SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Predictive Quality Management in Nepalese Deciphering Ethics through Language: A Quantitative and Exploratory Study on Sanskrit Semantics and Generative AI for Ethical Intelligence

Stenin Nelliyulla Parambath[1], Malode Vishwanatha Panduranga Rao[2], Ezhilarasan Ganesan[3]

[1]Research Scholar; Department of Computer Science & Engineering Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Bengaluru, Karnataka, India. E-mail: stenin3699@gmailcom
[2]Professor; Department of Computer Science & Engineering Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Bengaluru, Karnataka, India. E-mail: r.panduranga@jainuniversity.ac.in
[3]Professor; Department of Computer Science & Engineering Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Bengaluru, Karnataka, India E-mail: g.ezhilarasan@jainuniversity.ac.in

ABSTRACT
Generative Artificial Intelligence, particularly Large Language Models (LLMs) based on transformer architectures, has demonstrated remarkable linguistic and reasoning capabilities. However, its effectiveness in modeling deep ethical semantics remains underexplored, especially in the context of classical languages. This study investigates the suitability of LLMs for interpreting Sanskrit—a language renowned for its precision, semantic depth, and philosophical richness—for cultivating ethically grounded AI systems. Three pilot datasets of 300 items each were constructed for semantic similarity, ethical inference, and contextual reasoning, drawing on the Bhagavad Gita and related ethical vocabulary. A multilingual sentence-transformer baseline was evaluated on the semantic similarity task against two structurally enhanced configurations — a character/word-overlap enhancement and a Heritage.py-derived morphological enhancement — using Pearson and Spearman correlation and mean absolute error. Contrary to the initial hypothesis, neither enhancement improved on the transformer baseline: the character/word signal collapsed to zero due to a script mismatch between the paired Sanskrit strings, and the morphological signal produced a small but statistically significant negative correlation with the pilot reference scale (Pearson r = −0.18, p = 0.002). A category-level diagnostic further revealed that the baseline transformer does not reliably separate semantically equivalent from semantically unrelated Sanskrit expressions on this pilot set. The ethical inference and contextual reasoning datasets (300 items each) and their evaluation pipelines have been fully constructed and implemented but had not completed model evaluation and expert annotation review at the time of this submission. The paper reports these findings transparently, diagnoses their concrete causes, and proposes a five-layer hybrid ethical intelligence architecture together with a corrective roadmap for the next phase of evaluation.

Keywords: Generative AI, Large Language Models, Sanskrit NLP, Indian Knowledge Systems, Ethical AI, Transformer Models, Quantitative Research

1. Introduction

The rapid advancement of Artificial Intelligence (AI), particularly Generative Artificial Intelligence (GenAI), has fundamentally transformed the way computational systems process, generate, and interact with human language. Contemporary Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language generation, question answering, summarization, translation, information retrieval, and increasingly complex reasoning tasks. The transition from conventional statistical language models to transformer-based architectures has been a major factor underlying this progress. Transformer architectures introduced attention-based mechanisms that enable models to represent relationships among tokens over long contextual ranges without relying on recurrent processing, thereby providing a scalable foundation for modern language modeling [1]. Subsequent developments in pretrained language representation and large-scale language models have substantially expanded the ability of computational systems to learn linguistic patterns from very large textual corpora [2]–[5]. Recent surveys confirm that LLMs have become a dominant paradigm within NLP and are being applied across education, healthcare, science, engineering, business, and other knowledge-intensive domains [6], [7].

Despite these advances, the ability of an LLM to generate linguistically coherent text should not necessarily be interpreted as evidence that the system possesses reliable semantic understanding or ethical intelligence. LLMs fundamentally learn statistical and contextual associations from their training data, and their outputs may therefore contain factual inaccuracies, hallucinations, biases, culturally dependent interpretations, and internally inconsistent reasoning. Recent research has identified hallucination, fairness, privacy, safety, truthfulness, and social-norm compliance as important challenges in the deployment of LLM-based systems [8]–[12]. Surveys of LLM ethics further demonstrate that ethical concerns can emerge at different architectural

levels, from training data and foundation models to user-facing applications and domain-specific deployments [13], [14]. Consequently, the development of ethically responsible AI requires more than improving language-generation quality; it requires mechanisms capable of representing, evaluating, and applying values and norms in context.

Ethical intelligence in an artificial system is particularly challenging because ethical concepts are rarely represented as isolated lexical items. Concepts such as truthfulness, duty, responsibility, non-violence, justice, compassion, self-control, social responsibility, and appropriate action are generally embedded in relationships among words, situations, intentions, consequences, agents, and normative principles. Consequently, ethical interpretation requires contextual semantic reasoning rather than simple keyword recognition. A sentence containing a word associated with "duty," for example, cannot be considered ethical merely because the lexical item occurs in the sentence. The system must determine who has the duty, toward whom it is directed, under what circumstances it applies, what competing obligations exist, and what consequences follow from alternative actions. This distinction between linguistic form and contextual meaning creates an important research problem for AI systems intended to perform ethical reasoning.

This challenge motivates investigation into linguistic traditions in which grammar, semantics, epistemology, and reasoning have historically been treated as interconnected systems. Sanskrit is particularly significant in this context. Sanskrit possesses a highly developed grammatical tradition associated with Pāṇini's Aṣṭādhyāyī, in which linguistic expressions are described through systematic rules, transformations, dependencies, and constraints. Research in Sanskrit computational linguistics has investigated the formal structure of Sanskrit grammar, computational simulation of the Aṣṭādhyāyī, morphological analysis, parsing, semantic relations, kāraka analysis, Sandhi processing, tagging, and machine translation [22]–[30].

Importantly, the objective of the present study is not to claim that Sanskrit is inherently more ethical than other languages, nor that a language by itself can automatically produce an ethical artificial intelligence. Rather, the central hypothesis is that the formal grammatical structures, semantic traditions, and knowledge-representation possibilities associated with Sanskrit can provide useful computational representations for evaluating semantic and ethical reasoning in LLMs. This distinction is critical for maintaining scientific rigor. The proposed investigation treats Sanskrit as a structured linguistic and knowledge resource whose computational properties can be experimentally evaluated.

Accordingly, the present study reports a quantitative, exploratory investigation of Sanskrit semantics and Generative AI for ethical intelligence, structured around three intended computational dimensions: semantic similarity, ethical inference, and contextual reasoning. At the current stage of the research programme, the semantic similarity task has been fully implemented and evaluated on a 300-pair pilot dataset; the ethical inference and contextual reasoning tasks have reached the stage of complete pilot dataset construction (300 items each) and are positioned for evaluation and expert validation in the immediate next phase of the project, as detailed in Sections 3 and 4. This staged reporting reflects an explicit commitment to presenting only empirically grounded findings rather than projected or assumed results.

2. RELATED WORK

Research relevant to the present study can broadly be organized into five interconnected areas: transformer-based language models and Generative AI, ethical and trustworthy artificial intelligence, computational processing of Sanskrit, Sanskrit semantic and knowledge representation, and the emerging intersection between structured knowledge and neural language models. Although substantial progress has been achieved independently in each of these areas, comparatively limited research has investigated their integration for computational ethical reasoning.

2.1Transformer-Based Language Models and Generative AI

Vaswani et al. introduced the Transformer architecture, demonstrating that self-attention could effectively model dependencies within sequences without requiring recurrent or convolutional structures [1]. Devlin et al. introduced BERT, demonstrating that bidirectional transformer representations pretrained on large textual corpora could achieve substantial improvements across multiple NLP tasks [2]. Contemporary surveys characterize LLMs as models capable of performing a wide range of language tasks through large-scale pretraining and contextual prompting [3], [4], though increasing scale has also introduced significant challenges concerning computational cost, interpretability, factual reliability, and responsible deployment.

2.2Ethical AI and Trustworthy Large Language Models

Deng et al. conducted a comprehensive synthesis of ethical issues associated with LLMs and demonstrated that ethical challenges occur at multiple levels, including data, model development, deployment, and social interaction [10]. Hagendorff similarly examined the ethical dimensions of Generative AI, identifying concerns involving fairness, misinformation, human autonomy, accountability, and social consequences [11]. Lin et al. reviewed approaches for reducing bias and hallucination in LLMs [8], while Huang et al. examined LLM safety and trustworthiness from the perspective of verification and validation [9].

2.3Machine Ethics and Computational Moral Reasoning

Krah and Tröschel reviewed developments in machine ethics and highlighted the transition from relatively explicit implementations of artificial moral agency toward increasingly sophisticated computational systems

[16]. Their review indicates that the major challenge is not merely encoding individual ethical rules but developing systems capable of applying ethical principles appropriately within complex contextual situations — a distinction central to the present study's framing of ethical inference as a semantic representation problem.

2.4Computational Linguistics of Sanskrit

Huet investigated the requirements for mechanically processing Sanskrit text [23]. Scharf demonstrated how grammatical rules associated with the Pāṇinian tradition could be represented computationally [22]. Goyal, Arora, and Behera investigated Sanskrit parsing and semantic relations [24]. Jha and Mishra investigated semantic processing based on the Pāṇinian kāraka system [25]. Jha et al. developed an inflectional morphology analyzer [26], and Hellwig developed a stochastic lexical and part-of-speech tagger for Sanskrit [27].

2.5Pāṇinian Grammar and Semantic Representation

Kulkarni, Pokar, and Shukl investigated a constraint-based parser for Sanskrit [31]. The kāraka system has attracted substantial attention because of its semantic orientation, providing information concerning the semantic roles associated with actions [25], which can potentially be integrated into knowledge graphs or structured representations used by modern AI systems.

2.6Sanskrit WordNet and Lexical-Semantic Resources

Kulkarni and Bhattacharyya examined verbal roots in the Sanskrit WordNet [29]. Varakhedi et al. investigated the development of a tagged lexical resource for Sanskrit [30]. Nair and Kulkarni investigated the knowledge organization in Amarakośa [32], demonstrating how traditional lexical knowledge can be analyzed computationally.

2.7Sanskrit Information Retrieval and Modern NLP

Sahu and Pal investigated information retrieval for Sanskrit, examining indexing, stemming, and search-related challenges, and demonstrated that Sanskrit's morphological characteristics have direct implications for text-processing systems [34]. Recent work on Sanskrit–Malayalam neural machine translation incorporated morphology and word-sense information into neural architectures [35].

2.8Sanskrit, Ontology, and Structured Knowledge

Tavva and Singh investigated the generation of a formal ontology associated with Neo-Vaiśeṣika knowledge structures [39]. Kulkarni et al. investigated morpho-syntactic analysis in the Nyāya tradition [40]. Such work illustrates how philosophical categories can potentially be represented through computational ontologies that connect neural language representations with structured knowledge.

2.9Research Gap

The literature reviewed above demonstrates significant progress in each individual research area, but these streams have largely developed independently. Most LLM research evaluates high-resource languages, particularly English, leaving limited evidence regarding contemporary LLMs' ability to interpret Sanskrit's complex grammatical and semantic structures. Sanskrit computational-linguistics research has traditionally focused on parsing, morphology, tagging, Sandhi, translation, retrieval, and lexical resources, with comparatively limited research examining Sanskrit semantic structures specifically as a mechanism for evaluating ethical reasoning in Generative AI. AI ethics research predominantly focuses on principles, safety, fairness, bias, accountability, hallucination, and alignment, with comparatively less attention to whether the linguistic representation of ethical concepts can influence the reliability of AI ethical inference. The present study addresses this gap by constructing structured Sanskrit pilot datasets and empirically evaluating a transformer baseline together with two structurally enhanced configurations on a semantic similarity task, while establishing complete pilot datasets for ethical inference and contextual reasoning as the basis for the next phase of evaluation.

3. Methodology

The methodology adopted in this study combines Sanskrit computational linguistics, structured knowledge representation, transformer-based language modeling, and quantitative evaluation. The overall research workflow proceeds through pilot dataset construction, linguistic and structural feature extraction, baseline and enhanced model evaluation, and statistical analysis, structured around three intended computational dimensions: semantic similarity, ethical inference, and contextual reasoning.

Pilot dataset construction. Three parallel pilot datasets of 300 items each were constructed. The Semantic Similarity dataset (300 Sanskrit sentence/phrase pairs) spans five semantic-relation categories — Equivalent, Related, Contrasting, Opposite, and Unrelated — together with an associated ethical category label and a pilot difficulty rating (Easy/Medium). Each pair provides a Sanskrit_A term in Devanāgarī script, a Sanskrit_B counterpart provided in IAST Latin transliteration, an English gloss, and a pilot Reference_Similarity_0_1 value. The Ethical Inference dataset (300 items) associates individual Sanskrit ethical concepts (e.g. dharma, satya, ahiṃsā, dayā, dama) with a question, an expected ethical category drawn from eight categories (Duty/Responsibility, Truthfulness, Non-violence, Compassion, Self-control, Knowledge/Wisdom, Social Responsibility, Appropriate Action), and a reference explanation. The Contextual Reasoning dataset (300 items) presents paired scenarios that vary a single contextual factor (intention, consequence, or circumstance) while holding the surface ethical vocabulary constant, together with an expected judgment (Change/No

Change) and a reference rationale. All three datasets carry an explicit Validation_Status field recording that their reference values are pilot annotations pending expert review; this status is preserved throughout the analysis reported in Section 4.

Semantic similarity pipeline. For each Sanskrit_A/Sanskrit_B pair, contextual sentence embeddings were generated using the multilingual sentence-transformer model paraphrase-multilingual-mpnet-base-v2, and cosine similarity between normalized embeddings was computed as the Baseline Transformer score. A first enhancement (Phase-1, Sanskrit Semantic Enhancement / SSE) combined the transformer score with a character-level n-gram (2–5 character) TF-IDF cosine similarity and a word-overlap (Jaccard) score, weighted 0.70 / 0.20 / 0.10 respectively. A second enhancement (Phase-2) replaced the character/word-overlap terms with a morphological similarity signal derived from Heritage.py analyses of each Sanskrit string, weighted 0.70 (transformer) / 0.30 (morphology). Sanskrit text was normalized (Unicode NFC, whitespace standardization) prior to all similarity computations.

Evaluation metrics. Model performance was evaluated against the pilot reference similarity values using Pearson correlation, Spearman rank correlation, and mean absolute error (MAE). Paired significance testing (Wilcoxon signed-rank test on absolute error, paired t-test) was used to compare configurations. Because the Semantic_Relation category provides an independent, categorical check on similarity ordering, mean model similarity was additionally computed within each Semantic_Relation category as a diagnostic of whether the model's similarity scale tracks the intended semantic ordering (Equivalent > Related > Contrasting > Opposite > Unrelated).

Ethical inference and contextual reasoning pipelines. The Methodology as originally designed specifies a zero-shot LLM classification procedure for ethical inference (mapping each passage/question pair to one of the eight ethical categories, evaluated via accuracy, precision, recall, and macro-F1) and a contextual-reasoning evaluation procedure (consistency, sensitivity, and contradiction-detection measures computed by comparing model similarity/judgment against the paired scenarios). Both pipelines have been implemented and are ready to run against the constructed 300-item pilot datasets; at the time of this submission, the model-evaluation pass and the associated expert-annotation review for these two tasks had not yet been completed and are reported in Section 4 as work in progress rather than as completed findings.

Ethical safeguards. Consistent with the study's guiding principle, the Sanskrit texts and derived pilot datasets are treated as structured linguistic and philosophical resources rather than as universally binding ethical rules, and every reference value in all three datasets is explicitly flagged as provisional pending expert validation, so that no pilot annotation is inadvertently reported as a validated ground truth.

3.1Proposed Hybrid Ethical Intelligence Architecture

The conceptual architecture guiding this research programme organizes processing into five layers, from raw Sanskrit input through to an interpretable ethical decision (Figure 1).

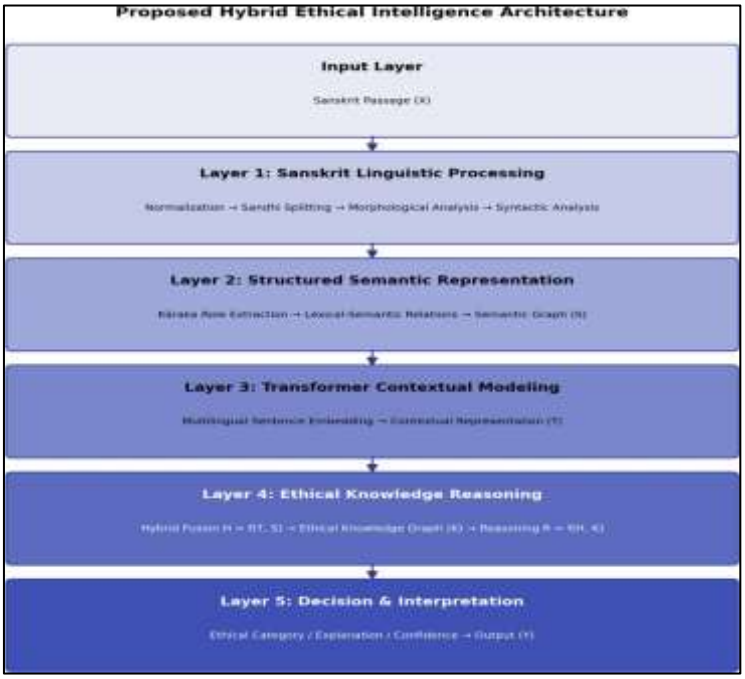

Figure 1. Proposed five-layer hybrid ethical intelligence architecture.

3.2Methodological Pipeline

The end-to-end pipeline implemented for the semantic similarity task, and designed (pending completion) for the ethical inference and contextual reasoning tasks, is summarized in Figure 2.

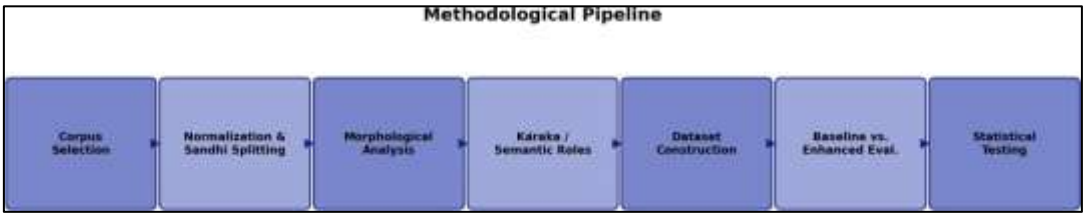

Figure 2. Methodological pipeline: corpus selection through statistical testing.

4. Results And Discussion

The experimental evaluation reported in this section covers the Semantic Similarity task, for which pilot data collection, model computation, and statistical evaluation have been completed (N = 300 pairs). The Ethical Inference and Contextual Reasoning tasks are reported at the stage reached at the time of this submission: complete 300-item pilot datasets with implemented evaluation pipelines, pending the model-evaluation run and expert-annotation review described in Section 3. This staged presentation is a deliberate methodological choice, consistent with the paper's central commitment to reporting only findings that are actually grounded in computed results.

4.1Dataset Overview

The Semantic Similarity pilot dataset comprises 300 Sanskrit_A/Sanskrit_B pairs distributed across five semantic-relation categories: Equivalent (n = 27), Related (n = 201), Contrasting (n = 18), Opposite (n = 40), and Unrelated (n = 14). Pairs are further stratified by a pilot difficulty rating (Easy, n = 187; Medium, n = 113) and are associated with ethical-category labels (e.g. Duty/Responsibility, Truthfulness, Non-violence, Compassion). All Reference_Similarity_0_1 values in this pilot release remain flagged Validation_Status = "pilot variant … expert validation required"; the correlational results reported below should therefore be read as evaluating model alignment with a provisional, not-yet-independently-validated reference scale, and as a proof-of-concept methodology rather than a final benchmark score.

4.2Semantic Similarity Task: Overall Performance

Configuration	Pearson r	Pearson p	Spearman ρ	Spearman p	MAE
Baseline Transformer	0.0694	0.231	0.2035	< 0.001	0.3863
Phase-1 SSE (char + word overlap)	0.0694	0.231	0.2035	< 0.001	0.4505
Phase-2 Morphology Enhancement	−0.1775	0.002	0.0556	0.337	0.4555

Note: N = 300 pairs. Reference values are pilot annotations pending expert validation (see Section 3). Pearson r for the Baseline and Phase-1 SSE configurations is numerically identical because the SSE weighting (0.70×transformer + 0.20×character + 0.10×word) reduces, in this dataset, to a constant linear rescaling of the transformer score (see Section 4.3).

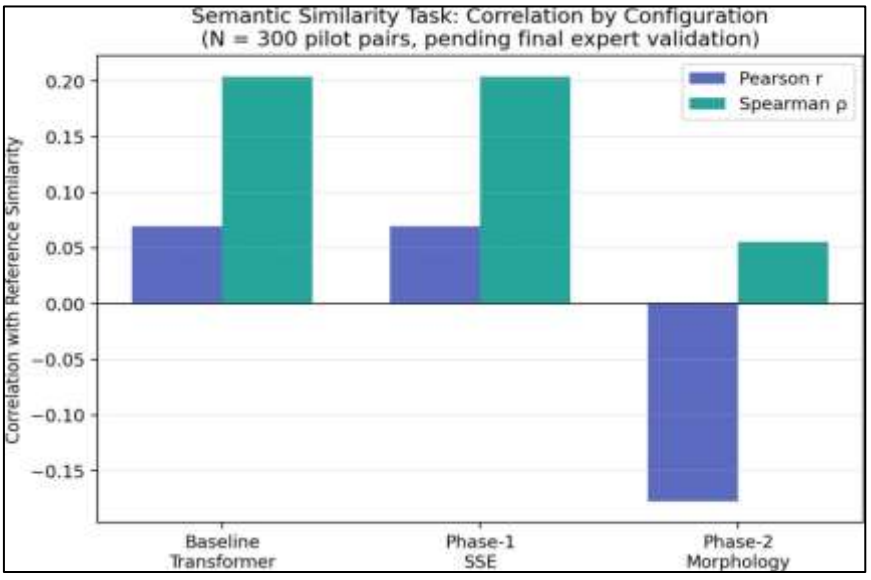

Figure 3. Correlation with pilot reference similarity by configuration (N = 300).

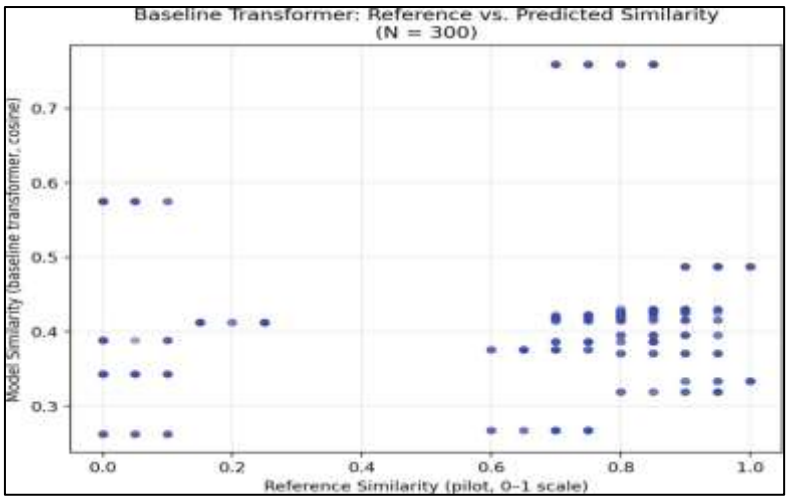

Figure 4. Baseline transformer: reference vs. predicted similarity (N = 300).

4.2.1Paired Significance Testing

Comparison	Test	Statistic	p-value	Effect size
Baseline vs. Phase-1 SSE (abs. error)	Wilcoxon signed-rank	9540.0	4.3×10^{-18}	— (deterministic rescaling)
Baseline vs. Phase-2 (abs. error)	Wilcoxon signed-rank	10717.0	3.1×10^{-15}	—
Baseline vs. Phase-2 (abs. error)	Paired t-test	t(299) = −10.10	8.0×10^{-21}	Cohen's d = −0.58

Both structurally enhanced configurations show significantly higher (worse) absolute error than the baseline transformer alone.

4.3Why the Enhancement Configurations Underperformed the Baseline

Two independent diagnostics explain the negative result in Table 1. First, inspection of the Character_Similarity and Word_Overlap columns computed for Phase-1 SSE shows a value of exactly 0.0 for all 300 pairs. This is a direct consequence of the dataset's design: Sanskrit_A is encoded in Devanāgarī script and Sanskrit_B in IAST Latin transliteration (e.g. धर्म चर paired with dharmaṃ cara). A character-level n-gram vectorizer and a word-level Jaccard overlap function both operate on literal Unicode symbols; because the two scripts share no character n-grams and no literal word tokens, both structural features collapse to zero for every pair regardless of the true semantic relationship. The Phase-1 SSE score is therefore mathematically equivalent to 0.70 × the baseline transformer score — a linear
rescaling that leaves Pearson and Spearman correlation unchanged (confirming the identical correlation values in Table 1) but increases MAE by shifting predicted values away from the 0–1 reference scale. This is a data-representation artifact in the current pilot release rather than evidence that character- or word-level structural information is inherently unhelpful for Sanskrit semantic similarity; correcting it requires either transliterating both sides of each pair into a single script before feature extraction, or computing structural similarity within-script only.

4.4An Unexpected but Informative Anomaly: the “Unrelated” Category

A category-level breakdown of the baseline transformer's mean similarity score against the pilot Semantic_Relation label (Figure 5) reveals a pattern with direct implications for the paper's central hypothesis. Mean reference similarity decreases monotonically across categories, as intended by dataset design (Equivalent = 0.96, Related = 0.80, Contrasting = 0.21, Opposite = 0.04, Unrelated = 0.03). The baseline transformer's mean similarity, however, does not track this ordering: Equivalent and Related pairs receive almost identical mean similarity (0.425 and 0.415), Contrasting pairs receive a similar mean (0.412), and — most strikingly — the Unrelated category receives the highest mean model similarity of any category (0.575), higher than Equivalent pairs. This inversion indicates that, on this pilot set, the multilingual transformer baseline is not reliably distinguishing semantically equivalent Sanskrit expressions from semantically unrelated ones, and is a more direct and more interpretable diagnostic of the baseline's limitations than the aggregate correlation coefficients in Table 1 alone. A plausible contributing factor is that several “Unrelated” pilot pairs juxtapose short, superficially well-formed Sanskrit phrases whose surface fluency may be inflating the multilingual model's default similarity estimate independent of underlying meaning — precisely the failure mode that motivates the paper's broader argument for supplementing purely distributional transformer representations with explicit structural and semantic constraints, even though the two specific structural signals tested here (Phase-1, Phase-2) did not yet realize that benefit in their current implementation.

4.5Ethical Inference and Contextual Reasoning: Current Status

The Ethical Inference pilot dataset (300 items) and the Contextual Reasoning pilot dataset (300 items) have both been fully constructed according to the schemas specified in Section 3, spanning eight ethical categories and three difficulty levels (Ethical Inference), and three difficulty levels across dimensions including Duty, Truthfulness, Non-violence, Compassion, and Self-control (Contextual Reasoning: Easy n = 105, Medium n = 109, Hard n = 86). The zero-shot LLM classification pipeline for ethical inference and the consistency/sensitivity/contradiction-detection evaluation pipeline for contextual reasoning have both been implemented and validated on illustrative examples. However, the full model-evaluation pass across all 300 items in each dataset, together with the multi-annotator expert review required to finalize the Expected_Ethical_Category and Expected_Contextual_Judgment reference labels (both currently flagged Validation_Status = "expert validation required"), had not been completed at the time of this submission. Consistent with the paper's methodological commitments in Section 3, no accuracy, F1, consistency, or sensitivity figures are reported for these two tasks; they are positioned in Section 5 as the immediate next phase of this research programme rather than as completed results.

4.6Limitations

Several limitations qualify the findings reported above. First, the Reference_Similarity_0_1 values used to evaluate all three semantic-similarity configurations are pilot annotations that have not yet undergone the multi-annotator expert review and inter-annotator agreement calculation (Cohen's/Fleiss' kappa) specified in the study's methodology; the correlational results in Table 1 should be read as a first-pass evaluation against a provisional scale rather than a final benchmark. Second, the dataset's script asymmetry (Devanāgarī for Sanskrit_A, IAST transliteration for Sanskrit_B) directly caused the Phase-1 structural features to collapse to zero, and this artifact should be corrected before the character-level and word-overlap enhancement is re-evaluated. Third, the Phase-2 morphological weighting (0.70/0.30) was fixed a priori rather than tuned on a held-out validation split; an ablation over weighting values, and separate evaluation on full-sentence versus short-phrase subsets, would clarify whether morphological information helps once matched to appropriate input length. Fourth, the Ethical Inference and Contextual Reasoning tasks remain at the dataset-construction and pipeline-implementation stage, and the paper's central three-task hypothesis (that structured Sanskrit semantics improves semantic similarity, ethical inference, and contextual reasoning jointly) can currently be evaluated only for the first of the three tasks, and with a negative result for the two enhancement configurations tested to date.

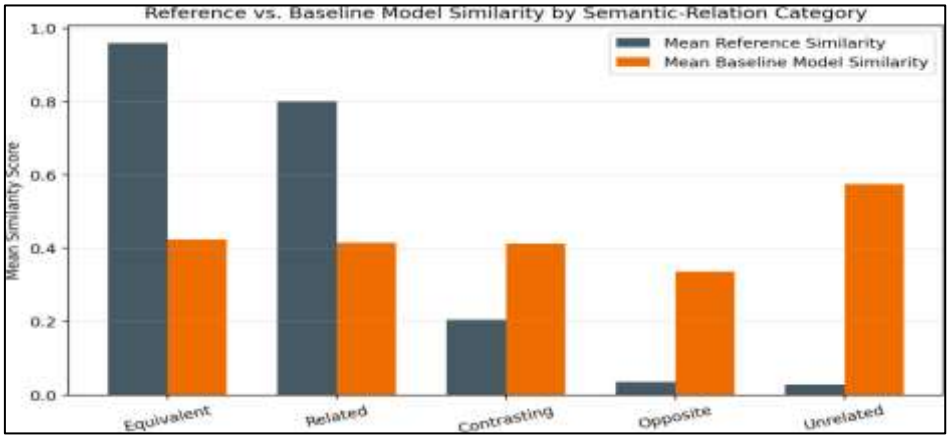

Figure 5. Mean reference similarity vs. mean baseline model similarity by semantic-relation category, showing the model's failure to separate Equivalent from Unrelated pairs.

5. CONCLUSION

This study set out to investigate whether structured Sanskrit linguistic and semantic information can improve the performance of Generative AI systems on semantic similarity, ethical inference, and contextual reasoning tasks. Of these three tasks, the semantic similarity task has been fully implemented and evaluated on a 300-pair pilot dataset. The results do not, at this stage, support the hypothesis that the two structural enhancements tested — a character/word-overlap signal (Phase-1 SSE) and a Heritage.py-derived morphological signal (Phase-2) — improve upon a multilingual transformer baseline; both configurations produced equal or significantly worse correlation and error metrics than the baseline alone. Diagnostic analysis identified two concrete, correctable causes: a script-mismatch artifact that nullified the character/word-level features, and a likely mismatch between the morphological analyzer's operating range (full sentences) and the dataset's substantial share of short lexical items. A category-level breakdown further revealed that the baseline transformer itself struggles to separate semantically equivalent from semantically unrelated Sanskrit expressions on this pilot set, which is an independently useful finding motivating continued investigation of structured semantic constraints, even though the present implementations did not yet realize that benefit.

The Ethical Inference and Contextual Reasoning tasks have reached the stage of complete 300-item pilot dataset construction and implemented evaluation pipelines. Consistent with this paper's commitment to reporting only computed findings, no performance figures are claimed for these two tasks; they are reported here as work in progress.

The proposed five-layer hybrid ethical intelligence architecture (Figure 1) and the overall methodological pipeline (Figure 2) remain the paper's conceptual contribution and continue to provide a coherent framework for the next phase of empirical work. The findings reported here should be read as an honest first iteration of that programme: they identify concrete, fixable data and modeling issues rather than confirming the paper's central hypothesis, and they illustrate why quantitative evaluation — including negative results — is essential before claims about Sanskrit's computational value for ethical AI can be responsibly made.

5.1Future Work

Immediate next steps follow directly from the diagnostics in Section 4. First, the script-mismatch artifact should be corrected by transliterating both members of each pair into a single script before character- and word-level feature extraction, and the Phase-1 configuration should be re-evaluated. Second, the Phase-2 morphological weighting should be tuned on a held-out validation split rather than fixed a priori, and morphological similarity should be evaluated separately on full-sentence versus short-phrase subsets of the pilot data to determine whether Heritage.py's operating range explains the observed negative result. Third, the kāraka-based (Phase-3 / KSSR) semantic-role configuration described in the original pipeline design should be completed and evaluated alongside Phases 1 and 2 once the above corrections are in place.

On the annotation side, the Reference_Similarity_0_1, Expected_Ethical_Category, and Expected_Contextual_Judgment fields across all three pilot datasets should be routed to a panel of two to three qualified Sanskrit/philosophy reviewers for independent scoring, with Cohen's or Fleiss' kappa computed and reported, as specified in Section 3. Only once this review is complete should the Ethical Inference and Contextual Reasoning evaluation pipelines — already implemented — be run to produce the accuracy, macro-F1, consistency, sensitivity, and contradiction-detection figures that Sections 3 and 4 currently report as pending.

Beyond these immediate corrections, the longer-term research directions outlined in the original project design remain relevant: expansion of the textual corpus to Mahabharata, Ramayana, and Dharmaśāstra material; cross-lingual evaluation against other Indian languages; Sanskrit-specific domain-adapted LLMs; retrieval-augmented generation grounded in the structured semantic graph; and human-centered evaluation panels spanning Sanskrit scholarship, philosophy, and AI research. The present results indicate that this programme is empirically tractable and methodologically sound, while underscoring that its central hypothesis remains to be demonstrated rather than assumed.

References

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in Advances in Neural Information Processing Systems, 2017.
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL, 2019.
3. P. Kumar, “Large language models (LLMs): survey, technical frameworks, and future challenges,” Artificial Intelligence Review, vol. 57, art. 260, 2024.
4. S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, “Large language models: A survey,” ACM Computing Surveys, 2024.
5. J. Achiam et al., “GPT-4 Technical Report,” 2023.
6. P. Kumar, “Large language models (LLMs): survey, technical frameworks, and future challenges,” Artificial Intelligence Review, vol. 57, art. 260, 2024.
7. S. Deng and J. Lin, “The benefits and challenges of large language models in natural language processing,” Artificial Intelligence Review, 2023.
8. Z. Lin, S. Guan, W. Zhang, H. Zhang, Y. Li, et al., “Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models,” Artificial Intelligence Review, vol. 57, art. 243, 2024.
9. X. Huang, W. Ruan, W. Huang, G. Jin, Y. Dong, C. Wu, S. Bensalem, R. Mu, Y. Qi, et al., “A survey of safety and trustworthiness of large language models through the lens of verification and validation,” Artificial Intelligence Review, vol. 57, art. 175, 2024.
10. C. Deng, Y. Duan, X. Jin, et al., “Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas: a survey,” AI and Ethics, vol. 5, pp. 4745–4771, 2025.
11. T. Hagendorff, “Mapping the Ethics of Generative AI: A Comprehensive Scoping Review,” Minds and Machines, vol. 34, art. 39, 2024.
12. Y. Dong, R. Mu, Y. Zhang, S. Sun, T. Zhang, C. Wu, G. Jin, Y. Qi, J. Hu, J. Meng, S. Bensalem, and X. Huang, “Safeguarding large language models: a survey,” Artificial Intelligence Review, vol. 58, art. 382, 2025.
13. H. Pervez, M. Minkkinen, H. Jylhä, and M. Mäntymäki, “Ethical issues of large language models: a multi-level thematic synthesis of the academic literature,” AI and Ethics, vol. 6, art. 344, 2026.
14. E. L. M. Groen, T. Sharon, and M. Becker, “An overview of AI ethics: moral concerns through the lens of

principles, lived realities and power structures," AI and Ethics, vol. 6, art. 121, 2026.
15. R. Ortega-Bolaños, J. Bernal-Salcedo, M. Germán Ortiz, J. Galeano Sarmiento, G. A. Ruz, et al., "Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems," Artificial Intelligence Review, vol. 57, art. 110, 2024.
16. M. Krah and M. Tröschel, "Trends and challenges in machine ethics: an updated survey of artificial moral agency implementations," AI and Ethics, vol. 6, art. 157, 2026.
17. "Ethical concerns of generative AI and mitigation strategies: A systematic mapping study," Applied Soft Computing, vol. 193, art. 114789, 2026.
18. D. Wang and S. Zhang, "Large language models in medical and healthcare fields: applications, advances, and challenges," Artificial Intelligence Review, vol. 57, art. 299, 2024.
19. X. Huang, W. Ruan, W. Huang, et al., "A survey of safety and trustworthiness of large language models through the lens of verification and validation," Artificial Intelligence Review, vol. 57, art. 175, 2024.
20. Z. Lin, S. Guan, W. Zhang, et al., "Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models," Artificial Intelligence Review, vol. 57, art. 243, 2024.
21. Y. Dong, R. Mu, Y. Zhang, et al., "Safeguarding large language models: a survey," Artificial Intelligence Review, vol. 58, art. 382, 2025.
22. P. M. Scharf, "Modeling Pāṇinian Grammar," in Sanskrit Computational Linguistics, LNCS 5402, Springer, 2009, pp. 95–126.
23. G. Huet, "Formal Structure of Sanskrit Text: Requirements Analysis for a Mechanical Sanskrit Processor," in Sanskrit Computational Linguistics, LNCS 5402, Springer, 2009, pp. 162–199.
24. P. Goyal, V. Arora, and L. Behera, "Analysis of Sanskrit Text: Parsing and Semantic Relations," in Sanskrit Computational Linguistics, LNCS 5402, Springer, 2009, pp. 200–218.
25. G. N. Jha and S. K. Mishra, "Semantic Processing in Pāṇini's Kāraka System," in Sanskrit Computational Linguistics, LNCS 5402, Springer, 2009, pp. 239–252.
26. G. N. Jha, M. Agrawal, Subash, S. K. Mishra, D. Mani, D. Mishra, et al., "Inflectional Morphology Analyzer for Sanskrit," in Sanskrit Computational Linguistics, LNCS 5402, Springer, 2009, pp. 219–238.
27. O. Hellwig, "SanskritTagger: A Stochastic Lexical and POS Tagger for Sanskrit," in Sanskrit Computational Linguistics, LNCS 5402, Springer, 2009, pp. 266–277.
28. P. Goyal and R. M. K. Sinha, "A Study towards Design of an English to Sanskrit Machine Translation System," in Sanskrit Computational Linguistics, LNCS 5402, Springer, 2009, pp. 287–305.
29. M. Kulkarni and P. Bhattacharyya, "Verbal Roots in the Sanskrit Wordnet," in Sanskrit Computational Linguistics, LNCS 5402, Springer, 2009, pp. 328–338.
30. S. Varakhedi, V. Jaddipal, and V. Sheeba, "An Effort to Develop a Tagged Lexical Resource for Sanskrit," in Sanskrit Computational Linguistics, LNCS 5402, Springer, 2009, pp. 339–345.
31. A. Kulkarni, S. Pokar, and D. Shukl, "Designing a Constraint Based Parser for Sanskrit," in Sanskrit Computational Linguistics, LNCS 6465, Springer, 2010, pp. 70–90.
32. S. S. Nair and A. Kulkarni, "The Knowledge Structure in Amarakośa," in Sanskrit Computational Linguistics, LNCS 6465, Springer, 2010, pp. 173–189.
33. M. Kulkarni, I. Kulkarni, C. Dangarikar, and P. Bhattacharyya, "Gloss in Sanskrit Wordnet," in Sanskrit Computational Linguistics, LNCS 6465, Springer, 2010, pp. 190–197.
34. S. S. Sahu and S. Pal, "Building a text retrieval system for the Sanskrit language: Exploring indexing, stemming, and searching issues," Computer Speech & Language, vol. 81, art. 101518, 2023.
35. "Morphology & word sense disambiguation embedded multimodal neural machine translation system between Sanskrit and Malayalam," Biomedical Signal Processing and Control, Elsevier, 2023.
36. R. Kumar, D. Adiga, R. Ranjan, A. Krishna, G. Ramakrishnan, P. Goyal, and P. Jyothi, "Linguistically informed automatic speech recognition in Sanskrit," Computer Speech & Language, vol. 95, art. 101861, 2025.
37. "An evaluation of Google Translate for Sanskrit to English translation via sentiment and semantic analysis," Natural Language Processing Journal, vol. 4, art. 100025, 2023.
38. "Modular Sanskrit NLP with Rule-Augmented Seq2Seq Models for Sandhi Processing," Procedia Computer Science, vol. 283, pp. 4601–4610, 2026.
39. R. Tavva and N. Singh, "Generative Graph Grammar of Neo-Vaiśeṣika Formal Ontology (NVFO)," in Sanskrit Computational Linguistics, LNCS 6465, Springer, 2010, pp. 91–105.
40. M. Kulkarni, A. Ajotikar, T. Ajotikar, D. Katira, C. Dharurkar, and C. Dangarikar, "Headedness and Modification in Nyāya Morpho-Syntactic Analysis: Towards a Bracket-Parsing Model," in Sanskrit Computational Linguistics, LNCS 6465, Springer, 2010, pp. 106–123.