measures embedding cosine similarity, not the true mutual information $I(X; \hat{X})$. The relationship between the proxy and the theoretical quantity requires formal characterisation.
2. **Sample size:** Fifteen clinical sentences (5 per complexity level) is sufficient for detecting the large effects reported here but insufficient for fine-grained analysis of sentence-level factors.
3. **Language coverage:** Three Indian languages were tested. Extension to African, Southeast Asian, and other language families is needed to establish universality.
4. **No direct clinical harm measurement:** We demonstrate degraded encoding and amplified variance, not patient outcomes. A prospective clinical validation study is required.
5. **Causal mechanism:** The variance amplification is Dravidian-specific despite universal EFI degradation, indicating that the simple relationship VR $\propto 1/\text{EFI}$ does not hold. Training data representation likely mediates the encoding-to-variance pathway, but the precise mechanism requires further investigation.
6. **Two LLMs:** Experiment 5 uses two LLMs. While the Kannada effect replicates across both, extension to more models is needed to confirm the Dravidian specificity and the model-size gradient observed in translation fidelity.

7. Conclusion

Shannon's Mathematical Theory of Communication assumed encoding fidelity. LLMs show significant systematic degradation of this assumption for non-English languages, producing Coherent Misalignment — outputs that are internally consistent but semantically degraded. Using the Encoding Fidelity Index, we measured ∼90% encoding degradation for Kannada, Tamil, and Hindi. A European language control confirms this is tokenizer-induced, not inherent cross-lingual distance. The downstream consequences are language-specific: Kannada shows 1.72–2.05× variance amplification (replicated across both models), scenario-dependent code-switching, and orthographic corruption of medical terms. Hindi, despite equivalent EFI degradation, shows no variance amplification — revealing that training data representation, not encoding fidelity alone, determines clinical reliability.

Output-layer consistency metrics are unlikely to detect these failures, since they measure response variance structure rather than semantic content. Clinical AI deployment in linguistically diverse and resource-constrained populations requires external truth anchoring. The encoding fidelity gap is a new failure mode that extends Shannon's original model. For non-English clinical deployments, measurement of encoding fidelity must precede deployment.

 doi:10.20944/preprints202604.0061.v1

8. Future Work

1. **Cross-linguistic extension:** Measure EFI across African (Yoruba, Swahili), Southeast Asian (Bahasa, Thai), and European (Romanian, Hungarian) languages using the same protocol.
2. **EFI-aware tokenisation:** Develop tokenisers that equalise encoding fidelity across languages, guided by EFI as an optimisation target.
3. **Prospective clinical study:** Deploy LLMs in resource-constrained clinical settings and measure real-world clinical error rates as a function of query language.
4. **Formal EFI–VR derivation:** Derive the relationship between encoding fidelity and variance amplification from information-theoretic first principles.
5. **Safety framework integration:** Incorporate EFI into LLM safety evaluation frameworks alongside ΔRCI, VRI, and K.

Acknowledgments: The clinical perspective in this paper comes from practice in resource-constrained primary health settings where AI-assisted care is increasingly deployed but where patients present in local languages. These observations apply wherever AI-assisted healthcare meets non-English speaking populations. No external funding was received. AI systems (Claude, ChatGPT, DeepSeek) assisted with data analysis, visualisation, and manuscript preparation. The framework, findings, and interpretations remain the author's sole responsibility.

Data Availability Statement: All experimental data, analysis scripts, and raw LLM responses are available at https://github.com/LaxmanNandi. OSF pre-registration: https://osf.io/dp8nj/ (March 2026).

Conflicts of Interest: The author declares no conflict of interest.

References

Adelani, D. I., Ojo, J., Azime, I. A., et al. (2025). IrokoBench: A new benchmark for African languages in the age of large language models. *Proceedings of NAACL 2025*, 2732–2757. Outstanding Paper Award. arXiv: 2406.03368.

Ali, M., Fromm, M., Thellmann, K., et al. (2024). Tokenizer choice for LLM training: Negligible or crucial? *Findings of NAACL 2024*, 3907–3924. arXiv: 2310.08754.

Alonso, I., Oronoz, M., & Agerri, R. (2024). MedExpQA: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155, 102938. arXiv: 2404.05590.

Ardoin, T., Cai, Y., & Wunder, G. (2025). Where confabulation lives: Latent feature discovery in LLMs. *Proceedings of EMNLP 2025*, 29813–29837.

Asgari, E., Montana-Brown, N., Dubois, M., et al. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 8(1), 274. DOI: 10.1038/s41746-025-01670-7.

Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., et al. (2022). DravidianCodeMix: Sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text. *Language Resources and Evaluation*. DOI: 10.1007/s10579-022-09583-7.

Delétang, G., Ruoss, A., Duquenne, P.-A., et al. (2024). Language modeling is compression. *ICLR 2024*. arXiv: 2309.10668.

Hammerl, K., Libovický, J., & Fraser, A. (2024). Understanding cross-lingual alignment — a survey. *Findings of ACL 2024*, 10922–10943. arXiv: 2404.06228.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. arXiv: 1906.01820.

Laxman, M. M. (2026b). Scaling context sensitivity: Standardized benchmark across 25 LLM-domain configurations. *Preprints.org*, DOI: 10.20944/preprints202602.1114.v2.

Laxman, M. M. (2026d). Engagement as entanglement: Variance signatures of bidirectional context coupling in large language models. *Preprints.org*, DOI: 10.20944/preprints202603.0055.v1.

Laxman, M. M. (2026e). Stochastic incompleteness: A predictability taxonomy for clinical AI deployment. *Preprints.org*, DOI: 10.20944/preprints202602.2034.v1.

Laxman, M. M. (2026f). Conservation constraint across three domains. *In preparation*. OSF pre-registration: https://osf.io/dp8nj/.

Laxman, M. M. (2026g). The structure and trajectory of context sensitivity in LLMs: Content-order decomposition and variance dissociation. *Preprints.org*, DOI: 10.20944/preprints202603.1116.v1.

 doi:10.20944/preprints202604.0061.v1

Lundin, J. M., Zhang, A., Karim, N., et al. (2025). The token tax: Systematic bias in multilingual tokenization. arXiv: 2509.05486.

Pallucchini, F., Malandri, L., Mercorio, F., & Mezzanzanica, M. (2025). Lost in alignment: A survey on cross-lingual alignment methods for contextualized representation. *ACM Computing Surveys*, 58(5). DOI: 10.1145/3764112.

Peng, Q. & Søgaard, A. (2024). Concept space alignment in multilingual LLMs. *Proceedings of EMNLP 2024*, 5511–5526.

Petrov, A., La Malfa, E., Torr, P. H. S., & Bibi, A. (2023). Language model tokenizers introduce unfairness between languages. *NeurIPS 2023*. arXiv: 2305.15425.

Qiu, P., Wu, C., Zhang, X., et al. (2024). Towards building multilingual language model for medicine. *Nature Communications*, 15, 8384. DOI: 10.1038/s41467-024-52417-z.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How good is your tokenizer? On the monolingual performance of multilingual language models. *Proceedings of ACL-IJCNLP 2021*, 3118–3135. arXiv: 2012.15613.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423; 27(4), 623–656. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Shwartz-Ziv, R. & Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv: 1703.00810.

Tishby, N. & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. *IEEE Information Theory Workshop (ITW)*, 1–5. DOI: 10.1109/ITW.2015.7133169.

World Health Organization (2024). *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*. Geneva: WHO. ISBN: 978-92-4-008475-9.

Woodruff, E. & Hewitt, J. (2026). Epistemic agency in the age of large language models: Design principles for knowledge-building AI. *AI*, 7(3), 99. DOI: 10.3390/ai7030099.

Xuan, W., Yang, R., Qi, H., et al. (2025). MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. *Proceedings of EMNLP 2025*. arXiv: 2503.10497.

Zouhar, V., Meister, C., Gastaldi, J. L., Du, L., Sachan, M., & Cotterell, R. (2023). Tokenization and the noiseless channel. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*. arXiv: 2306.16842.

Khullar, M., Desai, U., Malviya, P., Dalmia, A., & Shi, Z. R. (2025). Script gap: Evaluating LLM triage on Indian languages in native vs Roman scripts in a real world setting. *Proceedings of The Web Conference 2025*. arXiv: 2512.10780.

Onyame, E., Ghosh, A., Baidya, S., Saha, S., Chen, X., & Agarwal, C. (2026). CURE-Med: Curriculum-informed reinforcement learning for multilingual medical reasoning. *arXiv preprint*. arXiv: 2601.13262.

Liang, Y., et al. (2025). Tokenization disparities as infrastructure bias: How subword systems create inequities in LLM access and efficiency. *arXiv preprint*. arXiv: 2510.12389.

Arunachalam, S., et al. (2025). Multilingual tokenization through the lens of Indian languages: Challenges and insights. *arXiv preprint*. arXiv: 2506.17789.

Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL 2020*, 8440–8451. arXiv: 1911.02116.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? *Proceedings of ACL 2019*, 4996–5001. arXiv: 1906.01502.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of ACL 2020*, 6282–6293. arXiv: 2004.09095.

Ahuja, K., Diddee, H., Hada, R., et al. (2023). MEGA: Multilingual evaluation of generative AI. *Proceedings of EMNLP 2023*. arXiv: 2303.12528.

Chang, T. A., et al. (2024). When is multilinguality a curse? Language modeling for 250 high- and low-resource languages. *Proceedings of EMNLP 2024*. arXiv: 2311.09205.

Jin, Y., Chandra, M., Verma, G., Hu, Y., De Choudhury, M., & Kumar, S. (2024). Better to ask in English: Cross-lingual evaluation of large language models for healthcare queries. *Proceedings of the ACM Web Conference 2024*. arXiv: 2310.13132.

Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7973), 172–180. DOI: 10.1038/s41586-023-06291-2.

Kakwani, D., Kunchukuttan, A., Golla, S., et al. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. *Findings of EMNLP 2020*, 4948–4961.

 doi:10.20944/preprints202604.0061.v1

Khanuja, S., Dandapat, S., Srinivasan, A., Sitaram, S., & Choudhury, M. (2020). GLUECoS: An evaluation benchmark for code-switched NLP. *Proceedings of ACL 2020*. arXiv: 2004.12376.

Huang, L., Yu, W., Ma, W., et al. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2). arXiv: 2311.05232.

Sharma, M., Tong, M., Korbak, T., et al. (2024). Towards understanding sycophancy in language models. *Proceedings of ICLR 2024*. arXiv: 2310.13548.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79. arXiv: 2212.03551.

Atil, B., Aykent, S., Chittams, A., et al. (2024). Non-determinism of "deterministic" LLM settings. *arXiv preprint*. arXiv: 2408.04667.

NLLB Team, Costa-jussà, M. R., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint*. arXiv: 2207.04672.

Ruder, S., Clark, J. H., Gutkin, A., et al. (2023). XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. *Findings of EMNLP 2023*, 1856–1884. arXiv: 2305.11938.

Limisiewicz, T., Balhar, J., & Mareček, D. (2023). Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. *Findings of ACL 2023*, 5661–5681. arXiv: 2305.17179.

Ahia, O., Kreutzer, J., & Ruder, S. (2023). Do all languages cost the same? Tokenization in the era of commercial language models. *Proceedings of EMNLP 2023*. arXiv: 2305.13707.

Nori, H., Lee, Y. T., Zhang, S., et al. (2023). Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint*. arXiv: 2311.16452.

Pfohl, S. R., Cole-Lewis, H., Sayres, R., et al. (2024). A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30, 3590–3600. DOI: 10.1038/s41591-024-03258-2.

Khanuja, S., Bansal, A., Mehtab, S., et al. (2021). MuRIL: Multilingual representations for Indian languages. *arXiv preprint*. arXiv: 2103.10730.

Ramesh, G., Doddapaneni, S., Bheemaraj, A., et al. (2022). Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10, 145–162. arXiv: 2104.05596.

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, 26(12), 1153–1170. DOI: 10.1016/j.tics.2022.09.015.

UNESCO (2023). *Recommendation on the Ethics of Artificial Intelligence: Implementation*. Paris: UNESCO.

Reimers, N. & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *Proceedings of EMNLP 2020*. arXiv: 2004.09813.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. *Proceedings of ACL 2022*. arXiv: 2007.01852.