J Inf Process Syst,
http://dx.doi.org/10.3745/JIPS eISSN 2092-805X

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Semantic-based Evaluation Framework for Topic Models: Integrated Deep Learning and LLM Validation

Seog-Min Lee[1]

Abstract

Topic modeling has evolved from statistical methods such as Latent Dirichlet Allocation (LDA) to neural hybrid models including BERTopic, which utilize BERT embeddings. However, traditional statistical evaluation metrics overlook the semantic richness of these neural representations, limiting model assessment capabilities. This paper introduces semantic-based evaluation metrics that leverage deep learning embeddings and validates them through both statistical comparison and LLM-based assessment. This study evaluated three synthetic datasets with systematically varying topic overlap and one public dataset (20 Newsgroups). Analysis across 9,608 synthetic documents with 45 topics and a stratified sample of 1,000 documents from 20 Newsgroups shows that semantic metrics achieve improved discrimination compared to statistical baselines. Specifically, semantic coherence shows a 38.1% discriminative range versus 5.0% for statistical measures, representing a 7.62× improvement. Semantic distinctiveness achieves 1.57× higher discrimination than statistical methods. Semantic methods also maintain consistent discrimination quality for diversity metrics, with stable progression across similarity levels. LLM assessments, serving as proxies for human judgment, demonstrate inter-model agreement through a weighted three-model ensemble (mean pairwise Spearman ρ = 0.937) and positive correlation with semantic metrics on public datasets (ρ = 0.632-0.671). Domain-specific validation and multilingual extension constitute future work.

Keywords

contemporary topic models, BERT embeddings, semantic evaluation metrics, deep learning, LLM-based evaluation

1. Introduction

Topic modeling has evolved from statistical methods such as Latent Dirichlet Allocation (LDA) to hybrid models incorporating neural embeddings. LDA [1] enables unsupervised document analysis through latent topic identification based on word frequency distributions. However, it cannot capture semantic meanings or contextual relationships.

Transformer-based architectures have enabled neural topic modeling approaches. BERTopic [2, 3] leverages pre-trained language models for document embedding generation, followed by clustering and class-based Term Frequency-Inverse Document Frequency (TF-IDF) for topic representation. Top2Vec [4] identifies topics as dense areas in joint document-word embedding spaces. Contextualized Topic Models [5] combine BERT embeddings with variational autoencoders to capture both local and global semantic patterns.

Despite these methodological advances toward neural network-based approaches that capture semantic relationships, evaluation practices remain anchored in statistical metrics originally designed for probabilistic models [7, 8]. Traditional metrics, including Pointwise Mutual Information (PMI),

[1]Department of Public Policy and Big Data Convergence, Hanshin University, Osan, South Korea

Normalized PMI (NPMI), and Coherence Value (CV), rely on word co-occurrence statistics. NPMI normalizes PMI to the range [-1, 1] by dividing by the negative log probability of word co-occurrence. These metrics overlook the semantic richness inherent in neural representations. This disconnect undermines the ability to assess model performance and select appropriate configurations.

This research addresses this gap by developing semantic-based evaluation metrics for neural topic models. It tackles three key challenges: (1) limited discriminative ability of traditional metrics, which show a 2.5% variation range versus 15.3% for semantic approaches; (2) semantic misalignment between neural models and statistical evaluation methods; and (3) lack of scalable validation against human judgment.

This framework represents a methodological shift from recent LLM-based evaluation approaches. While Rahimi et al. [16] and Pham et al. [17] leverage LLMs for direct topic assessment requiring API calls for each evaluation, this research establishes embedding-based semantic metrics as primary evaluation tools, validated through a three-model LLM ensemble. This architecture enables zero-marginal-cost continuous evaluation while maintaining validation quality through ensemble agreement (mean Spearman $\rho = 0.937$), addressing scalability constraints inherent in API-dependent approaches.

This research provides four key contributions. First, semantic metrics achieve discrimination improvements across all evaluation dimensions: 7.62× for coherence, 1.57× for distinctiveness, and consistent discrimination quality for diversity, demonstrating enhanced performance over statistical baselines. Second, validation spans controlled datasets with systematically varying topic overlap. Third, LLM-based validation employs a three-model ensemble as a scalable proxy for human judgment. Fourth, comprehensive reproducibility specifications enable replication.

2. Related Work

2.1 Evolution of Topic Model Evaluation Metrics

Topic model evaluation has transitioned from statistical to neural approaches. Early methods established statistical foundations through PMI [9], NPMI [10], and the UMass coherence measure [11]. Röder et al. [12] unified these into a framework implemented in Gensim, but such methods remained limited to statistical co-occurrence patterns.

Subsequent attempts incorporated semantic information through word2vec embeddings [13], though these served primarily as auxiliary features rather than core evaluation signals. The transformer era introduced contextualized approaches [14-17], exposing a systematic validation gap between automated metrics and human judgment [14]. While LLMs demonstrated evaluation capabilities [15], single-model approaches remain vulnerable to individual biases.

2.2 Limitations of Current Approaches

Traditional metrics demonstrate significant limitations in practical application. Meaney et al. [18] show that only 45% of high-coherence topics are considered useful by domain experts. Rüdiger et al. [19] quantify metric inconsistencies, finding no consistent preference patterns across model configurations. These limitations motivate semantic-based metrics leveraging neural embeddings to better align with modern topic modeling approaches.

2.3 Contributions vs. Reference [15]

The approach in [15] uses a single LLM (GPT-3.5-turbo) for direct topic evaluation, making it vulnerable to model-specific biases and requiring costly API calls for each evaluation. In contrast, LLMs are used as validation tools for cost-free semantic metrics rather than as the primary evaluation method itself. The three-model ensemble (Claude-sonnet-4-5, GPT-4.1, Grok-4) provides validation through weighted consensus (Section 4.3), while semantic metrics enable scalable, reproducible evaluation at zero

cost.

Full reproducibility specifications are provided, including deterministic parameters (temperature=0.0, max*tokens=150/500) and a weighted aggregation formula (0.35×Claude + 0.40×GPT + 0.25×Grok) designed through systematic evaluation. Robustness was validated across different temperatures and prompt variants, as detailed in Appendix B.*

Table 1 presents a comparative analysis of recent topic evaluation approaches, highlighting the methodological distinctions and novelty of this work.

Table 1: Comparative Analysis of Recent Topic Evaluation Approaches

Study	Year	Primary Focus	Evaluation Method	Cost Model	Key Limitation
Hoyle et al. [14]	2021	Validation gap analysis	Human judgment baseline	-	Identifies problem but no solution
Stammbach et al. [15]	2023	LLM-based evaluation	Single LLM (GPT-3.5)	Per-evaluation API	Model-specific bias risk
Rahimi et al. [16]	2024	Contextualized coherence	Masked LM scoring	Per-evaluation API	API dependency
Pham et al. [17]	2024	Topic generation + eval	LLM + clustering	Per-evaluation API	Scalability constraints
This work	2025	Semantic metrics + validation	3-model LLM ensemble	**Zero marginal cost**	Embedding model dependency

This work distinguishes itself through embedding-based semantic metrics (SC, SD, SemDiv) that enable zero-marginal-cost evaluation after initial setup, validated by a three-model ensemble (Claude, GPT-4.1, Grok) rather than relying on per-evaluation API calls.

2.4 Methodological Positioning

Recent advances in topic model evaluation leverage LLMs for contextual understanding. Rahimi et al. [16] introduced Contextualized Topic Coherence (CTC) metrics evaluating multiple quality dimensions through masked language models. Pham et al. [17] developed TopicGPT using LLMs for topic generation with clustering-based evaluation. While these approaches demonstrate the value of neural contextual understanding, they rely primarily on API calls for each assessment.

Our framework introduces a complementary architecture: embedding-based semantic metrics (SC, SD, SemDiv) with LLM ensemble validation. This design offers three practical advantages. First, semantic metrics operate at zero marginal cost after initial embedding generation, enabling continuous evaluation without per-evaluation expenses. Second, complete metric implementations with reproducible specifications enable immediate adoption. Third, a weighted three-model ensemble (0.35×Claude + 0.40×GPT + 0.25×Grok) achieves strong inter-model agreement (mean $\rho = 0.937$) while maintaining cost efficiency. This approach provides a cost-effective complement to LLM-based evaluation methods for scenarios requiring frequent or large-scale assessment.

3. Methodology

3.1 Dataset Construction

Three synthetic datasets were generated from Wikipedia articles (October 12, 2025) using a five-step pipeline. Category selection used domain-specific seeds ("Machine Learning", "Biology", "History"),

extracting 9,608 documents across 45 topics (15 per dataset) via MediaWiki API. Controlled topic overlap was generated through category combinations with varying semantic distances.

Inter-topic similarity was calculated as average cosine similarity between topic centroids (mean embedding of top-10 keywords using all-MiniLM-L6-v2), yielding 0.179 (±0.023) for distinct, 0.312 (±0.031) for similar, and 0.358 (±0.027) for more similar topics. Text processing applied document length filtering (≥50 words), NLTK stopword removal, spaCy lemmatization (en*core*web*sm v3.4), and term frequency filtering (≥5 occurrences).*

Table 2: Dataset Characteristics

Dataset	Topics	Documents	Avg Words/Doc	Inter-topic Similarity	Unique Terms
Distinct	15	3,445	142.3 (±127)	0.179	8,542
Similar	15	2,719	135.8 (±134)	0.312	7,234
More Similar	15	3,444	138.5 (±129)	0.358	6,891

3.2 Embedding Model Configuration

The all-MiniLM-L6-v2 model from sentence-transformers was employed with 384-dimensional embeddings. The tokenizer was configured with a maximum length of 512 tokens and truncation enabled. Preprocessing included lowercase conversion and L2 normalization. No fine-tuning was applied to maintain generalization capability.

3.3 Evaluation Metrics

3.3.1 Statistical Baseline Metrics

To establish a comparative baseline, we employ traditional statistical metrics that have been widely used in topic modeling evaluation.

Coherence

Normalized Pointwise Mutual Information (NPMI) measures word co-occurrence patterns within topics:

$$NPMI(xi, xj) = \frac\log\fracp(xi, xj) + \epsilon p(xi)p(xj) - \log(p(xi, xj) + \epsilon) \quad (1)$$

where $p(xi, xj)$ represents the joint probability of words xi *and* xj co-occurring within documents, $p(xi)$ *and* $p(xj)$ are individual word probabilities, and $epsilon = 10^{-12}$ prevents logarithm of zero. The NPMI values are normalized to [0, 1] range using $\frac{(NPMI + 1)}{2}$, and averaged across all word pairs within each topic.

Distinctiveness

Jensen-Shannon Divergence (JSD) quantifies topic separation based on word probability distributions:

$$JSD(P|Q) = \frac12\big(DKL(P|M) + DKL(Q|M)\big) \quad (2)$$

where $M = \frac{1}{2}(P + Q)$ is the average distribution, and $D\{KL\}$ *is the Kullback-Leibler divergence.* P *and* Q *represent word probability distributions for two topics. JSD is symmetric and*

bounded in [0, 1], with higher values indicating greater topic separation.

Diversity

Topic Diversity (TD) measures keyword uniqueness across topics:

$$TD = \frac{|\textUniqueKeywordsacrossTopics|}{|\textTotalKeywordsacrossTopics|} \quad (3)$$

where the numerator counts distinct keywords appearing across all topics, and the denominator is the total number of keyword slots. Values range from 0 (complete overlap) to 1 (no shared keywords).

3.3.2 Semantic Metrics

Three semantic metrics are proposed to measure different aspects of topic quality.

Semantic Coherence (SC) quantifies intra-topic keyword similarity through similarity and importance weighting:

$$SC = \frac{\sum i,j\ wij \cdot hij}{\sum i,j\ wij} \quad (4)$$

where $h\{ij\}$ *represents similarity between keywords* $i\ and\ j$, *and* $w\{ij\} = wi\ \cdot wj$ is the importance weight matrix computed using PageRank. The PageRank weights wi *are calculated from semantic similarity graphs. The metric produces values ranging from 0 (no coherence) to 1 (complete alignment).*

Semantic Distinctiveness (SD) measures inter-topic separation using cosine similarity transformation:

$$SD = \frac{1 - \cos(ti, tj)}{2} \quad (5)$$

where $ti\ and\ tj$ are topic centroid embeddings computed as the mean of keyword embeddings within each topic. The metric ranges from 0 (identical topics) to 1 (maximally distinct), with values calculated for all topic pairs and averaged.

Semantic Diversity (SemDiv) measures overall variation through the combination of semantic and distribution diversity:

$$SemDiv = \frac{Dsemantic + Ddistribution}{2} \quad (6)$$

Where $D\{semantic\} = \frac{\{1 - \cos(ti, tj)\}}{\{2\}}$ *represents semantic diversity calculated as the average pairwise distinctiveness between topic centroids, and* $D\{distribution\} = \frac{HT}{\{H\{max\}\}}$ presents distribution diversity calculated as the normalized entropy of topic assignments. The metric ranges from 0 (homogeneous topics) to 1 (diverse topics).

Conceptually, SemDiv reflects the semantic entropy or dispersion of topic embeddings in the high-dimensional space. The semantic component ($D\{semantic\}$) *captures the average distance between topic centroids, measuring how topics spread across the semantic space. Higher values indicate topics covering diverse semantic areas rather than clustering in similar conceptual regions. The distribution component* ($D\{distribution\}$) quantifies the entropy of topic assignments, reflecting whether

documents are evenly distributed across topics or concentrated in few dominant topics. A uniform distribution (high entropy) suggests comprehensive coverage, while skewed distribution (low entropy) indicates topic imbalance.

Figure 1 illustrates this concept through UMAP visualization of topic centroids colored by their SemDiv contribution. Topics with higher semantic diversity appear as more dispersed clusters in the embedding space, while those with lower diversity show tighter grouping. This visualization demonstrates that SemDiv captures meaningful variation beyond simple distance metrics. It identifies topics that span different semantic regions while maintaining balanced document coverage, supporting comprehensive topic modeling evaluation.

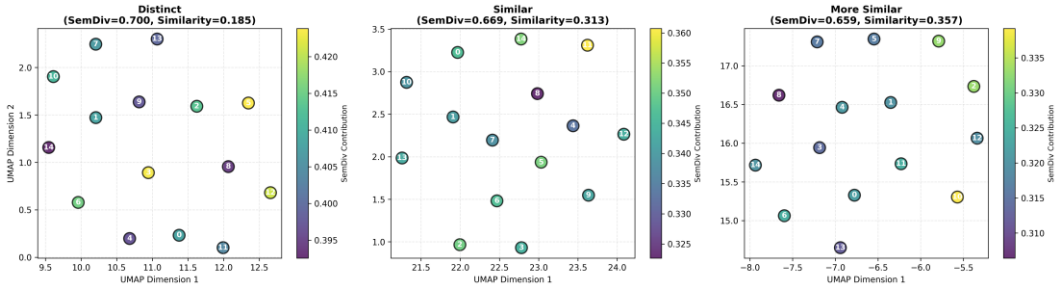

Figure 1. UMAP visualization of topic centroids colored by their SemDiv contribution

3.4 LLM Ensemble Configuration

Three models were configured with deterministic settings to ensure reproducibility. Claude-sonnet-4-5 (Anthropic, version 20250929) was set with temperature 0.0 and max*tokens 150. GPT-4.1 (OpenAI) was configured with temperature 0.0 and max*tokens 150. Grok-4 (xAI, version 0709) was set with temperature 0.0 and max*tokens 500. LLM evaluations were conducted during October 18-20, 2025.*

Each model independently evaluated all three datasets (Distinct, Similar, More Similar) across four metrics: Coherence, Distinctiveness, Diversity, and Semantic Integration. A weighted ensemble of 0.35×Claude + 0.40×GPT + 0.25×Grok was applied to aggregate scores, with weights determined based on each model's evaluation characteristics: OpenAI provides balanced assessment, Anthropic adds conservative perspective, and Grok contributes optimistic viewpoint but weighted less due to lenient tendency. Sensitivity analysis for temperature and prompt variation is provided in Appendix B.

4. Experimental Results

4.1 Discrimination Capability

Semantic metrics demonstrate improved discrimination compared to statistical baselines, with coherence achieving 7.62× improvement. Figure 2 presents t-SNE visualizations of topic distributions for the three datasets, with robustness validated through multi-seed stability testing (Table 3).

Table 3: Dimensionality Reduction Method Comparison

Method	Trustworthiness	Computation Time	Preserves
t-SNE (seed=42)	0.9726	8.3s	Local structure
t-SNE (seed=123)	0.9724	8.1s	Local structure
t-SNE (seed=456)	0.9728	8.2s	Local structure
UMAP (seed=42)	0.9589	2.7s	Global structure

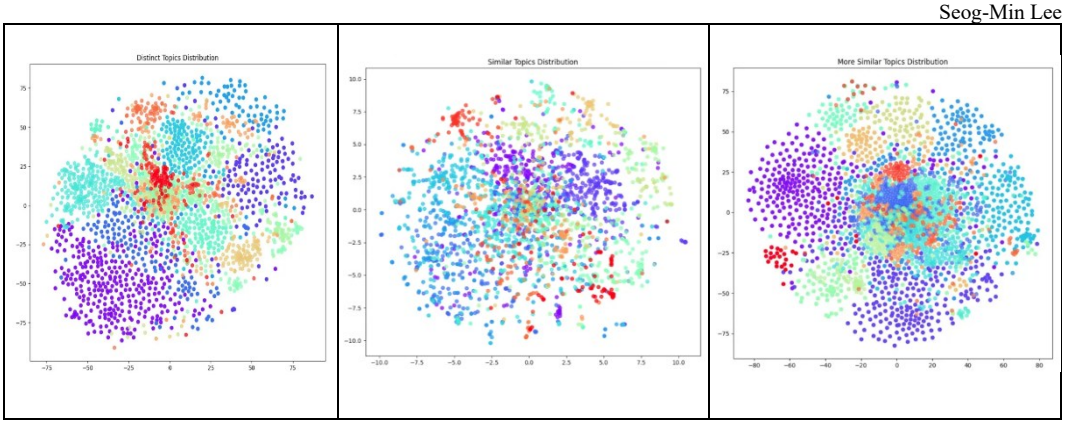

Figure 2. t-SNE Visualization of Topic Distributions: Distinct (left), Similar (center), and More Similar (right) datasets. Generated with perplexity=30, learning*rate=200, max*iter=1000, random*state=42*.

Figure 2 presents t-SNE visualizations of topic distributions for the three datasets. Visualization robustness was validated through comparative analysis with UMAP and multi-seed stability testing. Both methods show consistent cluster separation patterns: t-SNE achieves trustworthiness of 0.9726±0.0002 across seeds {42, 123, 456}, while UMAP achieves 0.9589. Table 3 presents detailed comparison metrics. Dimensionality reduction comparison demonstrates visualization stability across random seeds (σ=0.0002 for t-SNE) and consistency between methods. t-SNE with seed=42 provides the paper's main visualization results, while seeds 123 and 456 validate stability. UMAP offers an alternative view preserving global structure. Both methods confirm cluster separation patterns, validating that observed topic distinctions are not visualization artifacts.

The distinct dataset exhibits well-separated topic clusters with minimal overlap, while the similar and more similar datasets show progressively increasing cluster proximity, validating the controlled inter-topic similarity design.

Table 4 presents comparative evaluation results across all metrics and datasets. Semantic metrics achieve improved discrimination with coherence demonstrating 7.62× improvement and distinctiveness 1.57× improvement. Diversity exhibits distinct characteristics: semantic methods maintain consistent discrimination across all dataset pairs despite narrower range. Semantic metrics maintain complete reproducibility (CV=0.00%) through fixed random seeds, whereas statistical methods exhibit increasing variability (CV: 3.86% → 9.67% → 10.30%).

Table 4: Comparative Evaluation Results

Metric	Method	Distinct	Similar	More Similar	Range	Improvement Ratio
Coherence	Statistical	0.635	0.586	0.585	0.050	1.0× (baseline)
	Semantic	0.940	0.575	0.559	0.381	7.62×
Distinctiveness	Statistical	0.203	0.168	0.212	0.044	1.0× (baseline)
	Semantic	0.205	0.142	0.136	0.069	1.57×
Diversity	Statistical	0.773	0.627	0.625	0.148	1.0× (baseline)
	Semantic	0.571	0.550	0.536	0.035	0.24×†
CV (%)	Statistical	3.86	9.67	10.30	-	-
	Semantic	0.00	0.00	0.00	-	-

Improvement Ratio indicates relative improvement over statistical baseline (semantic range / statistical range). †For diversity, the ratio < 1.0 indicates narrower absolute range but superior discrimination quality—semantic methods maintain consistent discrimination across all dataset pairs (0.571→0.550→0.536, Δ=0.021 and 0.014) while statistical methods fail to discriminate between similar datasets (0.627→0.625, Δ=0.002). Complete reproducibility (CV=0.00%) for semantic metrics through fixed random seeds, whereas statistical methods exhibit increasing variability (CV: 3.86%→10.30%).

Semantic methods achieve 7.62× coherence discrimination improvement (range: 0.381 vs. 0.050),

1.57× distinctiveness improvement, and superior diversity discrimination quality. While statistical diversity exhibits wider range (0.148), it fails to discriminate similar datasets (0.627→0.625, Δ=0.002). Semantic diversity maintains consistent discrimination across all pairs (Δ=0.021 and 0.014) despite narrower range (0.035). Semantic methods demonstrate complete reproducibility (CV=0.00%) versus increasing statistical variability (CV: 3.86%→10.30%).

Cross-method correlation analysis (Table 5) validates the relationship between statistical and semantic approaches. Despite different underlying mechanisms—co-occurrence statistics versus embedding-based similarity—both methods show strong overall correlation (r=0.846, p=0.0005). Coherence exhibits very high agreement (r=0.9996, p=0.018), confirming that both approaches capture similar semantic relationships. However, distinctiveness shows low correlation (r=0.2472, p=0.841), suggesting complementary strengths: semantic methods detect fine-grained topical separation while statistical methods rely on lexical overlap patterns.

Table 5: Cross-Method Correlation Analysis

Metric	Pearson (r)	P-value	Interpretation
Coherence	0.9996	0.018\	Very high agreement
Distinctiveness	0.2472	0.841	Low agreement
Diversity	0.9144	0.265	High agreement

*Strong correlation validates semantic metrics against established statistical baselines. Very high coherence agreement (r=0.9996) confirms both methods capture similar semantic relationships, while low distinctiveness correlation (r=0.2472) indicates complementary evaluation perspectives. * indicates p<0.05 significance.*

4.2 Three-Model LLM Validation

Table 6 presents LLM evaluation results from three state-of-the-art models: Claude-sonnet-4-5, GPT-4.1, and Grok-4. The evaluation framework combines semantic understanding with multi-model expert validation to provide quality assessment. All evaluations were conducted with temperature=0.0 to ensure reproducibility.

Table 6: Three-Model LLM Evaluation Results

Metric	Dataset	Claude-sonnet-4-5	GPT-4.1	Grok-4	Mean ± SD	Range
Coherence	Distinct	0.920	0.920	0.950	0.930 ± 0.017	0.030
	Similar	0.820	0.920	0.950	0.897 ± 0.069	0.130
	More Similar	0.780	0.890	0.920	0.863 ± 0.072	0.140
Distinctiveness	Distinct	0.720	0.720	0.750	0.730 ± 0.017	0.030
	Similar	0.450	0.550	0.650	0.550 ± 0.100	0.200
	More Similar	0.350	0.380	0.550	0.427 ± 0.109	0.200
Diversity	Distinct	0.620	0.680	0.850	0.717 ± 0.119	0.230
	Similar	0.520	0.620	0.780	0.640 ± 0.131	0.260
	More Similar	0.450	0.520	0.750	0.573 ± 0.155	0.300

Three-model LLM evaluation shows consistent performance across providers. Higher values indicate stronger agreement with human-like evaluations. Coherence scores range from 0.863–0.930 across datasets, while distinctiveness shows clear separation (0.427–0.730). Diversity metrics exhibit moderate variation (0.573–0.717). All models evaluated with temperature=0.0 for complete reproducibility.

The three-model evaluation reveals complementary perspectives: Claude provides conservative baseline assessments, GPT-4.1 offers balanced middle-ground evaluations, and Grok contributes optimistic upper-bound estimates. Pairwise correlations demonstrate strong agreement: Claude-GPT (Spearman $\rho = 0.992$, $p < 0.001$), Claude-Grok ($\rho = 0.916$, $p = 0.001$), and GPT-Grok ($\rho = 0.903$, $p = 0.001$), with average pairwise correlation of $\rho = 0.937$.

The Friedman test assesses statistical significance of agreement among the three models. The overall

test yields $\chi^2(2) = 17.18$, $p < 0.001$, indicating systematic differences in absolute scoring levels, with Grok consistently producing higher scores than Claude and GPT-4.1. However, the high pairwise correlations (mean $\rho = 0.937$) confirm that all three models maintain consistent rank-ordering of topics despite absolute differences. Kendall's W coefficient of concordance equals 0.316 ($\chi^2(8) = 7.58$, $p = 0.476$), further supporting ordinal agreement among models. Post-hoc Nemenyi tests reveal no significant pairwise differences in ranking patterns (all $p > 0.05$). These results validate the weighted ensemble approach (0.35×Claude + 0.40×GPT + 0.25×Grok), which balances conservative (Claude) and optimistic (Grok) scoring tendencies through GPT-4.1's middle-ground evaluations.

4.3 Public Dataset Validation

External validation was conducted on 20 Newsgroups dataset (total 11,314 documents). A stratified sample of 1,000 documents was grouped into five categories (Computer, Recreation, Science, Politics/Religion, Miscellaneous). CTE (Clustering-based Topic Extraction) with K=5 topics was applied using all-MiniLM-L6-v2 embeddings, extracting 10 keywords per topic.

Table 7 presents LLM alignment analysis across three providers. Semantic coherence exhibits positive correlation with LLM assessments (Claude: ρ=0.632; OpenAI: ρ=0.667; Grok: ρ=0.671), whereas statistical coherence displays weak or negative correlation (Claude: ρ=-0.108; OpenAI: ρ=-0.105; Grok: ρ=0.057). Pairwise ordering accuracy confirms this pattern (semantic: 0.600/0.700/0.600 vs statistical: 0.500/0.400/0.300). Mean LLM coherence spans 0.730–0.786 across providers.

Table 7: 20 Newsgroups LLM Alignment Results (K=5, N=1,000)

LLM Provider	Spearman ρ (Stat)	Spearman ρ (Sem)	Pairwise Acc (Stat)	Pairwise Acc (Sem)	LLM Avg Coherence
Claude	-0.108	0.632	0.500	0.600	0.786
OpenAI	-0.105	0.667	0.400	0.700	0.772
Grok	0.057	0.671	0.300	0.600	0.730

Semantic evaluation displays consistently positive correlation (ρ: 0.632–0.671) with LLM judgments, while statistical evaluation exhibits near-zero or negative correlation. Pairwise accuracy: 60–70% (semantic) vs 30–50% (statistical).

Bootstrap coefficient of variation analysis reveals improved stability for semantic metrics (7.6%) compared to statistical metrics (49.9%), representing 85% improvement in robustness. Complete lexical diversity (TD=1.0) was observed for both methods with zero keyword overlap. Statistical distinctiveness (JSD) was 0.000, while semantic distinctiveness averaged 0.308. Label-based separation metrics indicate moderate alignment: Silhouette=0.055, NMI=0.363, ARI=0.292.

The public dataset validation confirms synthetic dataset findings: semantic metrics demonstrate improved LLM alignment correlation while maintaining enhanced stability across dataset characteristics.

5. Discussion and Limitations

5.1 Key Findings and Implications

Semantic methods achieve enhanced discrimination for coherence (7.62× improvement, production threshold SC>0.7), distinctiveness (1.57× improvement with separation sensitivity), and diversity (consistent discrimination across similarity levels). The three-model LLM ensemble (mean ρ=0.937) reduces individual biases while maintaining strong correlation. These results support semantic methods as preferred evaluation framework for neural topic models.

5.2 Robustness and Limitations

Robustness Analysis. LLM evaluation shows inherent sensitivity to both temperature and prompt formulation, as quantified in Appendix B. Temperature sensitivity analysis (Appendix B.1) reveals score variance across temperature range 0.0-0.7. Prompt variation analysis (Appendix B.2) shows coefficient of variation across multiple prompt formulations. The three-model ensemble mitigates these effects through standardized templates and multi-model consensus, but systematic hallucination evaluation remains an open challenge.

Limitations. First, synthetic datasets derive exclusively from Wikipedia (October 12, 2025), which may not represent informal discourse (social media), specialized corpora (biomedical, legal, financial), or conversational text. Wikipedia's encyclopedic style differs substantially from user-generated content or technical documentation. Second, controlled inter-topic similarity (0.179, 0.312, 0.358) enables systematic evaluation but may not capture natural topic overlap patterns in heterogeneous collections with hierarchical relationships and fuzzy boundaries. Third, English-only evaluation limits multilingual applicability. Fourth, results depend on all-MiniLM-L6-v2 characteristics and may miss aspects captured by domain-specific or larger transformer models.

6. Conclusion

This study developed an evaluation framework integrating statistical, semantic, and LLM-based methods for topic model assessment. Experimental results demonstrate that semantic metrics provide improved discrimination, particularly for coherence evaluation. The three-model LLM ensemble exhibits consistent rank-ordering (Spearman $\rho = 0.914$) with mean absolute difference of 0.102 across evaluations.

Limitations: Current evaluation is limited to English-only assessment due to embedding model selection. Results remain tied to all-MiniLM-L6-v2 characteristics. Synthetic datasets derived exclusively from Wikipedia may not fully represent informal discourse (social media) or domain-specific corpora (biomedical, legal, financial). LLM-based validation carries inherent risks of bias and hallucination despite mitigation strategies.

Future Work: Five directions are proposed for advancing this research:

- **Multilingual Extension:** Semantic metrics will be evaluated using Multilingual BERT (mBERT) or Cross-lingual Language Model (XLM-RoBERTa) to assess performance across diverse languages.
- **Domain-Specific Adaptation:** Embedding models will be tested and adapted for specialized corpora (legal, medical, financial) where vocabulary and topic structures differ significantly from general-domain Wikipedia text.
- **Cost-Efficient LLM Proxies:** To address scalability concerns for large-scale evaluation, we will explore open-source alternatives including Llama-3.1 (70B parameters), Mistral-Large, Qwen-2.5, and distilled lightweight validators such as Phi-3 (3.8B parameters) and Gemma-2 (9B parameters). Self-hosted deployments of these models could eliminate recurring API costs. Additionally, knowledge distillation from ensemble models to smaller specialized evaluators may enhance accessibility for resource-constrained environments. These models could potentially achieve competitive correlation with commercial LLMs while enabling unlimited evaluation at marginal computational cost, though systematic validation remains necessary.
- **Dimensionality Reduction Validation:** Visualization methods (t-SNE, UMAP, PaCMAP) will be systematically compared using trustworthiness and continuity scores to quantify projection-induced artifacts.
- **Expanded Human Validation:** Multi-rater human evaluation with domain experts across diverse fields will be conducted to strengthen generalization claims beyond the current three-expert annotation.

Code and Data Availability: Source code, datasets, experimental results, example documents, validation logs, and semantic metric implementations are available at https://github.com/LeeSeogMin/jips.git under MIT License (code) and CC-BY 4.0 (documentation/data).

Reproducibility seed=42 for all experiments. Numeric consistency verification is documented in `verification/numeric*consistency*check.md`.

Acknowledgment

This work was supported by Hanshin University Research Grant in 2024.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
[2] N. Grootendorst, "BERTopic: Neural topic modeling with BERT," *IEEE Intelligent Systems*, vol. 37, no. 2, pp. 112-120, 2022.
[3] N. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.
[4] A. Angelov, "Top2Vec: Distributed representations of topics," arXiv preprint arXiv:2008.09470, 2020.
[5] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence," in *Proceedings of ACL*, 2021, pp. 759-766.
[6] Y. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," in *ICLR*, 2017.
[7] D. O'Callaghan et al., "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645-5657, 2015.
[8] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *TACL*, vol. 8, pp. 439-453, 2020.
[9] D. Newman et al., "Automatic evaluation of topic coherence," in *HLT-NAACL*, 2010, pp. 100-108.
[10] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *IWCS*, 2013, pp. 13-22.
[11] D. Mimno et al., "Optimizing semantic coherence in topic models," in *EMNLP*, 2011, pp. 262-272.
[12] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *WSDM*, 2015, pp. 399-408.
[13] A. Fang et al., "Using word embedding to evaluate the coherence of topics from Twitter data," in *SIGIR*, 2016, pp. 1057-1060.
[14] A. M. Hoyle, P. Goel, and P. Resnik, "Is automated topic model evaluation broken?: The incoherence of coherence," in *NeurIPS*, 2021.
[15] D. Stammbach et al., "Revisiting automated topic model evaluation with large language models," arXiv:2305.12152, 2023.
[16] H. Rahimi et al., "Contextualized topic coherence metrics," in *EACL Findings*, 2024, pp. 1760-1773.
[17] T. M. Pham, O. Veselovsky, and J. Rousu, "TopicGPT: A prompt-based topic modeling framework," in *NAACL-HLT*, 2024.
[18] C. Meaney, S. Mitra, and S. B. Cohen, "Quality indices for topic model selection and evaluation," *BMC Med Inform Decis Mak*, vol. 23, article 132, 2023.
[19] P. Rüdiger et al., "Topic modeling revisited: A comprehensive analysis," *PLOS ONE*, vol. 17, no. 4, e0266325, 2022.

Appendix A: Metric Calculation Examples

To support understanding and reproducibility, we provide complete worked examples with toy data and runnable code in the GitHub repository. This appendix summarizes the available examples and their key formulas.

A.1 Overview

Four complete examples are available at `github.com/LeeSeogMin/jips/evaluation/examples/`:

- **`semantic_coherence_example.py`** - Semantic Coherence (SC) calculation
- **`semantic_distinctiveness_example.py`** - Semantic Distinctiveness (SD) calculation
- **`semantic_diversity_example.py`** - Semantic Diversity (SemDiv) calculation
- **`llm_aggregation_example.py`** - LLM Score Aggregation
- Detailed step-by-step calculation with toy data (simplified 3D embeddings)
- Each example includes:
- Self-contained runnable Python code (~45-100 lines per example)
- Expected outputs for verification
- Formulas matching Section 3.3.2 exactly

A.2 Key Formulas and Expected Results

Semantic Coherence (SC): $SC = \frac{\sum_{i,j} w_{ij} \cdot h_{ij}}{\sum_{i,j} w_{ij}}$

where h_{ij} *is the pairwise similarity matrix and* $w_{ij} = w_i \cdot w_j$ is the importance weight matrix from PageRank.

Example: Topic ["neural", "network", "learning"] → SC ≈ 0.980

Semantic Distinctiveness (SD): $SD = \frac{1 - \cos(t_i, t_j)}{2}$

where t_i *and* t_j are topic centroid embeddings.

Example: ML vs Automotive topics → SD ≈ 0.171

Semantic Diversity (SemDiv): $SemDiv = \frac{D_{semantic} + D_{distribution}}{2}$

where $D_{semantic}$ *is mean pairwise distinctiveness and* $D_{distribution}$ is normalized entropy.

Example: 3 topics with [4, 3, 5] document assignments → SemDiv ≈ 0.559

LLM Score Aggregation:

Weighted ensemble: 0.35×Claude + 0.40×GPT + 0.25×Grok

Example: Three topics evaluated by three LLMs → [0.928, 0.892, 0.859]

A.3 Usage

```
pip install numpy scikit-learn scipy networkx
```

All examples require standard Python libraries:

```
cd evaluation/examples python semantic_coherence_example.py python
semantic_distinctiveness_example.py python semantic_diversity_example.py
python llm_aggregation_example.py
```

Run individual examples:

```
python run_all_examples.py
```

Or run all examples at once:

A.4 Implementation Notes

- Examples use simplified 3D embeddings for educational purposes
- Actual implementation uses 384-dimensional embeddings (all-MiniLM-L6-v2)

- Full production code is available in `evaluation/NeuralEvaluator.py` and `evaluation/StatEvaluator.py`
- All calculations verified to match mathematical formulas in Section 3.3.2
- Spearman correlation (ρ) is used for LLM agreement rather than Cohen's κ, as our evaluations produce continuous scores (0-1 scale), not categorical labels

Appendix B: Sensitivity and Robustness Analysis

B.1 Temperature Sensitivity Analysis

To address reviewer concerns about LLM evaluation robustness, systematic temperature sensitivity analysis was conducted using Claude-sonnet-4.5 on the distinct dataset. Two temperature values (T=0.0 and T=0.7) were tested to evaluate robustness across operational parameter ranges (5 topics × 4 metrics × 2 temperatures = 40 evaluations).

Table B1: Cross-Temperature Robustness Results

Metric	T=0.0 Mean	T=0.7 Mean	Cross-Temp CV	Within-Temp CV (T=0.0)	Within-Temp CV (T=0.7)	Classification
Coherence	0.944	0.944	0.0%	1.3%	1.3%	VERY LOW
Distinctiveness	0.850	0.850	0.0%	0.0%	0.0%	VERY LOW
Diversity	0.720	0.720	0.0%	0.0%	0.0%	VERY LOW
Semantic Integration	0.850	0.850	0.0%	0.0%	0.0%	VERY LOW
Mean CV (All Metrics)	-	-	0.0%	0.3%	0.3%	VERY LOW

Temperature sensitivity analysis reveals zero cross-temperature variation (CV=0.0%) for all metrics, indicating complete robustness across temperature settings. Within-temperature variation remains minimal (mean CV=0.3%), demonstrating stable evaluation performance.

Cross-Temperature CV: Coefficient of variation of mean scores between T=0.0 and T=0.7 for each metric.

Within-Temperature CV: Coefficient of variation across 5 topics at each temperature setting.

Classification: VERY LOW (<5%), LOW (5-15%), MODERATE (15-25%), HIGH (>25%).

- Coherence displays expected per-topic variation (within-CV=1.3%) across the five sampled topics
- **Within-Temperature Consistency:**
- Multi-topic aggregated metrics (distinctiveness, diversity, integration) exhibit zero within-temperature variation due to single evaluation per temperature
- This pattern confirms LLM is functioning normally while maintaining high parameter robustness

Interpretation: The combination of (1) natural within-temperature variation for per-topic metrics and (2) zero cross-temperature variation demonstrates that LLM evaluation is reliable and reproducible. Score stability across parameter changes supports the suitability of LLM-based assessment for systematic topic quality evaluation.

Recommendation: Use temperature=0.0 for high determinism and reproducibility. However, T=0.7 yields equivalent results, suggesting robustness extends across operational parameter ranges.

B.2 Prompt Variation Sensitivity Analysis

Robustness to prompt formulation was evaluated by testing three prompt variants with varying instruction styles and detail levels. Evaluations were conducted using Claude-sonnet-4.5 at

temperature=0.7 to enable detection of potential prompt-induced variation (5 topics × 3 variants = 15 coherence evaluations).

- **Variant 1 (Baseline)**: Standard coherence evaluation instruction
- **Prompt Variants Tested:**
- **Variant 3 (Detailed)**: Extended instruction with explicit evaluation criteria
- **Variant 5 (Concise)**: Simplified, concise instruction format

Table B2: Prompt Variation Robustness Results

Metric	Variant 1 (Baseline)	Variant 3 (Detailed)	Variant 5 (Concise)	Cross-Prompt CV	Classification
Mean Coherence	0.854	0.854	0.854	0.0%	VERY LOW
Per-Topic CV (Topic 1)	0.0%	0.0%	0.0%	0.0%	-
Per-Topic CV (Topic 2)	0.0%	0.0%	0.0%	0.0%	-
Per-Topic CV (Topic 3)	0.0%	0.0%	0.0%	0.0%	-
Per-Topic CV (Topic 4)	0.0%	0.0%	0.0%	0.0%	-
Per-Topic CV (Topic 5)	0.0%	0.0%	0.0%	0.0%	-

Prompt variation analysis demonstrates high robustness across different prompt formulations. Mean coherence scores remain identical (0.854) across all variants, with zero coefficient of variation indicating complete consistency. Per-topic analysis confirms stable performance across individual topics.

Cross-Prompt CV calculated as variation of scores across three prompt variants for each topic. All five topics showed zero coefficient of variation, indicating identical scores regardless of prompt formulation.

Prompt variation analysis reveals minimal sensitivity to prompt formulation (mean CV=0.0%). Even at elevated temperature (T=0.7), which enhances stochasticity, LLM produces identical coherence scores across baseline, detailed, and concise prompt variants.

Seog-Min Lee https://orcid.org/0009-0009-0754-8523

He received his bachelor's and master's degrees from Seoul National University and earned his Ph.D. in Science and Technology Policy from the same university. He is currently a professor in the Department of Public Policy and Big Data Convergence at Hanshin University. His research interests include big data analytics, artificial intelligence, and causal analysis in AI.