Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Uncertainty-Aware Assessment of LLM-Enhanced Topic Models: An Experton-Based Approach to Interpretability

Eddy Soria

eddy.soria@urv.cat

Universitat Rovira i Virgili

Antonio Moreno

Universitat Rovira i Virgili

Jordi Pascual

Universitat Rovira i Virgili

Ana Beatriz Hernández-Lara

Universitat Rovira i Virgili

Research Article

Keywords: topic modeling, LLM-enhanced topic models, BERTopic, topic interpretability, M-expertons, tourism

Posted Date: July 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10208366/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Uncertainty-Aware Assessment of LLM-Enhanced Topic Models: An Experton-Based Approach to Interpretability

Eddy SORIA[a,1], Antonio MORENO[b], Jordi PASCUAL[b] and Ana Beatriz HERNÁNDEZ-LARA[a]
[a] *Department of Business Management*
Universitat Rovira i Virgili, Catalonia, Spain
[b] *Department of Computer Science and Mathematics*
Universitat Rovira i Virgili, Catalonia, Spain
ORCiD ID: Eddy Soria https://orcid.org/0000-0001-9999-5850, Antonio Moreno https://orcid.org/0000-0003-3945-2314, Jordi Pascual https://orcid.org/0000-0002-7528-5819 and Ana Beatriz Hernández-Lara https://orcid.org/0000-0002-8110-9328

Abstract. This study evaluates traditional, neural, and Large Language Model (LLM)-enhanced topic modeling approaches in a specialized tourism corpus. We compare Latent Semantic Indexing, Non-negative Matrix Factorization, Latent Dirichlet Allocation, Structural Topic Modeling, Neural LDA, and BERTopic, and then examine how LLM-based topic refinement affects topic representation, coherence, diversity, and interpretability. The results show that the LLM-enhanced BERTopic pipeline outperforms the traditional and neural approaches tested, achieving the strongest combined performance in topic coherence and topic diversity. The analysis also shows that decoding temperature influences both topic representation and LLM-based interpretability assessment, indicating that temperature should be treated as a substantive modeling parameter rather than a purely technical generation setting. Beyond model comparison, the study introduces an uncertainty-aware evaluation framework based on the Theory of Expertons, operationalized through M-expertons, to aggregate human and LLM judgments of topic interpretability while preserving uncertainty. Although the empirical application focuses on tourism, the proposed framework is relevant for specialized corpora in which topic interpretation involves semantic ambiguity, evaluator subjectivity, and uncertain boundaries between related themes.

Keywords. topic modeling; LLM-enhanced topic models; BERTopic; topic interpretability; M-expertons; tourism.

1. Introduction

In an era of unprecedented data saturation, extracting actionable insights from vast, heterogeneous document collections has become a critical analytical challenge. Topic modeling addresses this by identifying latent themes, organizing unstructured information, and facilitating the analysis of research patterns, knowledge domains, and emerging trends [1]. These capabilities have made computational text analysis increasingly relevant for synthesizing fragmented evidence and supporting systematic knowledge discovery across disciplines.

[1] Corresponding Author: Eddy Soria Leyva, eddy.soria@urv.cat

Topic modeling has evolved from algebraic and probabilistic approaches, such as Latent Semantic Indexing (LSI), Non-negative Matrix Factorization (NMF), Latent Dirichlet Allocation (LDA), and Structural Topic Modeling (STM), toward neural and transformer-based methods [1, 2]. BERTopic represents a major development in this trajectory because it combines contextual embeddings, dimensionality reduction, clustering, and class-based TF-IDF to generate semantically coherent topic representations [3–5]. Recent studies have further extended this pipeline by incorporating Large Language Models (LLMs). These have been used either as standalone models [6] or through hybrid approaches (e.g., LLM-refined neural models) to improve document embeddings or to improve topic labels and descriptions [7–10].

These models have been applied across a wide range of domains, including computer science, artificial intelligence, and information systems [1, 2], law [4, 5], biomedicine and healthcare [10, 11], education [12, 13], transportation [14], sustainability, energy, urban studies, and environmental sciences [15, 16], business, management, and consumer behavior [17, 18], as well as tourism and hospitality [19].

In particular, tourism provides a useful domain for evaluating these models. This field encompasses a rich diversity of research areas, ranging from sustainability and gastronomy to more specialized domains such as space tourism. As a result, tourism generates heterogeneous textual corpora that combine interdisciplinary terminology with multicultural and multilingual characteristics while addressing a wide variety of applied problems [19, 20].

Traditional topic modeling approaches, such as LDA and STM, continue to dominate tourism research, while state-of-the-art neural and transformer-based models, such as BERTopic, have received comparatively less attention [19]. Against this background, a small number of recent studies have begun to compare these well-established approaches with emerging models in order to provide insights into their performance within the tourism domain.

Kirilenko and Stepchenkova [6] compared LDA, BERTopic, and GPT-4 across tourism corpora varying in coherence, document length, and corpus size, showing that GPT can produce highly interpretable topics but also raises concerns regarding trust and explainability. Similarly, Kamal et al. [21] reviewed and experimentally compared probabilistic, matrix-factorization, and neural embedding-based models on digital tourism datasets, concluding that no single model is universally optimal and that performance depends on corpus characteristics and evaluation criteria.

Despite these advances, three important research gaps remain. First, existing studies have primarily compared standalone topic modeling approaches, leaving the potential of hybrid LLM-enhanced pipelines largely unexplored. Second, although LLMs are increasingly incorporated into topic-modeling workflows, the influence of decoding temperature on topic representations has received little attention. Third, topic interpretability continues to be assessed through deterministic evaluation procedures despite being an inherently subjective and uncertain semantic judgment.

Guided by these gaps, this study addresses the following research questions:

- **RQ1**: How do traditional, neural, and LLM-enhanced topic modeling approaches differ in modeling a specialized tourism corpus?
- **RQ2**: How does temperature influence both topic representation and topic interpretability assessment?
- **RQ3**: How can uncertainty-aware appraisal capture the interpretability of the topics produced by the best-performing model?

This study makes several contributions to the topic modeling literature. First, it provides a comparative benchmark on a specialized tourism corpus, showing that a hybrid LLM-enhanced pipeline outperforms traditional and neural approaches while preserving methodological traceability. Second, it redefines temperature as a key modeling parameter in hybrid LLM-enhanced pipelines by demonstrating its influence on both topic representation and interpretability assessment. Third, it reconceptualizes topic interpretability as a subjective and uncertain appraisal. To the best of our knowledge, this is the first study to introduce an evaluation framework based on the Theory of Expertons for assessing topic interpretability, enabling a unified evaluation across mixed panels of human coders and LLM-as-a-Judge evaluators while preserving different degrees of uncertainty.

2. Background

Topic modeling techniques have evolved from traditional algebraic and probabilistic models toward neural, transformer-based, and LLM-enhanced approaches that incorporate contextual representations and generative language models [1, 22].

2.1. Traditional topic modeling approaches

Traditional topic modeling approaches were developed to identify latent thematic structures from document collections using term-frequency patterns and statistical regularities [23–25]. These methods generally rely on bag-of-words representations, where documents are modeled through word occurrences or document-term structures [25, 26]. They can be grouped into algebraic and probabilistic models.

2.1.1 Algebraic models

Algebraic approaches transform document-term matrices into simplified, meaningful representations, with methods including LSI, NMF, and matrix decomposition variants.

Latent Semantic Indexing. LSI, introduced by [23], was one of the first algebraic approaches designed to overcome the limitations of exact word matching in information retrieval, particularly problems related to synonymy and polysemy. Instead of analyzing documents only through isolated word frequencies, LSI assumes that documents contain an underlying latent semantic structure that can be inferred from global term-document relationships. LSI applies Singular Value Decomposition (SVD) to the document-term matrix (X), generating a reduced semantic representation:

$$X \approx U_k \Sigma_k V_k^T \quad (1)$$

where the decomposition compresses the original matrix into a lower-dimensional latent space that preserves the most relevant semantic patterns while reducing statistical noise. In this latent space, documents with related meanings can appear close to each other even if they do not share the same vocabulary, typically using cosine similarity to measure semantic proximity [23].

A major advantage of LSI is its ability to capture higher-order semantic associations beyond direct word overlap. However, because it relies on linear matrix factorization rather than a probabilistic generative process, LSI has limited capacity to model

contextual semantics and topic uncertainty. These limitations motivated the later development of new approaches such as probabilistic LSI (pLSI) and LDA [27].

Non-negative Matrix Factorization. NMF, proposed by [28], decomposes a non-negative document-term matrix (X) into two lower-dimensional matrices representing document-topic and topic-word relationships:

$$\underset{N \times P\ matrix}{\begin{bmatrix} x_{1,1} & \cdots & \cdots & \cdots & x_{1,P} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N,1} & \cdots & \cdots & \cdots & x_{N,P} \end{bmatrix}} \approx \underset{N \times K\ matrix}{\begin{bmatrix} w_{1,1} & \cdots & w_{1,K} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ w_{N,1} & \cdots & w_{N,K} \end{bmatrix}} * \underset{K \times P\ matrix}{\begin{bmatrix} h_{1,1} & \cdots & \cdots & \cdots & h_{1,P} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{K,1} & \cdots & \cdots & \cdots & h_{K,P} \end{bmatrix}}$$

where (X) represents the original document-term matrix containing word frequencies, (W) is the document-topic matrix describing the relevance of each topic within documents, and (H) is the topic-word matrix representing the importance of words within each topic. The parameter (K) denotes the number of latent topics. Unlike LSI, NMF constrains all matrix values to remain non-negative, producing additive and more interpretable semantic representations [28].

In topic modeling, NMF represents documents as combinations of latent semantic components, where each topic is characterized by highly related terms. This representation often improves interpretability and performs effectively on sparse text corpora, particularly when combined with TF-IDF weighting schemes [29]. However, because NMF relies on linear matrix factorization, it has limited ability to capture contextual semantics and complex word dependencies compared with more recent neural and transformer-based approaches.

2.1.2 Probabilistic models

Probabilistic models implement Bayesian frameworks to model how documents are generated from underlying topics, encompassing methods such as LDA and STM.

Latent Dirichlet Allocation. LDA, introduced by [25], is a probabilistic generative model that represents documents as mixtures of latent topics, where each topic is characterized by a probability distribution over words. Unlike matrix factorization approaches such as LSI or NMF, LDA assumes that documents are generated through a hierarchical Bayesian process in which topics and words are sampled probabilistically. Formally, LDA assumes that each document (d) has a topic distribution (θ_d), while each topic (k) has a word distribution (ϕ_k). The model generates documents according to the following process:

$$\theta_d \sim \mathrm{Dir}(\alpha)$$
$$z_{dn} \sim \mathrm{Multinomial}(\theta_d)$$
$$w_{dn} \sim \mathrm{Multinomial}(\phi_{z_{dn}})$$

where (θ_d) represents the probability distribution of topics within document (d), (z_{dn}) denotes the latent topic assigned to word (n) in document (d), and (w_{dn}) corresponds to the observed word generated from the selected topic. The Dirichlet prior (α) controls the sparsity of topic distributions across documents, while (ϕ_k) defines the probability of words within each topic.

The probability of a document under LDA is obtained by marginalizing over all possible topic assignments and topic mixtures:

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta), p(w_n \mid z_n, \beta) \right) d\theta \qquad (2)$$

where $(p(z_n \mid \theta))$ represents the probability of assigning a topic to a word, and $(p(w_n \mid z_n, \beta))$ denotes the probability of generating a word conditioned on the selected topic.

LDA provides interpretable topic representations and generalizes effectively to unseen documents because topic mixtures are modeled as latent random variables rather than fixed parameters. In addition, the Bayesian structure reduces overfitting compared with pLSI and enables robust topic discovery in large corpora. However, LDA relies on the bag-of-words assumption and therefore ignores word order and contextual semantics. Consequently, it may struggle to capture complex linguistic relationships and contextual dependencies compared with recent neural embedding and transformer-based approaches.

Structural Topic Modeling. STM, developed by [30], extends LDA by incorporating document-level metadata into the topic modeling process. While LDA assumes that topic distributions depend only on latent variables, STM allows observed document characteristics, such as time, author, political orientation, or publication source, to influence both topic prevalence and topic content.

In STM, the topic distribution for each document is modeled through a logistic-normal prior conditioned on document covariates:

$$\theta_d \mid X_d \sim LogisticNormal(X_d\gamma, \Sigma) \qquad (3)$$

where (θ_d) represents the topic proportions of document (d), (X_d) denotes the metadata associated with the document, (γ) contains regression coefficients, and (Σ) models topic covariance structure. This formulation enables the estimation of how metadata variables affect the prevalence of topics across documents.

A major advantage of STM is its ability to reveal relationships between thematic structures and external document attributes while maintaining the probabilistic framework of LDA. Additionally, STM supports hypothesis testing and topic correlation analysis through metadata-aware inference. However, the model increases computational complexity and requires carefully selected metadata, since irrelevant or noisy covariates may negatively affect topic interpretability and stability.

2.2. Neural, transformer-based, and LLM-enhanced topic modeling

Neural, transformer-based, and LLM-enhanced topic modeling approaches extend traditional methods by incorporating distributed and contextual representations of text. Instead of relying only on bag-of-words structures, these models use neural architectures, embeddings, or language models to capture richer semantic relationships between documents and terms [3, 9, 24].

2.2.1 Neural variational inference models (NVIF)

NVIF extend probabilistic topic modeling by combining latent topic representations with deep neural architectures based on variational autoencoders (VAEs). Neural LDA replaces traditional variational inference with neural inference networks and logistic-normal approximations to the Dirichlet prior, enabling scalable and efficient topic learning [31]. Building on this framework, ProdLDA substitutes the mixture assumption

of LDA with a product-of-experts formulation, producing more coherent and interpretable topics through sharper word distributions [31]. More recent extensions such as Neural Labeled LDA (NL-LDA) incorporate supervised or semi-supervised label information within the VAE framework, allowing metadata-guided topic discovery and document classification while preserving efficient black-box inference [32].

2.2.2 Transformer-based and hybrid topic modeling

Transformer-based approaches represent a major shift in topic modeling by replacing bag-of-words representations with contextualized semantic embeddings generated by pre-trained language models such as BERT.

BERTopic, introduced by [3], combines transformer embeddings with density-based clustering and class-based term weighting to generate semantically coherent and interpretable topics. Unlike traditional probabilistic models, contextual embeddings capture semantic similarity and polysemy by encoding words according to their surrounding linguistic context, allowing semantically related documents to be positioned closely in vector space.

BERTopic typically follows a modular pipeline [4] consisting of: (1) contextual embedding generation using transformer models such as Sentence-BERT, (2) dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP), (3) clustering through Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), (4) topic representation using class-based TF-IDF (c-TF-IDF), and (5) optional topic refinement [5].

UMAP reduces the dimensionality of semantic embeddings while preserving both local and global relationships between documents, thereby improving clustering performance in high-dimensional spaces. Subsequently, HDBSCAN identifies dense semantic regions and automatically detects outlier documents without requiring a predefined number of topics, unlike centroid-based approaches such as K-Means. Finally, the c-TF-IDF procedure estimates term importance at the cluster level by emphasizing words that are highly representative of a specific topic while reducing the influence of globally frequent terms, thus improving topic interpretability [3].

This modular architecture allows topic discovery to rely on semantic proximity between documents rather than only on lexical co-occurrence. Egger and Yu [24] emphasized that BERTopic is particularly useful for short and unstructured corpora because it supports semantic search, topic reduction, and flexible topic inspection. Khodeir and Elghannam [12] also found that BERTopic generated more granular topics than traditional models in MOOC forum data, whereas NMF remained competitive when fewer and broader topics were required.

Recent work has extended BERTopic by incorporating LLMs into different stages of the modeling process. One line of research uses LLMs to improve document embeddings. Invernici et al. [7] proposed TETYS, an LLM-based BERTopic pipeline that replaces default SBERT embeddings with LLM-based embeddings and optimizes UMAP and HDBSCAN parameters for large scientific corpora. Their results suggest that LLM-based embeddings can improve topic assignment and topic quality when compared with non-LLM embedding baselines. Similarly, Yang and Kim [8] compared BERTopic with embeddings generated by S-BERT, DistilBERT, Falcon, LLaMA2, and LLaMA3, finding that LLaMA3 achieved the strongest coherence in several datasets, although no single model dominated across coherence, diversity, and significance.

A second line of research uses LLMs after topic discovery to enhance topic interpretation. Aamir et al. [9] combined BERTopic with XLM-R and GPT, using GPT to generate and refine topic labels after clustering. In this design, BERTopic identifies the latent topic structure, while GPT improves the semantic specificity and readability of the resulting labels. Ma et al. [10] similarly compared ChatGPT-assisted LDA and ChatGPT-assisted BERTopic, reporting that BERTopic produced more compact and context-specific clusters, although the comparison did not support an unconditional superiority of BERTopic over traditional models. These findings suggest that hybrid BERTopic-LLM approaches are most defensible when each component performs a distinct role: BERTopic structures the corpus through embedding-based clustering, and the LLM strengthens semantic representation, labeling, and interpretation.

2.3. Topic quality and interpretability

The evaluation of topic quality and interpretability relies on multiple categories of metrics, such as coherence (e.g., NPMI) and topic diversity [4, 5], as well as evaluation techniques such as the word intrusion and topic intrusion tests [22, 33].

Topic modeling methods share core hyperparameters, with the number of topics (K) being the most critical (**Fig.1**), alongside other parameters such as distribution controls, convergence metrics, neural network settings, and temperature for controlling randomness in text generation.

In practice, topics are typically represented as lists of relevant terms ranked by importance (e.g., via probability in LDA/STM or c-TF-IDF scores in BERTopic). LLM-enhanced approaches further refine these representations into more descriptive labels or summaries [22, 26].

Automatic metrics such as coherence, diversity, and significance capture different properties of a topic model, and improvements in one metric do not necessarily imply improvements in the others. Yang and Kim showed that coherence, diversity, and significance involve trade-offs across models and datasets [8].

El-Gayar et al. [22], focusing specifically on human interpretability, found that BERTopic and RoBERTa produced more cohesive and easier-to-label topics than LDA, even though LDA achieved a high rate of labelable topics. This distinction is important because interpretability depends not only on statistical quality, but also on whether evaluators can understand, distinguish, and assign meaningful labels to the topics.

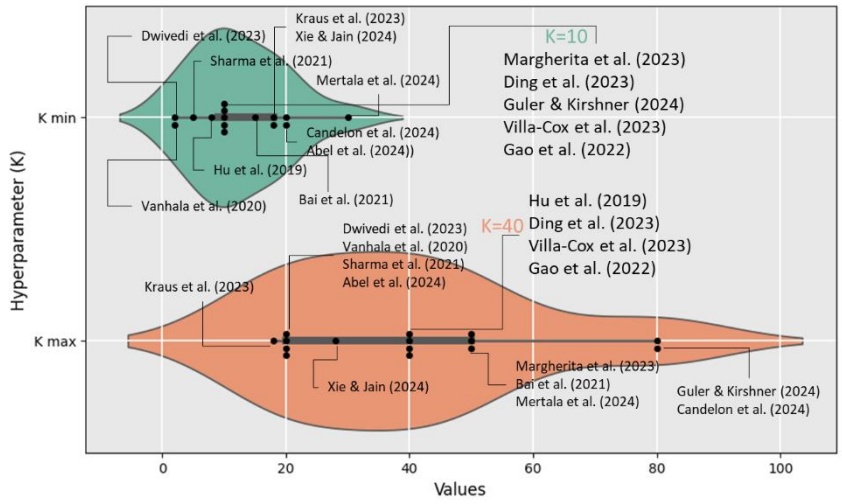


Fig.1. Distribution of K-values explored in recent topic modeling literature (K represents the number of topics). Studies analyzed: [13–18, 34–42].

2.4. Theory of Expertons and counter-appraisement

The aggregation of expert opinions has long been central to decision-making under uncertainty, particularly when the object of evaluation cannot be reduced to a single objective measurement. This problem is especially relevant in topic-model evaluation, where interpretability, semantic coherence, and label adequacy involve subjective appraisal, domain knowledge, and uncertainty.

Although many methods have been used to evaluate the human interpretability of topics [33], to the best of our knowledge, the potential advantages of the Theory of Expertons [43] have not yet been explored in this context. This theory provides a formal framework for representing and aggregating expert judgments while preserving the uncertainty contained in the original evaluations.

An experton is an extension of the concept of probabilistic sets, or random fuzzy sets, in which the probability associated with each α-cut is replaced by an interval of probabilities, or semi-probabilities, obtained from expert opinions through a statistical aggregation procedure [43–45].

Let E be a referential set, finite or not, and let $j = 1, \dots, r$ denote a finite set of experts. For each element $x \in E$, expert j provides a subjective appraisal by means of an interval of confidence: $\forall x \in E: [a_*^j(x), a^{*j}(x)] \subseteq [0,1]$, where $a_*^j(x)$ and $a^{*j}(x)$ denote, respectively, the lower and upper bounds of expert j's appraisal. After applying a statistical aggregation procedure to the lower bounds and upper bounds separately, two cumulative complementary laws are obtained: $F_*(\alpha, x)$ for the lower bounds and $F^*(\alpha, x)$ for the upper bounds. From this process, for each $x \in E$ and each $\alpha \in [0,1]$, the experton is obtained as

$$\tilde{A}(x, \alpha) = [F_*(\alpha, x), F^*(\alpha, x)]. \quad (4)$$

The referential experton is given by

$$\forall x \in E, \ \forall \alpha \in [0,1]: \quad [F_*(\alpha, x), F^*(\alpha, x)] = [1,1] \quad (5)$$

whereas the empty experton is defined as

$$\forall x \in E,\ \forall \alpha \in [0,1]: \qquad [F_*(\alpha, x), F^*(\alpha, x)] = \begin{cases} [1,1], & \alpha = 0, \\ [0,0], & \alpha \neq 0. \end{cases} \quad (6)$$

A central property of expertons is interval monotonicity. By construction, if $\alpha < \alpha'$, then the interval associated with the higher level is included in the interval associated with the lower level: $\forall x \in E,\ \forall \alpha, \alpha' \in [0,1]: (\alpha < \alpha') \Rightarrow [F_*(\alpha', x), F^*(\alpha', x)] \subseteq_i [F_*(\alpha, x), F^*(\alpha, x)]$, where $\subseteq_i$ denotes interval inclusion. Equivalently, $(\alpha < \alpha') \Rightarrow F_*(\alpha', x) \leq F_*(\alpha, x)$ and $F^*(\alpha', x) \leq F^*(\alpha, x)$.

This monotonic structure distinguishes the experton from a simple average of expert opinions. Rather than collapsing appraisals into a single central value, the experton preserves the cumulative distribution of the experts' confidence intervals across the evaluation scale, thereby retaining information about both central tendency and uncertainty.

Kaufmann also defined several algebraic operations over expertons, including complement, product, algebraic sum, minimum, and maximum [43, 44]. These operations allow expertons to be compared and combined while preserving their interval-valued structure. As a result, expertons form a coherent mathematical framework for fuzzy aggregation, supporting properties such as commutativity, associativity, involution under complement, distributivity under minimum and maximum, and De Morgan-type relations [43].

A further extension of the classical experton is the *M*-experton. In this generalization, the confidence interval is replaced by richer confidence structures, such as triplets or quadruplets. Thus, experts may express not only a lower and an upper bound, but also one or more internal values representing the zone of maximum presumption. In the quadruplet case, an appraisal can be represented as $(a_L, a_{C_1}, a_{C_2}, a_U)$; $0 \leq a_L \leq a_{C_1} \leq a_{C_2} \leq a_U \leq 1$, where a_L and a_U define the support interval, and a_{C_1} and a_{C_2} define the interval of maximum presumption. *M*-expertons therefore extend the classical experton while preserving its cumulative aggregation logic.

A topic may be broadly interpretable while still having uncertain semantic boundaries. Similarly, a topic label may be adequate but not perfectly exhaustive. *M*-expertons provide a way to preserve this uncertainty instead of forcing evaluators to produce a single crisp score.

Beyond aggregation, experton-based appraisal can also be used as a validation procedure through counter-appraisement. Counter-appraisement refers to the construction of an additional appraisal, obtained either through specific transformation procedures or through one or several counter-experts, in order to contrast the initial expert evaluation [44, 46]. In practical terms, an independent group of experts evaluates the same object without relying on the results of the first panel, and the resulting expertons are then compared. If both expertons lead to similar results, the initial appraisal can be regarded as more reliable; if they diverge substantially, the appraisal is weakened because the two groups do not converge on the same evaluation.

2.5. Research gaps and study positioning

Comparative studies in tourism indicate that topic model performance depends on corpus characteristics, including document length, corpus size, and textual noise. As a result, performance varies across coherence, diversity, and interpretability, suggesting that these metrics should be considered jointly when selecting the most appropriate model for a specific analytical objective.

Kirilenko and Stepchenkova [6] showed that LDA extracts detailed topics from large corpora of long, coherent documents but identifies topics for only about 30% of short, noisy social media posts, often producing semantically overlapping topics. BERTopic exhibits the opposite profile, performing well on short texts, including noisy social media posts, yet assigning interpretable topics to only about 9% of long reviews and showing limited performance on small corpora and long documents. Finally, these authors found that GPT-4 generated the most interpretable topics across different corpus types, although it offers limited transparency because it produces final topics without exposing the underlying derivation process [6].

Other studies show that NMF generally achieves higher topic coherence than LDA, producing better-separated and more stable topics, particularly on small-to-medium datasets [21, 24]. Moreover, BERTopic's main strength appears to lie in topic diversity rather than coherence, as NMF typically achieves higher coherence, whereas the variability introduced by BERTopic's clustering process reduces its reliability on smaller datasets [21, 47].

Despite these advances, three important research gaps remain. First, existing comparisons primarily evaluate standalone models, whether traditional (e.g., LDA), transformer-based neural models (e.g., BERTopic), or pure LLMs [6, 48], whereas much less is known about the performance of hybrid pipelines. In a hybrid design, models such as BERTopic retain responsibility for document embeddings, dimensionality reduction, and clustering, while an LLM refines topic labels and keywords. This gap is important because direct extraction based solely on LLMs such as ChatGPT may improve fluency and interpretability, but it also reduces methodological traceability [6], whereas a hybrid pipeline keeps topic discovery and topic representation analytically separable. It therefore remains unclear whether hybrid pipelines can better combine the strengths of these different approaches.

Second, existing comparative studies in tourism evaluate different model architectures [21] and, in some cases, explicitly acknowledge the stochastic nature of neural and embedding-based models [24]. However, they have not yet systematically examined how LLM decoding parameters, such as temperature, influence topic representations. Once LLMs are incorporated into the modeling pipeline, the generated topic labels and keywords become sensitive to temperature, yet its impact on the performance of topic modeling remains largely unknown.

Third, topic interpretability is inherently a subjective and imprecise assessment, as it depends on domain knowledge and on where each evaluator draws a topic's semantic boundaries. However, it is typically reduced to a single deterministic score, whether through an automatic coherence metric [21] or a binary correct/incorrect judgment [49]. Such crisp representations disregard the subjectivity of the assessment, as well as the disagreement and varying degrees of confidence that arise when different evaluators, whether human or LLM-based, interpret the same topic.

This study addresses these gaps by comparing traditional, neural, and LLM-enhanced topic modeling approaches on a specialized tourism corpus. Specifically, it quantifies performance differences across models (**RQ1**), introduces decoding temperature as a controllable parameter within a hybrid pipeline (**RQ2**), and reconceptualizes topic interpretability as a subjective and imprecise construct. To this end, it proposes an evaluation framework based on the Theory of Expertons, which enables the appropriate aggregation of the uncertainty inherent in evaluations provided by both human assessors and LLMs (**RQ3**). In doing so, the study provides practical,

empirically grounded guidance for selecting topic modeling methods in specialized tourism corpora.

3. Method

Several computational experiments were conducted to evaluate different algebraic, probabilistic, and neural topic modeling approaches, ultimately selecting the best-performing one.

This study utilized the TOURSKAD dataset [50], a corpus of scientific articles in the field of tourism, to evaluate the performance of topic modeling. Tourism, in particular, is a domain characterized by specialized terminology, in which the assessment of topic modeling performance has been comparatively less explored. This corpus underwent a preprocessing pipeline using spaCy. The process included noise removal through regular expressions (e.g., eliminating non-English characters, etc.), text normalization via lowercase conversion, lemmatization, and stop words removal.

A total of 138 distinct models were generated and evaluated through two rounds of experiments to optimize the cost-performance balance. All models underwent systematic evaluation across a range of topics (K) from 10 to 100 (in increments of 5). The performance of the models was assessed using three metrics, implemented through specialized Python libraries (OCTIS, scikit-learn, sentence-transformers and NLTK).

First, training time was measured to assess computational efficiency and scalability. This metric corresponds to the elapsed time required to fit each model under a given number of topics (K), and it was used to evaluate the practical feasibility of each approach across different topic configurations.

Second, topic coherence was evaluated using Normalized Pointwise Mutual Information (NPMI), computed through *Gensim's c_npmi* coherence implementation. NPMI measures the semantic relatedness of the most representative words within each topic based on their co-occurrence patterns in the reference corpus. In this study, coherence was computed over the top (n=10) words of each topic, excluding BERTopic's outlier topic (-1). Let ($W_k = w_1, \dots, w_n$) denote the set of the top (n) words associated with topic (k). For each pair of words (w_i) and (w_j) in (W_k), NPMI is defined as:

$$NPMI(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \qquad (7)$$

where ($P(w_i)$) and ($P(w_j)$) denote the probabilities of observing words (w_i) and (w_j), respectively, and ($P(w_i, w_j)$) denotes the probability that both words co-occur within a context window in the reference corpus. In the implementation, co-occurrence probabilities were estimated using Gensim's default boolean sliding window for *c_npmi*. The topic-level coherence score is obtained by averaging the pairwise NPMI values over all combinations of top words within each topic:

$$C_{NPMI}(k) = \frac{2}{n(n-1)} \sum_{1<i<j<n} NPMI(w_i, w_j) \qquad (8)$$

The final model-level coherence score is computed as the arithmetic mean of ($C_{NPMI}(k)$) across all (K) non-outlier topics. Higher NPMI values indicate that the top words of a topic tend to co-occur more frequently than expected by chance, suggesting greater intra-topic semantic coherence.

Third, topic diversity was employed to assess inter-topic quality by measuring the degree of lexical overlap among topics. Let (W^*) denote the set of unique words obtained from the top (n=10) words across all (K) non-outlier topics. Topic diversity is defined as:

$$TD = \frac{|W^*|}{nK} \qquad (9)$$

where ($|W^*|$) is the number of unique words across all topic representations, (n) is the number of top words considered per topic, and (K) is the number of non-outlier topics. A value close to 1 indicates that topics share few or no repeated words, suggesting greater diversity and lower redundancy among topics. Conversely, lower values indicate higher lexical overlap between topics.

While NPMI captures intra-topic coherence, topic diversity captures inter-topic distinctiveness; therefore, both metrics were jointly considered when selecting the optimal model configuration [11]. The estimation of the optimal number of topics for each model was based on maximizing the combined value of NPMI coherence and topic diversity metrics. The experiments were conducted on a workstation equipped with an Intel processor (~3.7 GHz), 64 GB of physical memory, and an NVIDIA GeForce RTX 2080 Ti graphics card with 11 GB VRAM, using CUDA version 12.7.

3.1. Experimental round 1: Baseline model comparison and selection

Baseline. In the first round, 19 combinations of the hyperparameter K were evaluated across six baseline models, consisting of two algebraic models (NMF, LSI), two probabilistic models (LDA, STM), and two neural models (Neural LDA, BERTopic). This selection of models ensured a varied exploration of different approaches to topic modeling (114 models in total).

Implementation Details. The algebraic models, NMF and LSI, were implemented using the OCTIS framework [51], leveraging their respective strengths in matrix decomposition and dimensionality reduction for topic identification. The probabilistic models included the traditional LDA, implemented through OCTIS with a symmetric Dirichlet prior (alpha=0.1), and the STM, implemented in R with temporal covariates to capture the evolution of topics over time. For each value of K, we fitted the STM model using spectral initialization [30] to ensure stable and reproducible results.

For the neural approaches, Neural LDA was implemented using the OCTIS framework with a three-layer architecture using softplus activation and dropout regularization, while BERTopic was implemented using the BERTopic library with default parameters, except for the number of topics which was manually controlled during model initialization. This manual topic number specification ensured methodological consistency with other models and enabled direct comparison across different topic number configurations.

3.2. Experimental round 2: Fine-tuning the selected model (BERTopic)

Building upon the BERTopic pipeline described in Section 2, our second experimental phase focused on enhancing performance by fine-tuning two critical components: a custom embedding generation model (step 1) and the final topic representations using LLMs (step 6).

Fine-tuning a custom encoder for document embeddings. For the custom encoder, we fine-tuned ModernBERT [52] on TOURSKAD dataset, leveraging its architectural

advantages including rotary positional embeddings and local-global alternating attention. The model was initialized from the answerdotai/ModernBERT-base checkpoint (80/20 split, 3 epochs, learning rate 2e-5, 15% masking) to generate dataset-specific embeddings, crucial for BERTopic's performance. The resulting model, designated as MoST-BERT, served as the embedding generator.

Fine-tuning of topic representations through LLMs. To refine the topic representations, we evaluated eight LLMs (Claude-3.5-Sonnet, DeepSeek-V3, Gemini-1.5-pro, Gemini-2.0-flash, GPT-4o, Gemma2-9b, Llama-3.3-70b, and Mixtral-8x7b). Each model was tested across three temperature settings (T=0.3, 1.0, 1.5), except for Claude-3.5-Sonnet (tested at T=0.3, 0.5, and 0.7 due to API constraints), resulting in 24 distinct LLM configurations. To orchestrate these interactions, we leveraged the structured input/output capabilities of each LLM's API. BERTopic, using custom MoST-BERT embeddings, first identified document clusters via HDBSCAN, extracted key terms with their c-TF-IDF scores, and selected representative documents for each topic. Afterward, this information was passed to each LLM as a standardized JSON schema. The models were then prompted to return a structured JSON output containing an enhanced topic representation, including: a concise, descriptive label and a refined list of keywords, both aimed at improving topic interpretability. Document processing was managed in batches, with automatic content reduction using tiktoken.

3.3. Multi-method assessment of topic interpretability

Topic interpretability refers to the extent to which a topic's underlying semantic theme can be accurately inferred from its representative terms by evaluators. Although NPMI coherence and topic diversity help discriminate interpretability [11, 53], this study conducted two additional evaluations involving the use of LLM judges and human coders (the authors) as a reference benchmark. First, word intrusion tests [33] were employed and implemented in two different phases. Subsequently, a novel technique based on the General Theory of Expertons [43] was proposed, specifically designed to aggregate interpretability judgments from multiple LLM evaluators.

3.3.1. Indirect approach: Word-intrusion evaluation

The word intrusion task measures topic interpretability indirectly by computing the proportion of evaluators who correctly identify an intruder word embedded within a topic representation [53]. Following [54], evaluators are presented with a set of representative topic terms together with a semantically unrelated distractor. Topics whose terms form a coherent semantic grouping are expected to yield higher intrusion detection rates, whereas less coherent topics make the intruder more difficult to identify.

The evaluation was structured into two phases. In the first phase, a panel of five LLM judges was employed to assess whether the intruder word could be consistently identified when both the intruder configuration and its semantic difficulty varied. The panel included compact and open-weight models (Panel A): *Llama 3.1 8B, Llama 3.3 70B, Llama 4 Scout 17B, gpt-oss-120b, and gpt-oss-20b*.

All models received the same stimulus bank and were asked to identify the intruder word in each item. This design allowed the examination of whether patterns were consistent across topics, rounds, and the two intruder-difficulty conditions. To this end, the stimulus bank was constructed with two conditions. **Table 1** presents illustrative examples of the intruder selection procedure.

Table 1. Illustrative examples of topic-specific intruder selection

Topic	Representation	Low-similarity random intruders	High-similarity random intruders	Random canonical intruders
T0	digital; guide; digital literacy; tour; information	wine; food; medical…	financial; coastal; vector…	language
T4	ecotourism; agritourism; farm; rural; community	vector; transmit; wine…	heritage; volunteer; food…	africa
T15	reality; virtual reality; vr; virtual; ar	heritage; wine; financial…	simulation; physical; airbnb…	forecast
T25	vector; virus; specie; transmission; epidemiology	migrant; language; internship…	transmit; big; digital…	entrepreneurship
T29	ontology; semantic; semantic web; natural language; query	entrepreneurship; event management; airbnb...	language; digital; robot…	volunteer

In the low-similarity random intruders condition (soft bank), the intruder was pseudo-randomly selected, using a fixed global seed, from the top words of another topic different from the target topic. This condition generated distractors that were relatively external to the semantic field of the evaluated topic. In the high-similarity random intruders condition, the intruder also came from another topic, but the source topic was selected from among the topics semantically closest to the target topic, making the distractor more plausible and therefore theoretically more difficult to detect.

To operationalize the semantic difficulty, the representation of each topic generated by BERTopic was encoded using the SentenceTransformers *all-MiniLM-L6-v2* embedding model [55]. Based on these vectors, cosine similarity (CS) was computed between pairs of topics:

$$CS(u,v) = \frac{\sum_{n=1}^{k} u_n v_n}{\sqrt{\sum_{n=1}^{k} u_n^2}\sqrt{\sum_{n=1}^{k} v_n^2}} \qquad (10)$$

where u and v are the embeddings of two topics, and k is the dimensionality of the vector space produced by *all-MiniLM-L6-v2*.

Higher values of CS indicate greater semantic proximity between topics. For each target topic, the hard bank selected intruders from the topics with the highest cosine similarity, rotating across the three nearest semantic neighbors in successive rounds. All random selections and option orderings were generated deterministically using a fixed *global_seed* = 42, combined with the bank, topic, and round identifiers, ensuring that the stimulus bank could be exactly reproduced. Each topic was evaluated across ten rounds per condition, with five words belonging to the target topic and one intruder word whose origin varied according to the bank [33, 54]. This structure made it possible to estimate not only the overall precision of the LLM judges, but also their robustness to low-similarity random intruders and semantically close intruders.

In the second phase, a human-aligned LLM-as-a-judge protocol was implemented using a single canonical word-intrusion item for each of the 30 topics, allowing the effects of different decoding temperatures on LLM judgments to be systematically examined.

Fifteen LLMs were used as judges and organized into three independent panels (**Table 2**). Panel A consisted of the five open-weight reference models used in Phase 1, enabling cross-phase comparison. Panel B comprised five OpenAI proprietary models (GPT-family), and Panel C included five models from alternative providers and model families. This design reduced dependence on any single model family or provider and follows recent evidence suggesting that diverse LLM juries can produce more stable

judgments than a single LLM judge [56]. The optimal model selected in the previous stage, Mixtral-8x7B, was excluded from all judging panels to avoid self-preference bias, a known limitation of LLM-based evaluation in which models may favor their own generations or outputs produced by related systems [57–59].

Each LLM judge evaluated the canonical word-intrusion items at three decoding temperatures: 0.3, 1.0, and 1.5. For each response, the selected word was compared against the canonical intruder, and invalid or unparsable responses were excluded from the corresponding calculations. By computing the precision of the models, Phase 2 results were summarized at the model, panel, temperature, and topic levels, with the human coders serving as an independent benchmark.

Table 2. Composition and technical characteristics of the LLM judges

Panel	Models	Developer / provider	Access type	Parameters	Context window	API price*, USD / 1M tokens
A	Llama 3.1 8B Instant	Meta / Groq	Open-weight	8B	131k	0.05 in / 0.08 out
A	Llama 3.3 70B Versatile	Meta / Groq	Open-weight	70B	131k	0.59 in / 0.79 out
A	Llama 4 Scout 17B-16E	Meta / Groq	Open-weight	17B, 16 experts	131k	0.11 in / 0.34 out
A	GPT-OSS 120B	OpenAI / Groq	Open-weight	120B MoE	131k	0.15 in / 0.60 out
A	GPT-OSS 20B	OpenAI / Groq	Open-weight	20B MoE	131k	0.075 in / 0.30 out
B	GPT-5.5	OpenAI	Proprietary	Not disclosed	1,050k	5.00 in / 30.00 out
B	GPT-5.2	OpenAI	Proprietary	Not disclosed	400k	1.75 in / 14.00 out
B	GPT-5.1	OpenAI	Proprietary	Not disclosed	400k	1.25 in / 10.00 out
B	GPT-5.4 mini	OpenAI	Proprietary	Not disclosed	400k	0.75 in / 4.50 out
B	GPT-5.4 nano	OpenAI	Proprietary	Not disclosed	400k	0.20 in / 1.25 out
C	DeepSeek-V4-Pro	DeepSeek	Open-weight	1.6T, 49B active	1,000k	0.435 in / 0.87 out
C	DeepSeek-V4-Flash	DeepSeek	Open-weight	284B, 13B active	1,000k	0.14 in / 0.28 out
C	Gemma 4 31B	Google DeepMind	Open-weight	31B	256k	-
C	Gemini 3.1 Flash-Lite	Google DeepMind	Proprietary	Not disclosed	1,000k	0.13 in / 0.75 out
C	Qwen3-32B	Alibaba Cloud / Groq	Open-weight	32B	131k	0.29 in / 0.59 out

**API prices and context windows correspond to the values available at the time of data collection. MoE = Mixture of Experts. B = billion parameters; T = trillion parameters. "Active" indicates the approximate number of parameters used per inference in MoE models.*

3.3.2. Proposed approach: Direct fuzzy appraisal through M-expertons

We introduced a direct fuzzy-evaluation procedure based on the Theory of Expertons [43]. Specifically, topic interpretability was operationalized as an uncertain appraisal problem in which human coders and LLM judges acted as evaluators, while *M*-expertons provided the aggregation framework. This approach was designed not only to obtain a representative interpretability value for each topic, but also to preserve the uncertainty structure contained in the evaluators' judgments.

Let $X = x_1, \dots, x_m$ denote the set of topics generated by the optimal topic model, and let $J = 1, \dots, n$ denote the evaluators in a given appraisal panel. Evaluators may correspond to human coders or LLM judges. Confidence judgments were expressed on Kaufmann's endecadary scale, defined as the ordered grid $\mathcal{A} = \{0, 0.1, 0.2, \dots, 0.9, 1.0.\}$.

The direct appraisal task was implemented through the quadruplet case of the *M*-experton. For each topic $x \in X$, evaluator $j \in J$ provided an ordered quadruplet $q_j(x) = \left(a_{L,j}(x), a_{C1,j}(x), a_{C2,j}(x), a_{U,j}(x)\right)$, such that $0 \leq a_{L,j}(x) \leq a_{C1,j}(x) \leq a_{C2,j}(x) \leq a_{U,j}(x) \leq 1$.

Here, $a_{L,j}(x)$ and $a_{U,j}(x)$ define the support interval of evaluator (j)'s judgment, whereas $a_{C1,j}(x)$ and $a_{C2,j}(x)$ define the interval of maximum presumption. Substantively, the support interval represents the plausible range of the evaluator's

appraisal, while the core interval represents the range in which the evaluator's confidence is strongest. For any component $r \in \mathcal{R} = \{L, C1, C2, U\}$, and any confidence level $\alpha \in \mathcal{A}$, the cumulative complementary proportion was computed as:

$$M_r(\alpha, x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left(a_{r,j}(x) \geq \alpha\right) \qquad (11)$$

where $(\mathbf{1}(\cdot))$ is the indicator function, which takes value 1 when the condition is satisfied and 0 otherwise. Therefore, for each topic (x), the *M*-experton was defined as the mapping $\widetilde{M}: X \times \mathcal{A} \to [0,1]^4$, given by:

$$\widetilde{M}(x, \alpha) = [M_L(\alpha, x), M_{C1}(\alpha, x), M_{C2}(\alpha, x), M_U(\alpha, x)] \qquad (12)$$

By construction, the following ordering condition holds for every $x \in X$ and $\alpha \in \mathcal{A}$:

$$0 \leq M_L(\alpha, x) \leq M_{C1}(\alpha, x) \leq M_{C2}(\alpha, x) \leq M_U(\alpha, x) \leq 1.$$

Moreover, each component $M_r(\alpha, x)$ is non-increasing with respect to α, so that for $\alpha < \alpha',\ M_r(\alpha', x) \leq M_r(\alpha, x), \qquad r \in \mathcal{R}$.

M-expertons were employed to quantify topic interpretability, defined as the extent to which a topic could be meaningfully understood from its representative words and documents. Human coders completed their assessments before the results from the LLM experiments were obtained, ensuring that the human benchmark remained independent of LLM outputs. The LLM assessments were subsequently performed across the predefined panels and decoding temperatures (T=0.3, T=1.0, T=1.5).

This formulation treats both human and LLM outputs as structured semantic appraisals. Such appraisals need not be deterministic, since they may reflect ambiguity, variability, and uncertainty in the interpretation of topic meaning. *M*-expertons therefore provide a suitable framework for representing, aggregating, and comparing human and LLM judgments while preserving the uncertainty associated with their evaluations.

3.3.3. Measurement and reliability metrics

For the word-intrusion task, topic interpretability was quantified using Model Precision (MP), defined as the proportion of judges that correctly identified the injected intruder word [53, 54]. Let (ω_t) denote the true intruder associated with stimulus *t*, and let $\widehat{\omega}_{t,j}$ represent the intruder selected by judge *j*. Model Precision was computed as:

$$MP_t = \frac{1}{J}\sum_{j=1}^{J} \mathbf{1}\left(\widehat{\omega}_{t,j} = \omega_t\right) \qquad (13)$$

where *J* is the total number of judges in the corresponding evaluator group and $\mathbf{1}(\cdot)$ is the indicator function. Higher MP_t values indicate that a larger proportion of judges identified the correct intruder and, therefore, that the topic was more interpretable under the intrusion criterion.

Inter-rater reliability was assessed using Krippendorff's alpha with the nominal metric, which is appropriate for unordered categorical responses. Krippendorff's alpha [60] was defined as:

$$\alpha = 1 - \frac{D_o}{D_e} \qquad (14)$$

where (D_o) denotes the observed disagreement among evaluators and (D_e) denotes the disagreement expected by chance. Because intrusion-test responses corresponded to unordered categorical selections, the nominal disagreement function was used:

$$\delta^2(c,k) = \begin{cases} 0, & \text{if } c = k \\ 1, & \text{if} c \neq k \end{cases}$$

Using Krippendorff's coincidence-matrix formulation, observed disagreement was computed as:

$$D_o = \frac{\sum_c \sum_k o_{ck}\, \delta^2(c,k)}{\sum_c \sum_k o_{ck}}, \qquad (15)$$

and expected disagreement was computed as:

$$D_e = \frac{\sum_c \sum_k n_c\, n_k \delta^2(c,k)}{n(n-1)} \qquad (16)$$

Here, (o_{ck}) denotes the observed coincidence between categories (c) and (k) in Krippendorff's coincidence matrix, (n_c) and (n_k) denote the corresponding marginal category totals, and (n) denotes the total number of valid pairable assignments. For each subgroup, a topic-by-evaluator response matrix was constructed, with topics as units of analysis and evaluators as raters. Krippendorff's alpha was estimated separately for the human panel, each LLM panel, all LLMs combined, and the combined human-LLM panel. Uncertainty was quantified using nonparametric bootstrap resampling at the topic level. Topics were resampled with replacement 1,000 times, Krippendorff's alpha was recomputed for each resample, and 95% percentile confidence intervals were obtained from the empirical bootstrap distribution.

Pairwise similarity between evaluators was assessed using the Simple Matching Coefficient (SMC). For two evaluators *i* and *j*, SMC was computed as the proportion of shared topics for which both evaluators selected the same intruder word:

$$SMC(i,j) = \frac{1}{|T_{ij}|} \sum_{t \in T_{ij}} \mathbf{1}\left(choice_{i,t} = choice_{j,t}\right) \qquad (17)$$

where T_{ij} denotes the set of topics answered by both evaluators and $\mathbf{1}(\cdot)$ is the indicator function. SMC matrices were computed separately for each LLM temperature level. In each matrix, the raters included the LLMs evaluated at that temperature together with the human coders, whose responses were temperature-independent.

3.3.4. M-expertons and counter-appraisement

For the direct fuzzy-appraisal procedure, metrics derived from the Theory of Expertons were computed from the *M*-expertons defined above. Separate *M*-expertons were constructed for the human coders, Panel A, Panel B, Panel C, and the full LLM panel.

We implemented a counter-appraisement procedure [44]. Panel A was treated as the primary LLM expert panel, while Panels B and C were used as independent counter-expert panels. Panel B represented proprietary GPT-family models, whereas Panel C represented a cross-provider panel. The human coders were retained as an external benchmark.

For each *M*-experton, the component-wise mathematical expectation was computed over the non-zero confidence levels. Let $\mathcal{A}^+ = \{0.1, 0.2, \ldots, 1.0\}$ denote the set of non-zero confidence levels and let $\mathcal{R} = \{L, C_1, C_2, U\}$ denote the four components of the quadruplet *M*-experton. For each component $r \in \mathcal{R}$, the expectation was computed as:

$$E_r(\widetilde{M}) = \frac{1}{|\mathcal{A}^+|} \sum_{\alpha \in \mathcal{A}^+} M_r(\alpha) \qquad (18)$$

This produced the expected quadruplet:

$$E(\widetilde{M}) = \left(E_L(\widetilde{M}), E_{C_1}(\widetilde{M}), E_{C_2}(\widetilde{M}), E_U(\widetilde{M})\right) \qquad (19)$$

Distances between *M*-expertons were computed in two complementary ways. First, expectation distance compared the expected quadruplets. Given two *M*-expertons $\widetilde{M}_A$ and $\widetilde{M}_B$, it was computed as:

$$d_E(\widetilde{M}_A, \widetilde{M}_B) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left| E_r(\widetilde{M}_A) - E_r(\widetilde{M}_B) \right| \qquad (20)$$

Second, full-law Hamming distance compared the complete cumulative *M*-experton structures across all non-zero confidence levels:

$$d_H(\widetilde{M}_A, \widetilde{M}_B) = \frac{1}{|\mathcal{A}^+|\,|\mathcal{R}|} \sum_{\alpha \in \mathcal{A}^+} \sum_{r \in \mathcal{R}} \left| M_{A,r}(\alpha) - M_{B,r}(\alpha) \right| \qquad (21)$$

In the endecadary quadruplet case used in this study, $|\mathcal{A}^+| = 10$ and $|\mathcal{R}| = 4$, so the operational full-law Hamming distance becomes:

$$d_H(\widetilde{M}_A, \widetilde{M}_B) = \frac{1}{40} \sum_{\alpha \in \{0.1, \ldots, 1.0\}} \sum_{r \in \{L, C_1, C_2, U\}} \left| M_{A,r}(\alpha) - M_{B,r}(\alpha) \right| \qquad (22)$$

The expectation distance compares the reduced expected quadruplets, whereas the full-law Hamming distance compares the complete cumulative *M*-experton structures. Formal definitions and metric properties of these distances are provided in **Appendix A**.

Small distances indicate stronger convergence between appraisal and counter-appraisal panels. Following the counter-appraisement criterion established in the experton literature, distances close to or below 5% were interpreted as strong convergence, whereas larger distances indicated weaker convergence between appraisal groups.

To summarize the expected quadruplet into a single representative value, we used Kaufmann's representative number for a confidence quadruplet:

$$K^* = \frac{E_L + 2(E_{C1} + E_{C2}) + E_U}{6} \qquad (23)$$

This scalar gives double weight to the interval of maximum presumption while retaining the contribution of the support bounds. Higher K^* values indicate higher expected topic interpretability.

Uncertainty was summarized through the width of the expected *M*-experton. The support width was defined as:

$$W_S(\widetilde{M}(x)) = E_U - E_L \qquad (24)$$

and the core width as:

$$W_C(\widetilde{M}(x)) = E_{C2} - E_{C1} \qquad (25)$$

Larger support widths indicate greater imprecision in the aggregated judgment, whereas smaller widths indicate more concentrated appraisals. These measures allowed us to compare not only whether humans and LLMs assigned similar expected interpretability values, but also whether they expressed similar levels of uncertainty.

4. Findings

4.1. *Round 1: Baseline model comparison and selection*

The comparative analysis of six topic modeling approaches (STM, LDA, LSI, NMF, Neural LDA, and BERTopic) revealed distinct performance patterns across computational efficiency (**Fig.2**) and topic quality metrics (**Fig.3**).

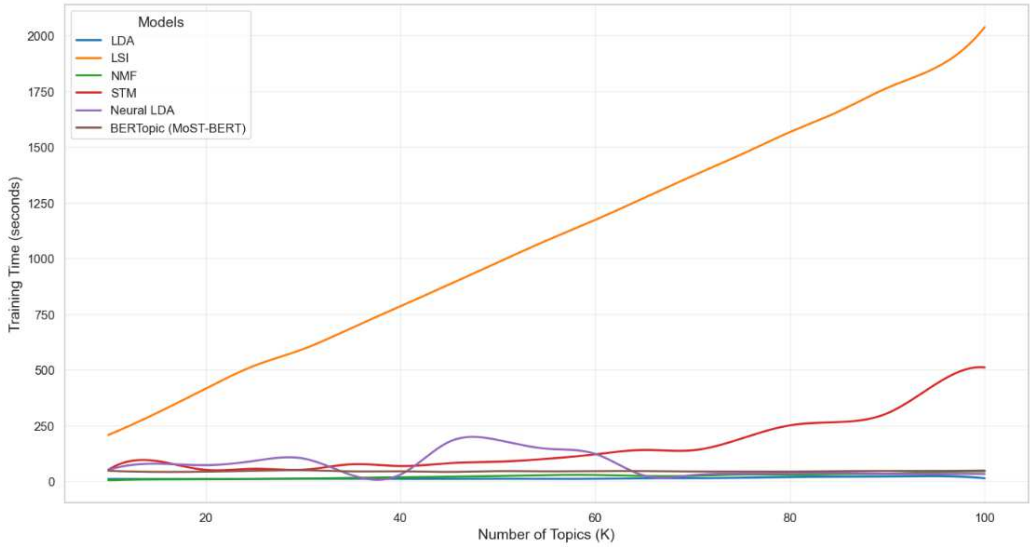


Fig. 2. Computational efficiency across different numbers of topics (K).

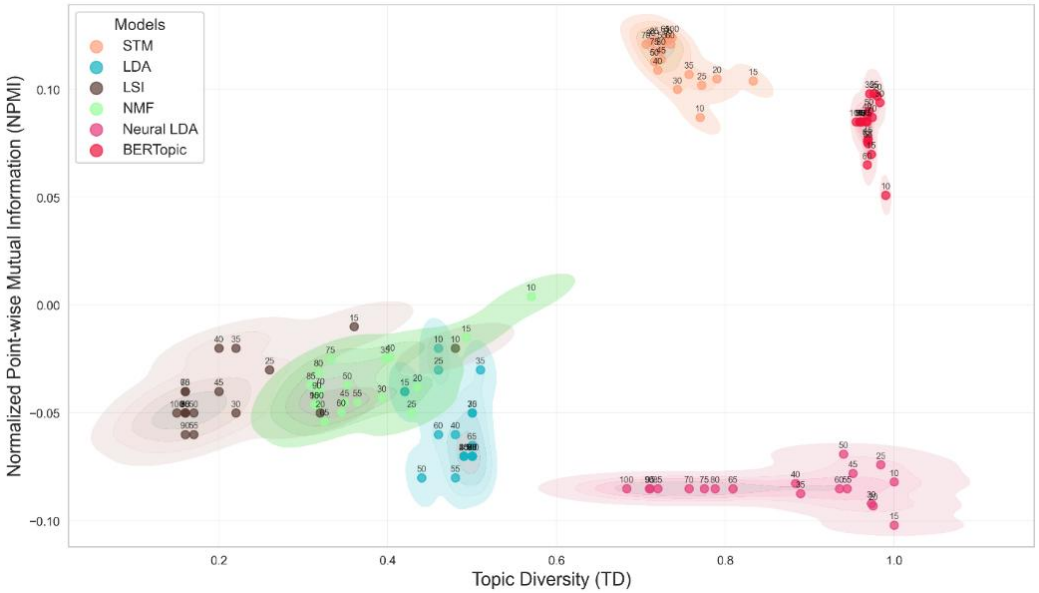


Fig. 3. NPMI-TD performance distribution of topic models. Numbers beside data points indicate the number of topics (K) for each model.

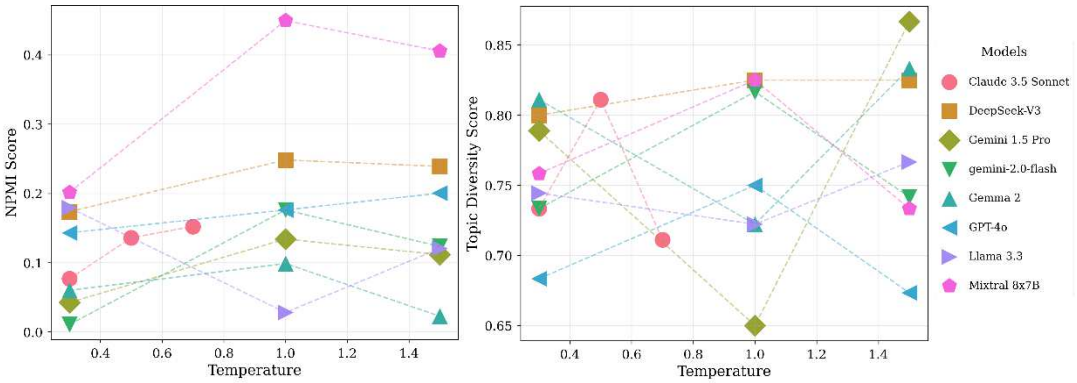


Fig. 4. Performance Comparison of Multiple Representations for BERTopic (K=30) across different temperature values (T).

Training time analysis (**Fig. 2**) showed that LSI exhibited a steady linear increase in computational cost as the number of topics (K) increased, while LDA, NMF, and BERTopic maintained relatively stable training times across the K range (10–100 topics). STM exhibited a slight upward trend at higher K values, while Neural LDA showed a modest increase around mid-range K values (K = 40–60).

According to the NPMI-TD analysis (**Fig. 3**), BERTopic and STM emerged as the models with comparatively better performance. While STM achieved marginally higher coherence scores (NPMI: 0.087-0.124), BERTopic demonstrated superior topic diversity (TD > 0.95) while maintaining comparable coherence levels. LDA, LSI, NMF and Neural LDA showed lower performance, with negative or near-zero NPMI scores and moderate to low TD.

Based on these findings, BERTopic emerged as the best model, offering the best balance of computational efficiency, topic diversity, and coherence across varying numbers of topics. BERTopic achieved optimal performance at K=30, balancing topic coherence and topic diversity.

4.2. Round 2: Fine-tuning the selected model (BERTopic)

We used the fine-tuned MoST-BERT model with BERTopic and evaluated eight LLMs to select the best Representation Model (**Fig. 4**). The evaluation revealed distinct performance patterns across different temperature settings (T). While Gemini-1.5-Pro achieved the highest topic diversity (0.87) at T=1.5, Mixtral-8x7b demonstrated superior performance, achieving the highest coherence (NPMI=0.45) and topic diversity (0.83) at T=1.0. Claude-3.5-Sonnet and DeepSeek-V3 showed stable performance across their temperature ranges, with Claude reaching notable diversity (0.82) at T=0.5 and DeepSeek maintaining consistent coherence scores (~0.25) at T=1.0-1.5.

Most models exhibited optimal performance at moderate temperatures (T=1.0), with diversity scores above 0.70. **Fig. 5** presents a comprehensive visualization of model performances at their optimal topic numbers (K), where BERTopic-Mixtral-8x7b clearly outperformed other approaches, achieving the highest combined performance score through optimal balance between coherence and diversity metrics. Traditional models clustered in the lower performance region, while other LLM-enhanced versions demonstrated intermediate performance levels.

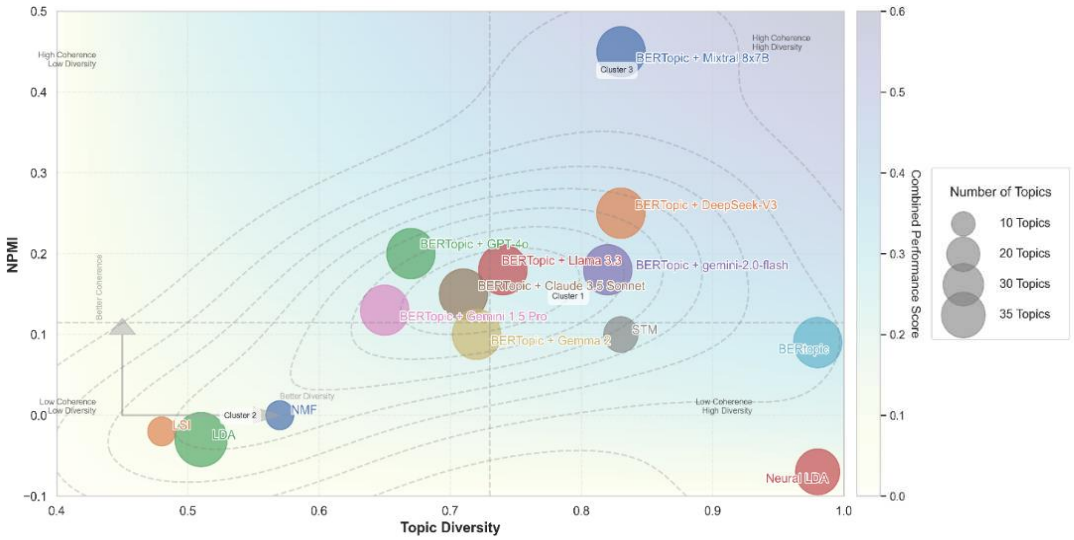


Fig. 5. Post-fine-tuning Comparison. Circle size indicates the optimal number of topics (K) for each model. The plot illustrates the trade-off between coherence and diversity, and the background contours represent the combined performance score.

4.3. Topic interpretability assessment

Before examining the LLM-as-a-judge results, **Fig.6** shows the semantic distribution of the 30 topics generated by the best-performing model (BERTopic-Mixtral). The projection indicates a heterogeneous topic space. Some topics, such as Culinary & Food Tourism (T13), Weather Forecasting (T23) and Wine Tourism (T24) formed compact and clearly separated regions. Others, including Digital Transformation, Vocational Education and Soft Skills, and Emotional Intelligence & Leadership appeared closer to neighboring semantic areas or showed more diffuse boundaries.

Fig.7 reports the word-intrusion precision by LLM and temperature. Overall, precision was high but varied across models. Gemma 4 31B achieved the highest average precision across temperatures (MP=0.98), followed by GPT-5.5 (MP= 0.90) and Llama 4 Scout 17B (MP=0.90), which matched or slightly exceeded the human reference (MP=0.89). GPT-5.2, GPT-5.1, gpt-oss-120b, and Llama 3.3 70B also performed strongly (MP=0.87-0.89), whereas smaller or less stable models, such as Llama 3.1 8B, GPT-5.4 nano, DeepSeek-V4-Flash, and gpt-oss-20b, obtained lower precision.

Precision generally decreased as temperature increased, although this effect was model-dependent. The highest aggregate performance was observed at (T=0.3). Some models remained stable across temperatures, particularly GPT-5.5, Llama 4 Scout 17B, and Gemma 4 31B, whereas DeepSeek-V4-Pro, gpt-oss-20b, DeepSeek-V4-Flash, and Llama 3.1 8B showed greater temperature sensitivity.

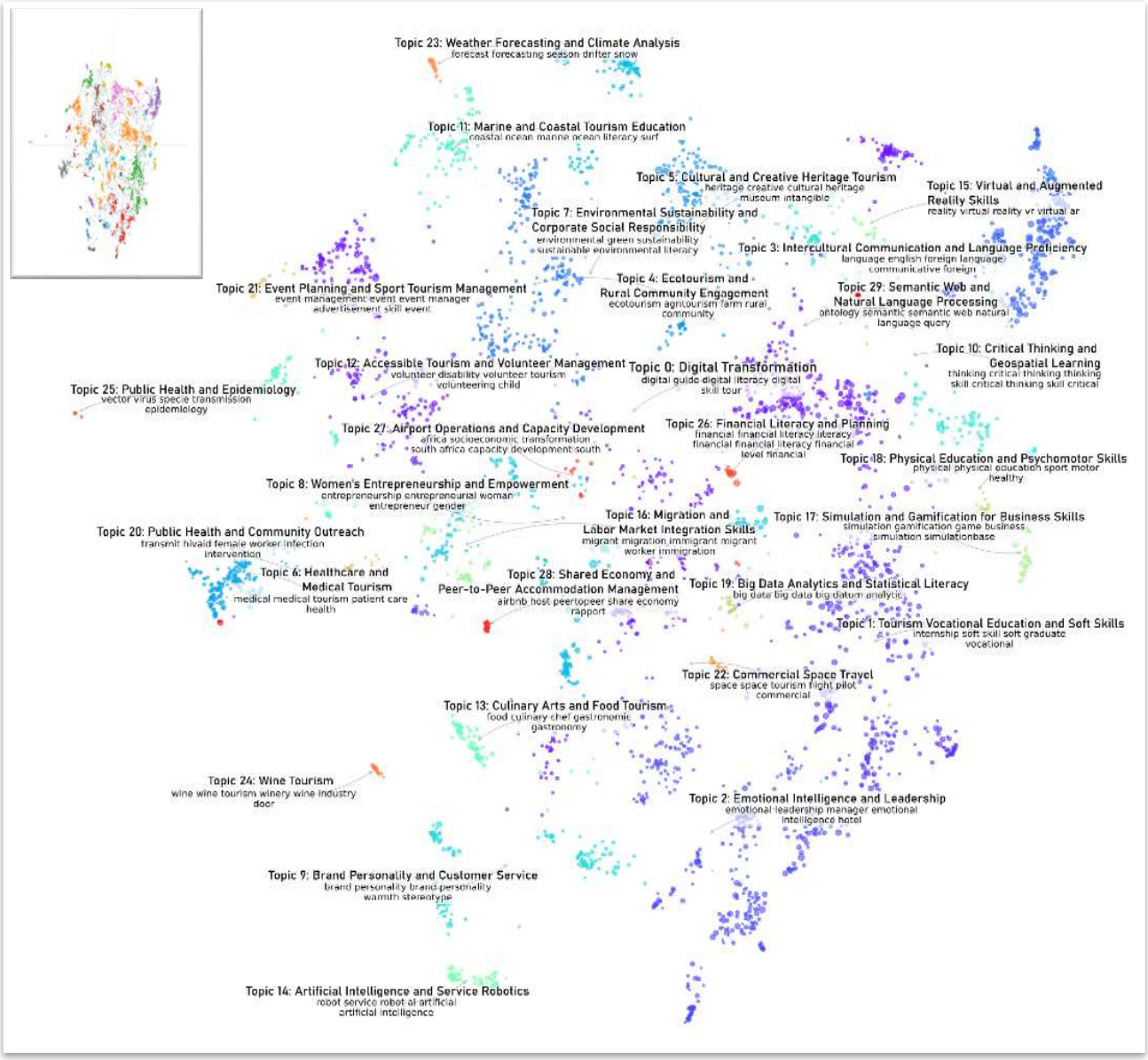


Fig. 6. Topic Distribution in 2D Space (BERTopic-Mixtral-8x7b)

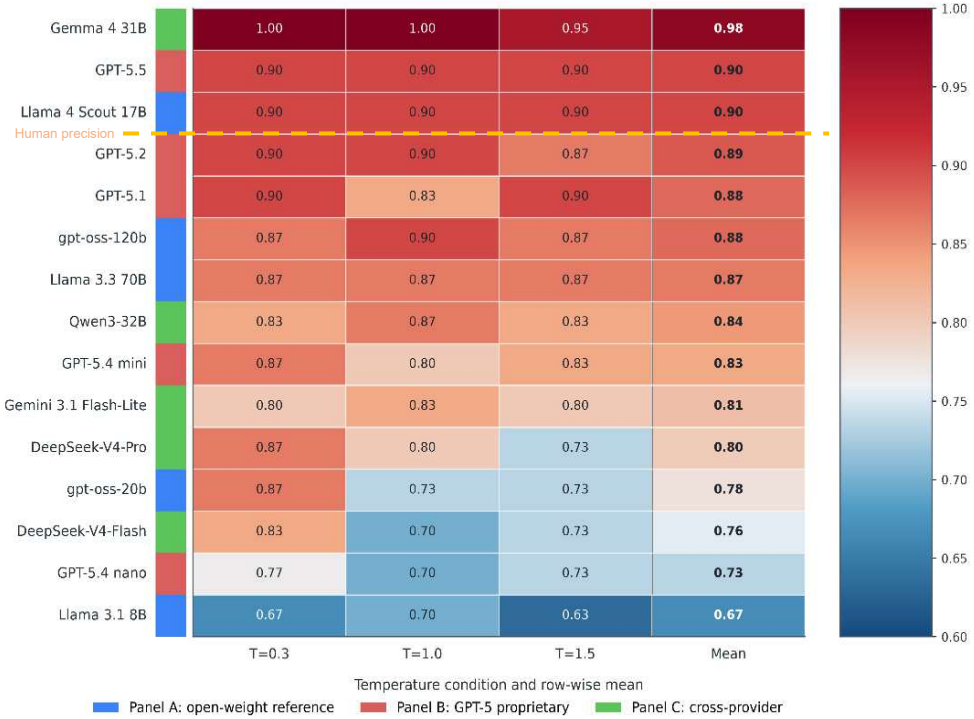


Fig. 7. Word-intrusion model precision (MP) by LLM and temperature condition.

Table 3 summarizes the intrusion-test results by evaluator group. Human coders obtained the highest group-level precision (MP=0.892), followed closely by Panel A (MP=0.820), Panel B (MP=0.827), Panel C (MP=0.831). The pooled LLM panel reached MP= 0.826, indicating that LLM-based evaluators were both internally consistent and generally successful at identifying the intruder term.

Table 3. Inter-rater reliability and model precision by evaluator group

Group	Model Precision	Krippendorff's alpha	95% CI	Pairwise agreement
Humans	0.892	0.818	0.705-0.909	0.828
Panel A	0.820	0.801	0.699-0.881	0.810
Panel B	0.827	0.832	0.745-0.906	0.840
Panel C	0.831	0.783	0.680-0.874	0.802
All LLMs	0.826	0.807	0.731-0.869	0.822
Humans + All LLMs	0.840	0.800	0.721-0.865	0.815

Reliability followed a similar pattern. Human coders reached a nominal Krippendorff's alpha of 0.818, while the combined LLM panel reached 0.807. Among the LLMs, Panel B showed the highest internal consistency (alpha=0.83) and pairwise agreement (0.840), similar to those obtained by Panel A (alpha=0.80) and Panel C (alpha =0.78). When human and LLM judgments were combined, reliability remained at the conventional acceptability threshold (alpha=0.800), suggesting that LLM evaluators generally achieved high word-intrusion precision and substantial inter-rater reliability, broadly consistent with human judgments. However, human coders remained the strongest benchmark.

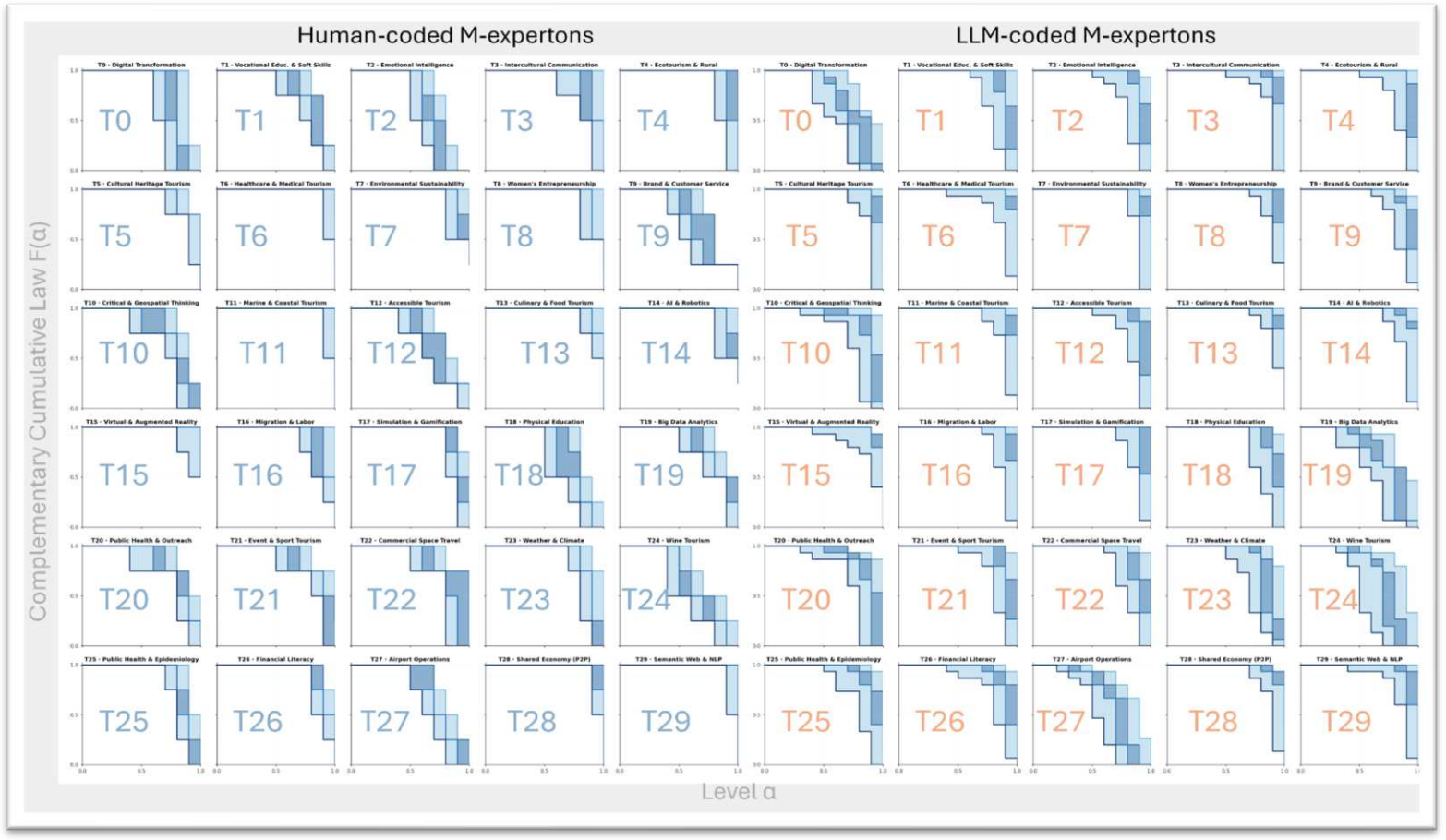


Fig. 8. *M*-experton cumulative laws for human-coded and LLM-coded topic interpretability judgments



Fig. 9. Per-topic interpretability across intrusion-based and *M*-experton-based evaluations

To complement the categorical evidence from the word-intrusion task, topic interpretability was then examined through the direct *M*-experton appraisal. **Fig.8** shows the *M*-expertons for the 30 topics, comparing human-coded and LLM-coded judgments. In both panels, most topics retain high cumulative values at upper experton (α-cut levels). Some topics show very similar cumulative functions in the human-coded and LLM-coded *M*-expertons, whereas others display larger differences. These findings indicate that the topics generally exhibit a high degree of interpretability, although interpretability levels are not uniform across the topic solution.

Given this heterogeneity, **Fig.9** examines topic-level interpretability in more detail across the evaluation phases and procedures. All measures are reported on a common 0-1 scale, where higher values indicate greater estimated topic interpretability.

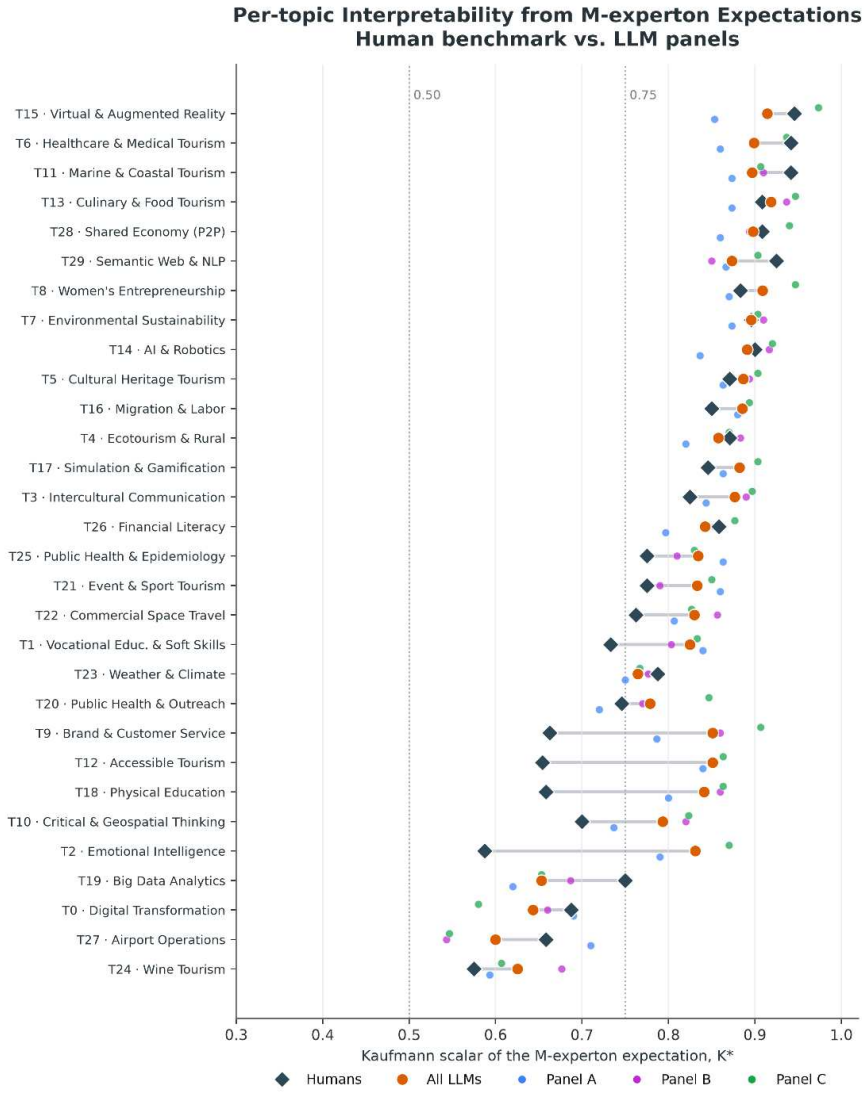


Fig. 10. Human and LLM panel *M*-experton expectations for topic interpretability

Several topics were consistently rated as highly interpretable across methods, including Culinary & Food Tourism, Environmental Sustainability, Marine & Coastal Tourism, Migration and Labor, and Intercultural Communication. These topics obtained both high intrusion model precision and high *M*-experton expectation scores. However, the findings also show that topic interpretability varied across evaluation procedures. In particular, some topics received low intrusion-test precision but higher *M*-experton scores, indicating that the two procedures did not always produce the same topic ranking. The largest within-topic ranges were observed for topics in which categorical intrusion accuracy and direct *M*-experton judgments diverged.

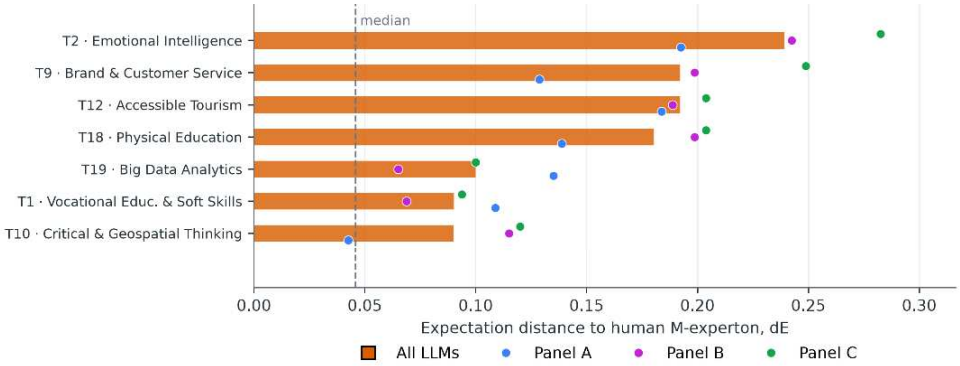


Fig. 11. Topics with the largest LLM-human divergence in M-experton expectations

More importantly, at the aggregate level, *M*-experton appraisals were more closely aligned with human judgments than word-intrusion scores, with a lower mean absolute LLM-human gap and a higher rank-order correlation across topics.

Self-stability results (**Appendix B**) show that the fuzzy judgments were highly reproducible across the initial and verification stages. Most models produced virtually identical quadruplets after verification, indicating that the protocol generated stable LLM judgments. Instability was concentrated in a small number of models, especially Llama 3.1 8B, and increased slightly at higher temperature (T=1.5).

Fig.10 further examines the *M*-experton expectations for topic interpretability by comparing the human benchmark with the LLM panels. The topics are ordered according to their average human and LLM values. Overall, both groups produced similar topic-level patterns. Topics such as Virtual and Augmented Reality, Healthcare and Medical Tourism, Marine and Coastal Tourism, Culinary and Food Tourism, Shared Economy, Semantic Web and NLP, Women's Entrepreneurship, and Environmental Sustainability were in the upper range of the endecadary scale.

Despite this broad similarity, the LLM panels tended to assign slightly higher values than the human benchmark for several topics. This pattern was especially visible for Emotional Intelligence, Accessible Tourism, Physical Education, Brand and Customer Service, and Vocational Education and Soft Skills. Nevertheless, the relative ordering of topics remained broadly comparable across humans and LLM panels.

Fig. 11 identifies the topics with the largest expectation distance between the human *M*-experton and the aggregated LLM *M*-experton. The greatest divergence was observed for Emotional Intelligence, followed by Brand and Customer Service, Accessible Tourism, Physical Education, Big Data Analytics, Vocational Education and Soft Skills, and Critical and Geospatial Thinking. For these topics, aggregated LLM expectations were farther from the human-coded *M*-expertons than the median topic-level distance. Panel-level markers indicate that this divergence was not homogeneous across LLMs. Panel C was the most distant in several high-gap topics, whereas Panels A and B were closer to the human benchmark in some cases.

Fig.12 reports the uncertainty width of the *M*-experton expectations. The support width was consistently larger for the LLM groups than for human coders. Human coders had a mean support width of 0.193, whereas the full LLM panel reached 0.259.

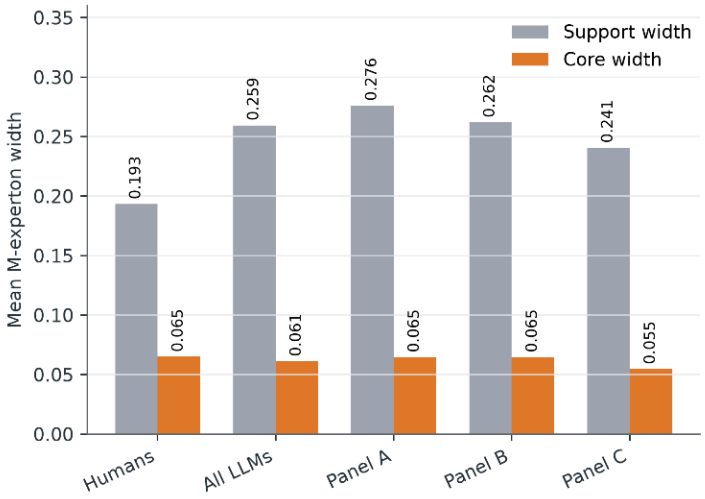


Fig. 12. Uncertainty width of human and LLM *M*-experton judgments

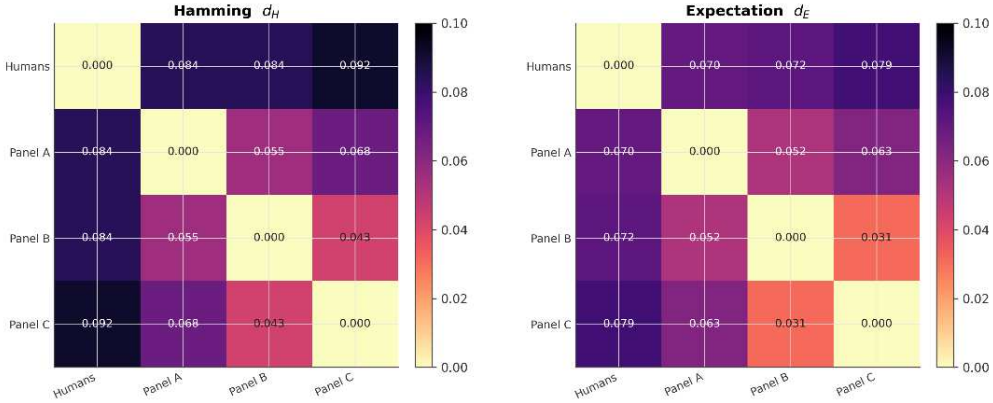


Fig. 13. Counter-appraisement distances between aggregated *M*-experton panels

Among the LLM panels, Panel A showed the widest support interval (0.276), followed by Panel B (0.262) and Panel C (0.241). By contrast, core widths were very similar across groups: humans obtained 0.065, the full LLM panel 0.061, Panel A 0.065, Panel B 0.065, and Panel C 0.055. Thus, the main difference between humans and LLMs was located in the outer support interval rather than in the core of maximum presumption.

Appendix B also provides a topic-level decomposition of *M*-experton uncertainty. The results show that support-width differences between humans and LLM panels were not uniform across topics, with the largest uncertainty widths concentrated in topics such as Wine Tourism, Digital Transformation, Critical and Geospatial Thinking, Commercial Space Travel, and Airport Operations.

Finally, **Fig.13** reports the counter-appraisement distances between aggregated *M*-expertons for topic interpretability. The Hamming and expectation distances produced consistent patterns. The smallest expectation distance was observed between Panel B and Panel C (d_E = 0.031), followed by Panel A - Panel B (d_E = 0.052), and Panel A - Panel C (d_E = 0.063). Distances between the human benchmark and the LLM panels were

marginally larger: Humans-Panel A (d_E = 0.070), Humans-Panel B (d_E = 0.072), and Humans-Panel C (d_E = 0.079). The same ordering was broadly reproduced by the Hamming distances.

Pairwise agreement heatmaps by temperature are included in **Appendix B** providing details on the SMC metric between all human and LLM evaluators at T=0.3, T=1.0, and T=1.5. The matrices show that agreement clusters were broadly stable across temperatures, while also identifying lower-agreement evaluators and models.

Overall, the results reveal heterogeneous interpretability across the topic solution. Human and LLM appraisals showed comparable ranking structures, with divergence concentrated in the support intervals and in a small group of topics.

5. Discussion

The comparative analysis of six topic modeling approaches (STM, LDA, LSI, NMF, Neural LDA, and BERTopic) revealed significant performance differences, highlighting the advancements of neural methods and the critical influence of the temperature parameter in LLM-enhanced models.

5.1. Round 1: Baseline model comparison and selection

The fact that BERTopic maintained high TD scores even as K increases suggests robust resistance to topic collapse. An unexpected finding was STM's slightly superior NPMI scores compared to BERTopic, despite its lower TD values.

This pattern, a traditional model retaining a slight coherence advantage while BERTopic leads on diversity, is consistent with comparisons outside this corpus. Mihajlov et al. [47] likewise reported that NMF achieved the best coherence and BERTopic the best diversity, and Kamal et al. [21] found matrix-factorization models competitive or superior on coherence across several tourism datasets. This may suggest that some degree of topic overlap could actually contribute to improved coherence, challenging the wisdom that maximally distinct topics are always preferable, which could be seen as a limitation in STM performance on TOURSKAD dataset.

Our results also suggest that the optimal number of topics may be model-dependent. The peak performance of BERTopic at K=20-35 contrasts with STM's slight improvement at higher K values, indicating that model selection and parameter tuning should be considered jointly rather than sequentially. In other words, we should avoid assuming that a fixed number of topics will perform equally well across different models.

5.2. Round 2: Fine-tuning the selected model (BERTopic)

To address the first research question (**RQ1**), the study demonstrated the effectiveness of advanced neural topic modeling techniques, specifically the LLM-enhanced BERTopic model. As evidenced in **Section 4.1**, BERTopic outperformed traditional models like LDA, NMF, and LSI in terms of topic coherence and diversity. Furthermore, the second round of experiments (**Section 4.2**) revealed that fine-tuning the embedding model (using MoST-BERT) and optimizing the LLM representation significantly improved the model's performance, as shown in **Fig. 4**.

The use of LLMs, particularly Mixtral-8x7b, led to more coherent representations, allowing for a more contextualized understanding of the topics. In fact, the concurrent identification of broad operational themes ("Digital Transformation"), soft skills ("Emotional Intelligence & Leadership"), and specialized technical competencies ("AI & Service Robotics") indicates the framework's capacity to provide a granular mapping of the tourism skill landscape, an outcome that can be challenging for traditional models.

To address the topic-representation component of **RQ2**, we analyzed different temperature settings (T). We found that the NPMI scores of BERTopic-Mixtral-8x7b varied from 0.20 (T=0.3) to 0.45 (T=1.0) while for traditional models these variations were not achieved by optimizing the number of topics, which is the most common in literature [15]. This significant variation in Topic Modeling performance observed as a result of different temperature settings is an important finding, because although a few researchers have already begun to explore the field of Hybrid Topic Modeling combining Neural Topic Models with LLMs [61–63], few studies have been conducted on the impact of the temperature hyperparameter on topic quality.

Our results also provide insights into the evolving landscape of LLM capabilities. While [6] identified GPT-4 as the optimal choice for topic modeling representations, our evaluation, which included other models like Gemini-1.5-Pro, DeepSeek-V3, Llama 3.3, Claude-3.5-Sonnet or GPT-4o, and employed hyperparameter tuning for optimal K, revealed different patterns. In fact, we found that the GPT-4o model did not perform as well on the TOURSKAD dataset as other models such as Mixtral-8x7b, Claude-3.5-Sonnet, Gemini-1.5-Pro, or DeepSeek-V3 when considering different levels of temperature. This discrepancy underscores the importance of regularly reassessing model selection.

The complementary relationship between LLMs and BERTopic also emerged as a key finding. The results showed that using BERTopic alone yields suboptimal results compared to its integration with advanced LLM representations. However, the data also suggests that relying solely on LLMs would be insufficient, as they typically demonstrate lower topic diversity compared to BERTopic's sophisticated dimensionality reduction and clustering techniques.

From a practical perspective, the strong performance of Mixtral-8x7b at T=1.0 has significant implications for accessibility and cost-effectiveness. While models like GPT-4o, DeepSeek-V3, Claude-3.5-Sonnet, and Gemini-1.5-Pro demonstrated competitive performance in specific metrics, their API access costs may limit widespread adoption. In contrast, Mixtral-8x7b's availability through open-access APIs (e.g., Groq's LPU-powered infrastructure) offers a favorable cost-performance balance for both researchers and practitioners. Previous studies, such as [6] focused on proprietary models; however, our results demonstrate that with balanced fine-tuning, it is possible to achieve superior outcomes even when using open-source models like Llama 3.3 or Mixtral. This finding paves the way for new research avenues that integrate the strengths of neural topic models with open-access LLMs.

Topic modeling with the best representation. The performance patterns observed across different topic modeling architectures reveal distinct capabilities and limitations of each approach. Traditional algebraic (NMF, LSI) and probabilistic approaches (LDA, STM) clustered in the lower-performance quadrant (NPMI<0.10, TD<0.60). However, LLM-enhanced BERTopic implementations demonstrated superior performance in both coherence and diversity metrics, indicating a significant advancement in topic modeling capabilities.

Importantly, this advance concerns coherence specifically. In Round 1, and consistent with prior comparisons [21, 47], BERTopic did not surpass traditional models on coherence; its consistent advantage was diversity. The LLM-refinement step is what raised coherence above the traditional baseline, yielding a model that leads on both metrics simultaneously. Because earlier tourism comparisons evaluated BERTopic and traditional models in their standalone form [6, 21], this gain from LLM-based representation refinement had not previously been observed against traditional baselines in a tourism corpus.

While classic probabilistic models like LDA have provided foundational frameworks for understanding document-topic relationships [25], their limited semantic understanding constrains their ability to identify complex thematic patterns. The superior performance of LLM-enhanced approaches indicates that incorporating contextual understanding and advanced semantic representations can significantly improve topic quality [22, 61, 64].

5.3. Topic interpretability assessment

With respect to the second component of **RQ2**, namely the effect of temperature on topic interpretability assessment, the results suggest that low temperature values may yield the highest average accuracy in LLM-based word-intrusion judgments, although several models, such as GPT-5.5, remained stable across temperature settings. These findings indicate that temperature should be reported not only as a topic-representation parameter but also as an evaluation parameter, since it influences the robustness of LLM-based interpretability judgments.

Having established the role of temperature in the evaluation process, **RQ3** focuses on how topic interpretability itself should be assessed. Recent neural topic modeling studies often operationalize interpretability through NPMI coherence and topic diversity [48, 65]. In contrast, this study argues that these metrics are insufficient on their own to assess topic interpretability in LLM-enhanced topic modeling.

Our results show that the two approaches produced similar patterns for several topics. Yet, the word-intrusion tests yielded more polarized evaluations: some topics were rated as highly interpretable, such as Culinary & Food Tourism and Environmental Sustainability, whereas others received very low scores, such as Emotional Intelligence. This may be partly explained by the structure of the intrusion task, which relies on isolated intruder words and requires a discrete choice, without giving evaluators the broader contextual evidence available in the full topic representation, such as representative documents. By contrast, the *M*-experton approach produced less polarized assessments and closer alignment between LLM and human appraisals at the aggregate level. This suggests that the proposed method is better suited to representing graded topic interpretability and the uncertainty associated with each topic.

This evaluation strategy also differs from how prior tourism studies have used LLMs to assess topics. Wang et al. [49] combined BERTopic with two LLMs (ChatGPT and DeepSeek) to validate hotel-review topics, but treated their agreement as a binary accept-or-reject decision and relied on human review only when the two models disagreed, noting that the absence of a systematic human benchmark may affect interpretive accuracy. Our protocol instead uses an independent human benchmark alongside multiple LLM panels and preserves the graded, uncertain nature of each judgment rather than reducing it to agreement. It thus offers a concrete way to address the trust and explainability concerns raised by Kirilenko and Stepchenkova [6]: rather than treating

LLM judgments as opaque, the *M*-experton makes their uncertainty explicit, and we find that LLMs express systematically wider support intervals than human coders.

At the aggregate level, our findings show that human coders remain the strongest benchmark, achieving the highest levels of accuracy. However, the LLM-as-a-judge approach reveals that LLMs can approximate or even partially outperform human judgments (e.g., Gemma 4 31B and GPT 5.5) in evaluating topic interpretability. These findings are consistent with previous studies demonstrating that LLMs constitute a viable alternative for topic modeling by providing reliable results under explicit evaluation protocols [8, 61, 62, 66].

Panel composition and model selection also influenced the evaluation. The three LLM panels produced broadly similar results, but Panel B showed the highest internal agreement, whereas Panel C achieved the highest group-level accuracy among the LLM panels. The best-performing individual model in the word intrusion task was Gemma 4 31B, an open-weight model, which outperformed both the average human accuracy and the proprietary GPT-5.5 model on this specific task. This finding demonstrates that access to proprietary models is not a prerequisite for robust topic interpretability evaluation.

Topic interpretability is heterogeneous. Even with state-of-the-art models such as BERTopic, which automatically exclude outliers [3, 9], interpretability should be assessed at the topic level rather than solely at the model level. Our results suggest that lexical overlap between semantically related domains can reduce word intrusion accuracy even when a topic remains meaningful in context. This supports recent evidence that AI-assisted, embedding-based topic modeling improves semantic representation in heterogeneous tourism corpora but still requires careful interpretability validation [6, 67]

In practice, recent LLM research has largely evaluated responses as either correct or incorrect, while paying comparatively little attention to tasks where judgments are inherently subjective. Our findings suggest that topic interpretability should not be treated solely as an accuracy problem but rather as one involving subjective and imprecise judgments.

6. Conclusion

This study makes three contributions to topic modeling research. First, this study provides practical guidance for selecting topic modeling approaches in specialized tourism corpora. The results indicate that an LLM-enhanced BERTopic pipeline with domain-adapted embeddings offers the best overall balance between topic coherence, diversity, and interpretability.

When computational resources or access to LLMs are limited, traditional models such as NMF or STM remain competitive alternatives depending on the analytical objective, whereas BERTopic without LLM refinement is preferable when maximizing topic diversity is the primary goal. The study offers evidence-based recommendations that help researchers select the most appropriate approach according to their analytical priorities.

From a practical perspective, the strong performance of open-weight models suggests that hybrid topic modeling can be implemented using locally deployable solutions, facilitating reproducible and cost-effective research under diverse computational and data-governance constraints.

Second, the study establishes temperature as a substantive modeling parameter in LLM-enhanced topic modeling by demonstrating its influence on both topic representation and interpretability assessment. Moderate temperatures generally improved the balance between coherence and diversity, while lower temperatures tended to produce more stable LLM-based interpretability judgments. This finding highlights that temperature is not merely a decoding setting but a methodological choice that shapes multiple stages of the topic-modeling workflow.

Finally, the main contribution of this study is the extension of the General Theory of Expertons to topic interpretability evaluation under LLM-as-a-Judge protocols. The proposed framework aggregates uncertain expert judgments in a setting where both human coders and LLM panels act as semantic evaluators.

More specifically, expertons have the advantage of preserving entropy until the final aggregation stage [43–45], thereby reducing information loss while retaining the support interval, the core of maximum presumption, and uncertainty widths. This may better reflect how both humans and LLMs generate semantic evaluations in natural language.

More importantly, our results show that LLMs produce wider fuzzy support intervals than human coders, highlighting the need for evaluation methods, such as *M*-expertons, that explicitly account for LLM uncertainty.

Conventional large-scale LLM-as-a-Judge protocols often require numerous repeated model calls to achieve reliable estimates [68]. For example, self-consistency (SC), a widely used decoding strategy for chain-of-thought reasoning [69], improves performance on multi-step reasoning tasks but incurs substantial computational, financial, and environmental costs because it relies on repeated sampling, even for questions that could be answered accurately in a single attempt [68, 70].

This issue is particularly relevant in tourism research, where increasingly large and heterogeneous corpora make manual topic inspection costly and difficult to reproduce. The proposed *M*-experton framework provides a more resource-efficient evaluation design. Rather than relying on extensive repeated sampling to maximize accuracy [68, 70], it explicitly models interpretability as an imprecise problem [44] by estimating uncertainty within a plausible range using only a small number of API calls. This reduces reliance on extensive repeated API calls while enabling the evaluation of 15 different LLMs across multiple temperature settings.

Ultimately, it should be noted that LLMs not only assigned scores with greater uncertainty but also, overall, gave higher ratings than human coders. This may reflect greater sensitivity to latent semantic associations among topic words rather than simple optimistic overestimation, as LLMs are specifically designed to capture patterns in the semantic structure of language. In this sense, LLMs may complement human coders by identifying semantic regularities that are less readily apparent in manual evaluation.

Limitations. This study has several limitations. First, the experiments were conducted on a single specialized tourism corpus, which limits the extent to which the results can be generalized to other domains, languages, document types, or levels of textual noise. Second, LLM-enhanced topic modeling depends on computational resources, API availability, and model-access conditions that may affect cost, scalability, and replicability. Third, the results are tied to the specific LLM versions, providers, and temperature settings available at the time of the study; performance and interpretability patterns may shift as models are updated or discontinued. Finally, although human coders were used as a benchmark for interpretability assessment, larger and more diverse expert panels would strengthen the validation of the proposed framework.

Future research. Future research should test the proposed framework across other specialized domains, multilingual corpora, and different types of documents to assess its robustness beyond tourism. Further studies should also examine the temporal stability of topic representations and LLM-based interpretability judgments across model versions, providers, and decoding configurations. In addition, future work could combine *M*-experton-based aggregation with richer qualitative expert evaluation to better understand why some topics are consistently interpretable while others remain semantically ambiguous. Finally, the integration of AI agents with neural topic modeling offers a promising direction for automating topic-quality diagnostics, verifying topic coherence, detecting unstable representations, and dynamically adjusting parameters during the modeling process.

Appendix A. Formal definitions and metric properties of *M*-experton distances

Definition 1. Expectation Distance. Let $\mathcal{A}^+ \subset (0,1]$ denote the set of non-zero confidence levels used to construct a discrete *M*-experton, and let $\mathcal{R} = \{L, C_1, C_2, U\}$ denote the set of components of a quadruplet *M*-experton. For a discrete *M*-experton $\widetilde{M}$, the component-wise mathematical expectation is defined as:

$$E_r(\widetilde{M}) = \frac{1}{|\mathcal{A}^+|} \sum_{\alpha \in \mathcal{A}^+} M_r(\alpha), \qquad r \in \mathcal{R}.$$

The expected *M*-experton is therefore the quadruplet:

$$E(\widetilde{M}) = \big(E_L(\widetilde{M}), E_{C_1}(\widetilde{M}), E_{C_2}(\widetilde{M}), E_U(\widetilde{M})\big).$$

Given two *M*-expertons $\widetilde{M}_A$ and $\widetilde{M}_B$, the expectation distance is defined as the normalized L_1 distance between their expected quadruplets:

$$d_E(\widetilde{M}_A, \widetilde{M}_B) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left| E_r(\widetilde{M}_A) - E_r(\widetilde{M}_B) \right|.$$

Proposition 1. Metric properties of the expectation distance. The expectation distance d_E is a metric on the space of expected *M*-expertons $E(\widetilde{M}) \in [0,1]^4$.

Proof

The expectation distance d_E is the normalized L_1 distance between the expected quadruplets $E(\widetilde{M}_A)$ and $E(\widetilde{M}_B)$. Therefore, non-negativity follows from the non-negativity of the absolute value: $d_E(\widetilde{M}_A, \widetilde{M}_B) \geq 0$. Moreover, $d_E(\widetilde{M}_A, \widetilde{M}_B) = 0$ if and only if $E_r(\widetilde{M}_A) = E_r(\widetilde{M}_B), \forall r \in \mathcal{R}$. Hence, $E(\widetilde{M}_A) = E(\widetilde{M}_B)$, which establishes the identity of indiscernibles in the expectation space.

Symmetry follows directly from the symmetry of the absolute value: $\left|E_r(\widetilde{M}_A) - E_r(\widetilde{M}_B)\right| = \left|E_r(\widetilde{M}_B) - E_r(\widetilde{M}_A)\right|$. Therefore, $d_E(\widetilde{M}_A, \widetilde{M}_B) = d_E(\widetilde{M}_B, \widetilde{M}_A)$.

For any third expected *M*-experton $E(\widetilde{M}_C)$, the triangle inequality of the absolute value gives $\left|E_r(\widetilde{M}_A) - E_r(\widetilde{M}_C)\right| \leq \left|E_r(\widetilde{M}_A) - E_r(\widetilde{M}_B)\right| + \left|E_r(\widetilde{M}_B) - E_r(\widetilde{M}_C)\right|$. Summing over all $r \in \mathcal{R}$ and multiplying by $1/|\mathcal{R}|$ yields: $d_E(\widetilde{M}_A, \widetilde{M}_C) \leq d_E(\widetilde{M}_A, \widetilde{M}_B) + d_E(\widetilde{M}_B, \widetilde{M}_C)$. ■

Thus, d_E satisfies non-negativity, identity of indiscernibles, symmetry, and the triangle inequality on the space of expected *M*-expertons. Therefore, d_E is a metric on the expectation space.

Definition 2. Full-law hamming distance. Let $\mathcal{A}^+ \subset (0,1]$ denote the set of non-zero confidence levels used to construct a discrete *M*-experton, and let $\mathcal{R} = \{L, C_1, C_2, U\}$ denote the set of components of a quadruplet *M*-experton. Given two discrete *M*-expertons $\widetilde{M}_A$ and $\widetilde{M}_B$, the full-law Hamming distance is defined as the normalized sum of absolute differences between their cumulative component functions:

$$d_H(\widetilde{M}_A, \widetilde{M}_B) = \frac{1}{|\mathcal{A}^+|\,|\mathcal{R}|} \sum_{\alpha \in \mathcal{A}^+} \sum_{r \in \mathcal{R}} \left| M_{A,r}(\alpha) - M_{B,r}(\alpha) \right|.$$

Proposition 2. Metric properties of the full-law hamming distance. The full-law Hamming distance d_H is a metric on the space of discrete M-expertons defined on the same confidence-level grid $\mathcal{A}^+$ and component set $\mathcal{R}$.

Proof

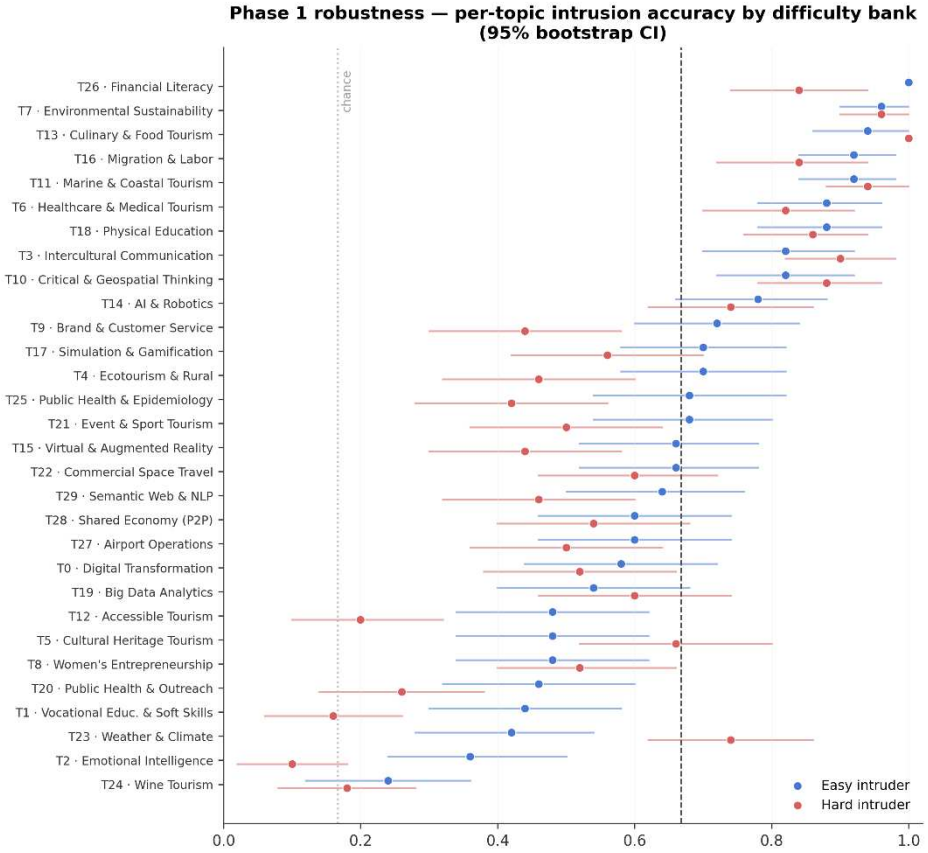
For any two M-expertons $\widetilde{M}_A$ and $\widetilde{M}_B$, each term $\left|M_{A,r}(\alpha) - M_{B,r}(\alpha)\right|$ is non-negative. Hence, $d_H(\widetilde{M}_A, \widetilde{M}_B) \geq 0$. Furthermore, $d_H(\widetilde{M}_A, \widetilde{M}_B) = 0$ if and only if $\left|M_{A,r}(\alpha) - M_{B,r}(\alpha)\right| = 0$ for every $\alpha \in \mathcal{A}^+$ and every $r \in \mathcal{R}$. Therefore, $M_{A,r}(\alpha) = M_{B,r}(\alpha), \quad \forall \alpha \in \mathcal{A}^+, \forall r \in \mathcal{R}$, which implies $\widetilde{M}_A = \widetilde{M}_B$.

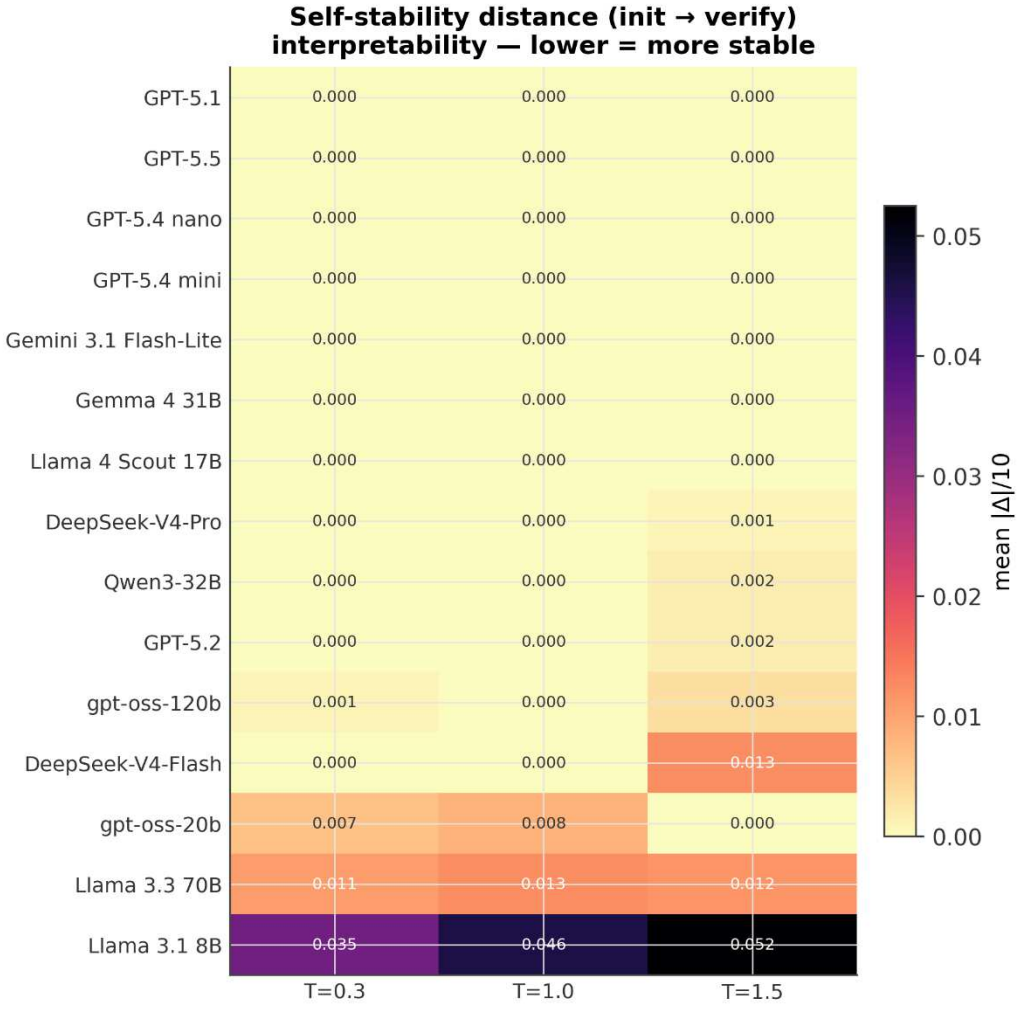
Symmetry follows from the symmetry of the absolute value: $\left|M_{A,r}(\alpha) - M_{B,r}(\alpha)\right| = \left|M_{B,r}(\alpha) - M_{A,r}(\alpha)\right|$. Consequently, $d_H(\widetilde{M}_A, \widetilde{M}_B) = d_H(\widetilde{M}_B, \widetilde{M}_A)$.

Finally, for any third M-experton $\widetilde{M}_C$, the triangle inequality of the absolute value gives $\left|M_{A,r}(\alpha) - M_{C,r}(\alpha)\right| \leq \left|M_{A,r}(\alpha) - M_{B,r}(\alpha)\right| + \left|M_{B,r}(\alpha) - M_{C,r}(\alpha)\right|$. Summing over all $\alpha \in \mathcal{A}^+$ and $r \in \mathcal{R}$, and multiplying by the positive normalization factor $1/(|\mathcal{A}^+|\,|\mathcal{R}|)$, yields $d_H(\widetilde{M}_A, \widetilde{M}_C) \leq d_H(\widetilde{M}_A, \widetilde{M}_B) + d_H(\widetilde{M}_B, \widetilde{M}_C)$. ■

Thus, d_H satisfies non-negativity, identity of indiscernibles, symmetry, and the triangle inequality. Therefore, d_H is a metric on the space of discrete M-expertons defined on the same confidence-level grid and component set.

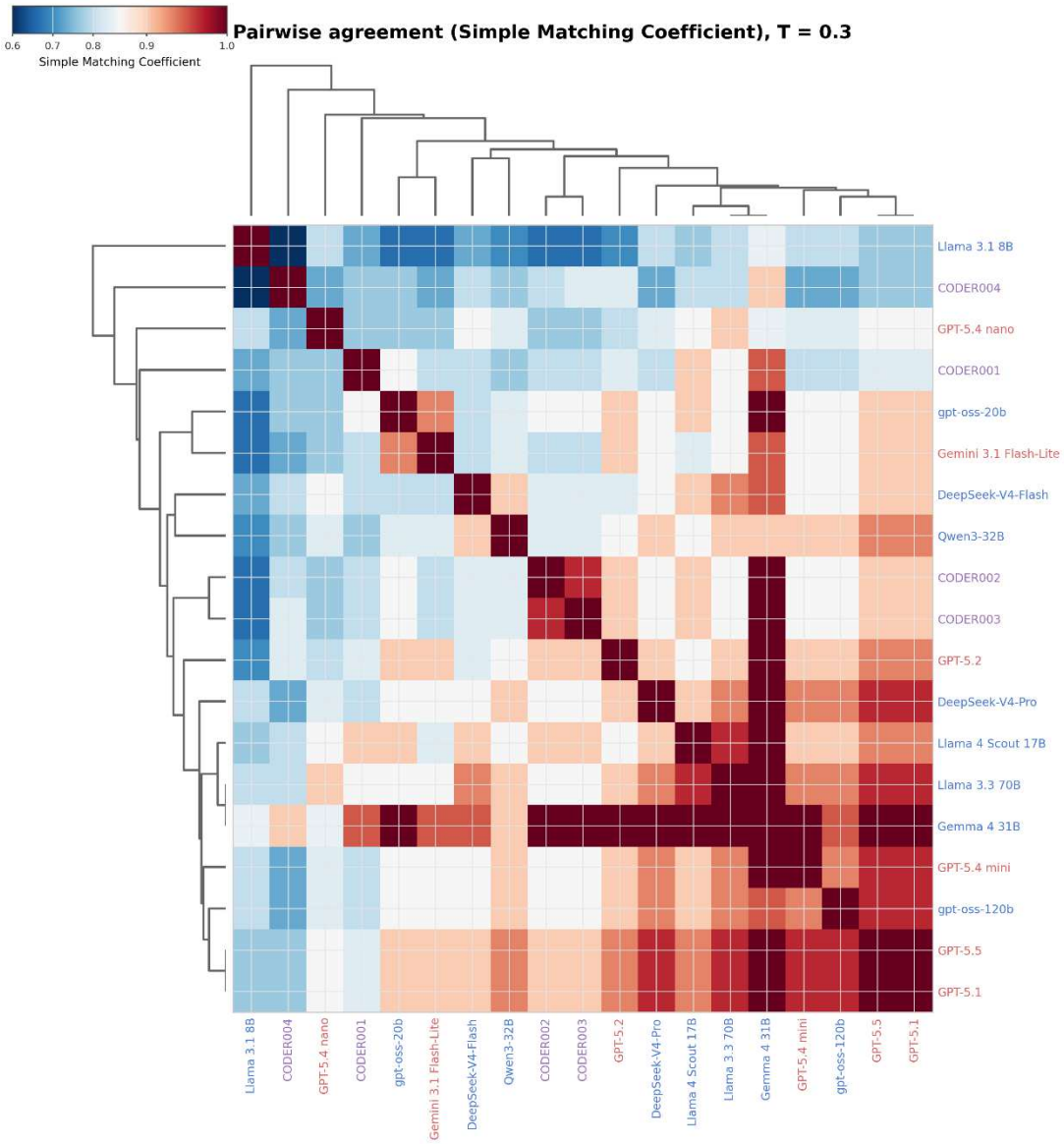
Appendix B. Supplementary robustness and agreement analyses for topic interpretability

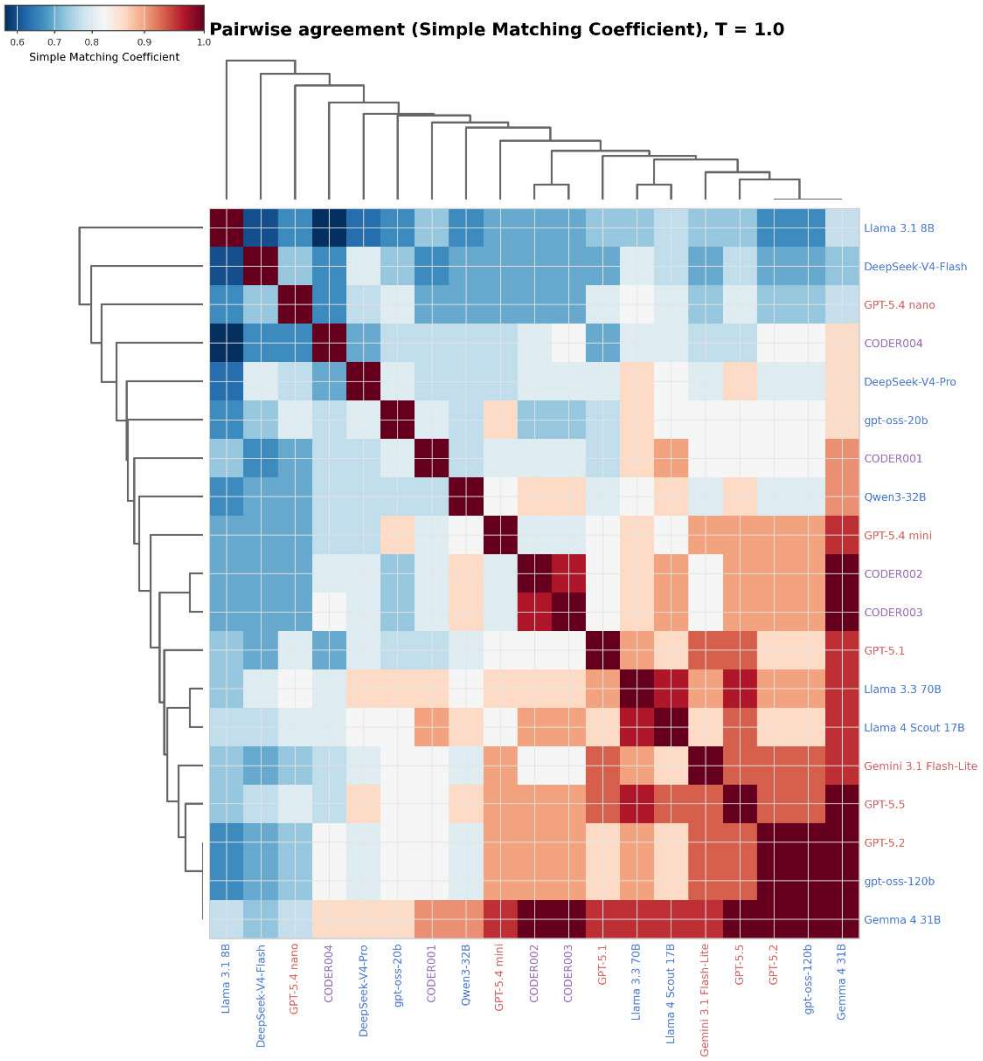


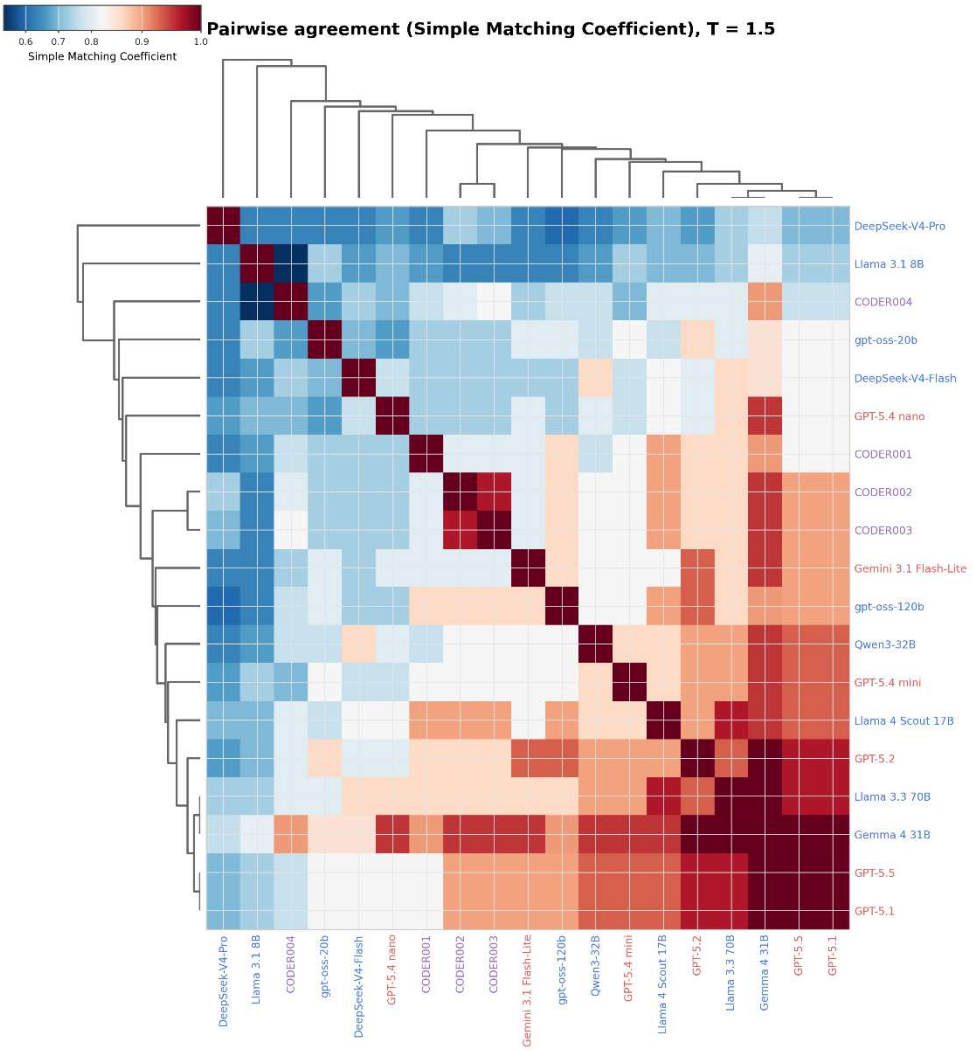

Self-stability distance (init → verify)
interpretability — lower = more stable
GPT-5.1
GPT-5.5
GPT-5.4 nano
GPT-5.4 mini
Gemini 3.1 Flash-Lite
Gemma 4 31B
Llama 4 Scout 17B
DeepSeek-V4-Pro
Qwen3-32B
GPT-5.2
gpt-oss-120b
DeepSeek-V4-Flash
gpt-oss-20b
Llama 3.3 70B
Llama 3.1 8B
0.000 0.000 0.000
0.000 0.000 0.000
0.000 0.000 0.000
0.000 0.000 0.000
0.000 0.000 0.000
0.000 0.000 0.000
0.000 0.000 0.000
0.000 0.000 0.001
0.000 0.000 0.002
0.000 0.000 0.002
0.001 0.000 0.003
0.000 0.000 0.013
0.007 0.008 0.000
0.011 0.013 0.012
0.035 0.046 0.052
T=0.3
T=1.0
T=1.5
mean |Δ|/10
0.05
0.04
0.03
0.02
0.01
0.00



Greater support width indicates a wider expected fuzzy interval, i.e., more uncertainty/imprecision in the aggregated judgment.


Pairwise agreement (Simple Matching Coefficient), T = 0.3
0.6
0.7
0.8
0.9
1.0
Simple Matching Coefficient
Llama 3.1 8B
CODER004
GPT-5.4 nano
CODER001
gpt-oss-20b
Gemini 3.1 Flash-Lite
DeepSeek-V4-Flash
Qwen3-32B
CODER002
CODER003
GPT-5.2
DeepSeek-V4-Pro
Llama 4 Scout 17B
Llama 3.3 70B
Gemma 4 31B
GPT-5.4 mini
gpt-oss-120b
GPT-5.5
GPT-5.1


Pairwise agreement (Simple Matching Coefficient), T = 1.0
0.6
0.7
0.8
0.9
1.0
Simple Matching Coefficient
Llama 3.1 8B
DeepSeek-V4-Flash
GPT-5.4 nano
CODER004
DeepSeek-V4-Pro
gpt-oss-20b
CODER001
Qwen3-32B
GPT-5.4 mini
CODER002
CODER003
GPT-5.1
Llama 3.3 70B
Llama 4 Scout 17B
Gemini 3.1 Flash-Lite
GPT-5.5
GPT-5.2
gpt-oss-120b
Gemma 4 31B


Pairwise agreement (Simple Matching Coefficient), T = 1.5
0.6
0.7
0.8
0.9
1.0
Simple Matching Coefficient
DeepSeek-V4-Pro
Llama 3.1 8B
CODER004
gpt-oss-20b
DeepSeek-V4-Flash
GPT-5.4 nano
CODER001
CODER002
CODER003
Gemini 3.1 Flash-Lite
gpt-oss-120b
Qwen3-32B
GPT-5.4 mini
Llama 4 Scout 17B
GPT-5.2
Llama 3.3 70B
Gemma 4 31B
GPT-5.5
GPT-5.1

7. Statements and Declarations

Funding

This research was conducted with the support of the Joan Oró FI predoctoral grant programme from the Department of Research and Universities of the Government of Catalonia and co-funded by the European Social Fund Plus, under Grant 2023 FI-1 00622. This work is also part of Grant PID2021-122575NB-I00, funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe". This work was also partially funded by Universitat Rovira i Virgili (URV) under project numbers 2023PFR-URV-114 and 2022PFR-URV-41, and by Generalitat de Catalunya (AGAUR) to the ITAKA consolidated research group under Grant 2021-SGR-00114. The support of the Spanish Research Network on Recommender Systems, ELIGE-IA, RED2022-134302-T, is also acknowledged.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

Eddy Soria contributed to conceptualization, methodology, software, formal analysis, investigation, data curation, writing the original draft, and visualization. Antonio Moreno contributed to methodology, supervision, validation, formal analysis, and writing, review and editing. Jordi Pascual contributed to methodology, validation, formal analysis, and writing, review and editing. Ana Beatriz Hernández-Lara contributed to conceptualization, formal analysis, methodology, supervision, validation, and writing, review and editing. All authors read and approved the final manuscript.

Data availability

The dataset analyzed in this study is publicly available in the CORA repository: [doi: 10.34810/data2133]. The dataset citation should be included in the reference list following the journal's data citation guidelines.

Code availability

The code used in this study is not publicly available. Additional implementation details may be provided by the corresponding author upon reasonable request, subject to institutional and project-related restrictions.

Ethics approval

Not applicable. This study did not involve external human participants, human biological material, animals, or clinical data.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Use of Large Language Models

Large Language Models were used as experimental components for topic representation refinement and interpretability assessment, as described in the Methods section. No Large Language Model is listed as an author. The authors take full responsibility for the study design, computational experiments, analyses, interpretation of results, and final manuscript content.

Acknowledgments

The authors acknowledge the support of the ITAKA research group at Universitat Rovira i Virgili, where the computational experiments were conducted.

8. References

[1] Abdelrazek A, Eid Y, Gawish E, et al. Topic modeling algorithms and applications: A survey. *Inf Syst* 2023; 112: 1–17, doi: 10.1016/j.is.2022.102131.

[2] Wu X, Nguyen T, Luu AT. A survey on neural topic models: methods, applications, and challenges. *Artif Intell Rev*. 2024; 57: 1–30, doi: 10.1007/s10462-023-10661-7.

[3] Grootendorst M. Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint* 2022; 1–10, doi: 10.48550/arXiv.2203.05794.

[4] Aouichaty S, Maleh Y, Mohtadi MT, et al. Sustainable topic modeling for legal Moroccan Arabic language: A challenging study on bertopic technique. *Procedia Comput Sci* 2024; 236: 582–588, doi: 10.1016/j.procs.2024.05.069.

[5] Ananta M, Utari R, Akhyar A, et al. Integrating bertopic and large language models for thematic identification of Indonesian legal documents. In: *11th International Conference on Advanced Informatics: Concept, Theory and Application, ICAICTA 2024*. IEEE, 2024, pp. 1–6, doi: 10.1109/ICAICTA63815.2024.10762984.

[6] Kirilenko AP, Stepchenkova S. Facilitating topic modeling in tourism research: Comprehensive comparison of new ai technologies. *Tour Manag* 2025; 106: 1–14, doi: 10.1016/J.TOURMAN.2024.105007.

[7] Invernici F, Curati F, Jakimov J, et al. Capturing research literature attitude towards sustainable development goals: an LLM-based topic modeling approach. *J Big Data* 2025; 12: 1–28, doi: 10.1186/s40537-025-01189-4.

[8] Yang C, Kim Y. Enhancing topic coherence and diversity in document embeddings using LLMs: A focus on BERTopic. *Expert Syst Appl* 2025; 281: 1–14, doi: 10.1016/j.eswa.2025.127517.

[9] Aamir N, Raza A, Iqbal MW, et al. Topic modeling empowered by a Deep Learning framework integrating BERTopic, XLM-R, and GPT. *Journal of Computing and Biomedical Informatics* 2025; 8: 1–18, doi: 10.56979/802/2025.

[10] Ma L, Chen R, Ge W, et al. AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. *Exp Biol Med* 2025; 250: 1–11, doi: 10.3389/ebm.2025.10389.

[11] Rijcken E, Kaymak U, Scheepers F, et al. Topic modeling for interpretable text classification from EHRs. *Front Big Data* 2022; 5: 1–11, doi: 10.3389/fdata.2022.846930.

[12] Khodeir N, Elghannam F. Efficient topic identification for urgent MOOC Forum posts using BERTopic and traditional topic modeling techniques. *Educ Inf Technol (Dordr)* 2025; 30: 5501–5527, doi: 10.1007/s10639-024-13003-4.

[13] Mertala P, López-Pernas S, Vartiainen H, et al. Digital natives in the scientific literature: A topic modeling approach. *Comput Human Behav* 2024; 152: 1–12, doi: 10.1016/J.CHB.2023.108076.

[14] Bai X, Zhang X, Li KX, et al. Research topics and trends in the maritime transport: A structural topic model. *Transp Policy (Oxf)* 2021; 102, doi: 10.1016/j.tranpol.2020.12.013.

[15] Abel D, Lieth J, Jünger S. Mapping the spatial turn in social science energy research: A computational literature review. *Renewable and Sustainable Energy Reviews* 2024; 201: 1–16, doi: 10.1016/J.RSER.2024.114607.

[16] Margherita EG, Escobar SD, Esposito G, et al. Exploring the potential impact of smart urban technologies on urban sustainability using structural topic modelling: evidence from Belgium. *Cities* 2023; 141, doi: 10.1016/j.cities.2023.104475.

[17] Guler N, Kirshner SN, Vidgen R. A literature review of artificial intelligence research in business and management using machine learning and ChatGPT. *Data Inf Manag* 2024; 8: 1–25, doi: 10.1016/J.DIM.2024.100076.

[18] Vanhala M, Lu C, Peltonen J, et al. The usage of large data sets in online consumer behaviour: A bibliometric and computational text-mining-driven analysis of previous research. *J Bus Res* 2020; 106: 46–59, doi: 10.1016/J.JBUSRES.2019.09.009.

[19] Grljević O. Topic modeling in hospitality and tourism research: Application areas, business insights, and managerial implications. *Menadzment u hotelijerstvu i turizmu* 2025; 13: 137–153, doi: 10.5937/menhottur2500010g.

[20] Das P, Mandal S, Nedungadi P, et al. Unveiling sustainable tourism themes with machine learning based topic modeling. *Discover Sustainability* 2025; 6: 1–28, doi: 10.1007/s43621-025-01065-4.

[21] Kamal M, Romani G, Ricciuti G, et al. Exploring digital tourism through topic models: A review and experimental study. *Journal of Data Science and Intelligent Systems* 2026; 4: 137–154, doi: 10.47852/bonviewJDSIS62024472.

[22] El-Gayar O, Al-Ramahi M, Wahbeh A, et al. A comparative analysis of the interpretability of LDA and LLM for topic modeling. In: *30th Americas Conference on Information Systems*. Research & Publications, pp. 1–10.

[23] Deerwester S, Dumais ST, Furnas GW, et al. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 1990; 41: 391–407, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.
[24] Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. *Frontiers in Sociology* 2022; 7: 1–16, doi: 10.3389/fsoc.2022.886498.
[25] Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 2003; 3: 993–1022, doi: 10.5555/944919.944937.
[26] El-Gayar O, Al-Ramahi M, Wahbeh A, et al. A comparative analysis of the interpretability of LDA and LLM for topic modeling. In: *30th Americas Conference on Information Systems*. Research & Publications, pp. 1–10.
[27] Hofmann T. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*; 1999, doi: 10.1145/312624.312649.
[28] Lee DD, Seung HS. Algorithms for Non-Negative Matrix Factorization. *Adv Neural Inf Process Syst* 2001; 13: 556–562.
[29] Chen Y, Zhang H, Liu R, et al. Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowl Based Syst* 2019; 163: 1–13, doi: 10.1016/j.knosys.2018.08.011.
[30] Roberts ME, Stewart BM, Tingley D. stm: An R package for structural topic models. *J Stat Softw* 2019; 91: 1–40, doi: 10.18637/jss.v091.i02.
[31] Srivastava A, Sutton C. Autoencoding variational inference for topic models. In: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. 2017, pp. 1–12, doi: 10.48550/arXiv.1703.01488.
[32] Wang W, Guo B, Shen Y, et al. Neural labeled LDA: a topic model for semi-supervised document classification. *Soft comput* 2021; 25: 14561–14571, doi: 10.1007/s00500-021-06310-2.
[33] Corey A, Oh A, Chen S, et al. Evaluating topic model interpretability from a primary care physician perspective. *Comput Methods Programs Biomed* 2016; 124: 67–75, doi: 10.1016/J.CMPB.2015.10.014.
[34] Candelon B, Joëts M, Mignon V. What makes econometric ideas popular: The role of connectivity. *Res Policy* 2024; 53: 105025, doi: 10.1016/J.RESPOL.2024.105025.
[35] Ding K, Niu Y, Choo WC. The Evolution of Airbnb Research: A Systematic Literature Review Using Structural Topic Modeling. *Heliyon* 2023; 9,doi: 10.1016/j.heliyon.2023.e17090.
[36] Dwivedi YK, Sharma A, Rana NP, et al. Evolution of artificial intelligence research in technological forecasting and social change: Research topics, trends, and future directions. *Technol Forecast Soc Change* 2023; 192: 1–7, doi: 10.1016/j.techfore.2023.122579.
[37] Gao B, Zhu M, Liu S, et al. Different voices between airbnb and hotel customers: An integrated analysis of online reviews using structural topic model. *Journal of Hospitality and Tourism Management* 2022; 51, doi: 10.1016/j.jhtm.2022.03.004.
[38] Hu N, Zhang T, Gao B, et al. What do hotel customers complain about? Text analysis using structural topic model. *Tour Manag* 2019; 72, doi: 10.1016/j.tourman.2019.01.002.
[39] Kraus S, Kumar S, Lim WM, et al. From moon landing to metaverse: Tracing the evolution of technological forecasting and social change. *Technol Forecast Soc Change* 2023; 189, doi: 10.1016/j.techfore.2023.122381.
[40] Sharma A, Rana NP, Nunkoo R. Fifty years of information management research: A conceptual structure analysis using structural topic modeling. *Int J Inf Manage* 2021; 58, doi: 10.1016/j.ijinfomgt.2021.102316, doi: 10.1016/j.ijinfomgt.2021.102316.
[41] Villa-Cox G, Bucaram-Villacís SJ, Goethals P, et al. Mapping linkages between ecosystem services and agricultural and landscape management using structural topic modeling of scientific literature. *Ecosyst Serv* 2023; 61, doi: 10.1016/j.ecoser.2023.101525.
[42] Xie J, Jain T. Unpacking micro-csr through a computational literature review: An identity heterogeneity view of internal stakeholders. *J Bus Res* 2024; 172: 114451, doi: 10.1016/J.JBUSRES.2023.114451.
[43] Kaufmann A. Theory of expertons and fuzzy logic. *Fuzzy Sets Syst* 1988; 28: 295–304, doi: 10.1016/0165-0114(88)90036-X.
[44] Kaufmann AJ. Expert appraisements and counter-appraisements with experton processes. In: *Proceedings of ISUMA 1990 - 1st International Symposium on Uncertainty Modeling and Analysis*. College Park, MD, USA, 1990, pp. 619–624, doi: 10.1109/ISUMA.1990.151326.
[45] González EL, Cuervo CM. The aggregation linguistic information for decision making based on expertons and 2-tuples. *Fuzzy Economic Review* 2003; 8: 39–53.
[46] Pérez-Romero ME, Flores-Romero MB, Alfaro-García VG. Tourism competitiveness theory perspectives through the application of counter-expertise methods. In: León-Castro E, Blanco-Mesa F, Alfaro-García V, et al. (eds) *Soft Computing and Fuzzy Methodologies in Innovation Management and Sustainability. Lecture Notes in Networks and Systems*. Springer, 2022, pp. 183–196, doi: 10.1007/978-3-030-96150-3_13.

[47] Mihajlov T, Nešić MI, Stanković R, et al. Topic modeling of the SrpELTeC corpus: A comparison of NMF, LDA, and BERTopic. In: *Annals of Computer Science and Intelligence Systems*. 2024. doi: 10.15439/2024F1593.
[48] Díaz-Pacheco Á, Guerrero-Rodríguez R, Álvarez-Carmona M, et al. A comprehensive deep learning approach for topic discovering and sentiment analysis of textual information in tourism. *Journal of King Saud University - Computer and Information Sciences* 2023; 35: 1–17, doi: 10.1016/j.jksuci.2023.101746.
[49] Wang B, Ge J, Lu F. From review filtering to topic modeling: a deep learning-based approach for analyzing negative online reviews of five-star hotels. *Journal of Hospitality and Tourism Technology* 2025; 1–25, doi: 10.1108/JHTT-04-2025-0328.
[50] Soria EL. TOURSKAD. 25 March 2025, doi: 10.34810/data2133.
[51] Terragni S, Fersini E, Galuzzi B, et al. OCTIS: Comparing and optimizing topic models is simple! In: *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the System Demonstrations*. 2021, pp. 263–70, doi: 10.18653/v1/2021.eacl-demos.31.
[52] Warner B, Chaffin A, Clavié B, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. 2024; 1–20, doi: 10.18653/v1/2025.acl-long.127.
[53] Lau JH, Newman D, Baldwin T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*. Gothenburg, Sweden: Association for Computational Linguistics, 2014, pp. 530–539, doi: 10.3115/v1/e14-1056.
[54] Chang J, Boyd-Graber J, Gerrish S, et al. Reading tea leaves: How humans interpret topic models. In: Bengio Y, Schuurmans D, Lafferty J, et al. (eds) *Advances in Neural Information Processing Systems 22 (NeurIPS 2009)*. Vancouver, Canada: Curran Associates, Inc., pp. 288–296, doi: 10.5555/2984093.2984126.
[55] Butploy N, Khiewwan K, Thongchai J, et al. Multi-transformer-based ensemble embedding model for enhanced vector search in NoSQL Database: A comparative statistical and performance analysis. *International Journal of Mathematical, Engineering and Management Sciences* 2025; 10: 1860–1879, doi: 10.33889/IJMEMS.2025.10.6.086.
[56] Verga P, Hofstatter S, Althammer S, et al. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint*, http://arxiv.org/abs/2404.18796 (2024).
[57] Ye J, Wang Y, Huang Y, et al. Justice or prejudice? Quantifying biases in LLM-as-a-Judge. In: *International Conference on Learning Representations (ICLR 2025)*. OpenReview.net, pp. 1–35, doi: 10.48550/arXiv.2410.02736.
[58] Panickssery A, Bowman SR, Feng S. LLM evaluators recognize and favor their own generations. In: *Advances in Neural Information Processing Systems (NeurIPS 2024).*, pp. 68772–68802, doi: 10.5555/3737916.3740113.
[59] Gu J, Jiang X, Shi Z, et al. A survey on LLM-as-a-Judge. *The Innovation* 2026; 7: 1–30, doi: 10.1016/j.xinn.2025.101253.
[60] Krippendorff K. *Content analysis: An introduction to its methodology*. 2nd ed. Thousand Oaks: SAGE Publications, Inc., 2004, doi: 10.4135/9781071878781.
[61] Mu Y, Dong C, Bontcheva K, et al. Large language models offer an alternative to the traditional approach of topic modelling. In: Calzolari N, Kan M-Y, Hoste V, et al. (eds) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia, pp. 10160–10171, doi: /10.48550/arXiv.2403.16248.
[62] Rahman MJ, Rahman S, Rahman MJ. Decoding trust: LLM in provider topic modeling. In: *30th Americas Conference on Information Systems, AMCIS 2024*, pp. 1–0.
[63] Uğuz S, Tülü ÇN. Topic Modeling Analysis in the field of Large Language Models with BERTopic (2020-2024). In: *2024 Innovations in Intelligent Systems and Applications Conference, ASYU 2024*. IEEE, 2024, pp. 1–6, doi: 10.1109/ASYU62119.2024.10757169.
[64] Sai MN, Royal NL, Chakravarthi N, et al. Automatic topic identification in destination reviews using hybrid topic modelling techniques. In: *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024;* 2024, doi: 10.1109/ICCCNT61001.2024.10726104.
[65] Gao X, Lin Y, Li R, et al. Enhancing topic interpretability for neural topic modeling through topic-wise contrastive learning. In: *IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024, pp. 584–597, doi: 10.1109/ICDE60146.2024.00051.
[66] Liu Y, Iter D, Xu Y, et al. G-EVAL: NLG evaluation using GPT-4 with better human alignment. In: Houda Bouamor, Juan Pino, Kalika Bali (eds) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 2511–2522, doi: 10.18653/v1/2023.emnlp-main.153.

[67] Sai MN, Royal NL, Chakravarthi N, et al. Automatic topic identification in destination reviews using hybrid topic modelling techniques. In: *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*. IEEE, 2024, pp. 1–8, doi: 10.1109/ICCCNT61001.2024.10726104.
[68] Wang X, Feng S, Li Y, et al. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning. In: Luis Chiruzzo, Alan Ritter, Lu Wang (eds) *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, 2025, pp. 6919–6932, doi: 10.48550/arXiv.2408.13457.
[69] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. In: *11th International Conference on Learning Representations, ICLR 2023*. 2023, pp. 1–24, doi: 10.48550/arXiv.2203.11171.
[70] Li Y, Yuan P, Feng S, et al. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In: *12th International Conference on Learning Representations, ICLR 2024*. 2024, pp. 1–18, doi: 10.48550/arXiv.2401.10480.