From Documents to Segments: A Contextual Reformulation for Topic Assignment

Anonymous ACL submission

Abstract

Traditional topic modeling assigns a single topic into each document. In practice, however, many real-world documents, such as product reviews or open-ended survey responses, contain multiple distinct topics. This mismatch often leads to *topic contamination*, where unrelated themes are merged into a single topic, making it difficult to identify documents that truly focus on a specific subject. We address this issue by introducing **segment-based topic allocation (SBTA)**, a reformulation of topic modeling that assigns topics not to entire documents, but to **segments**: short, coherent spans of text that each express a single theme. By modeling topical structure at the segment level, our approach yields cleaner and more interpretable topics and better supports analysis of multi-theme documents. To support systematic evaluation, we construct a **SemEval-STM**, a new dataset inspired by aspect-based sentiment analysis. Documents are first decomposed into topical segments using large language models (LLMs), followed by human refinement to ensure segment quality. We also propose a segment-level extension of the word intrusion task, enabling human evaluation of topical coherence at the granularity where topics are actually assigned. Across multiple models and evaluation metrics, we show that SBTA improves clustering quality and interpretability. Overall, this work provides a practical, scalable framework for fine-grained topic analysis in heterogeneous text corpora where documents naturally span multiple topics.

1 Introduction

Topic modeling is a core technique for discovering latent themes in a text corpus. Classical approaches, such as Latent Dirichlet Allocation (LDA; Blei et al., 2003), represent each topic as a distribution over words (*e.g.*, `stock`, `interest` for an `economics` topic) and each document as a mixture of these topics (*e.g.*, `politics` and `economics` topics within a news article). More recent approaches leverages large language models (LLMs) to generate and validate topic representations, often improving topic quality and interpretability (Stammbach et al., 2023; Pham et al., 2024).

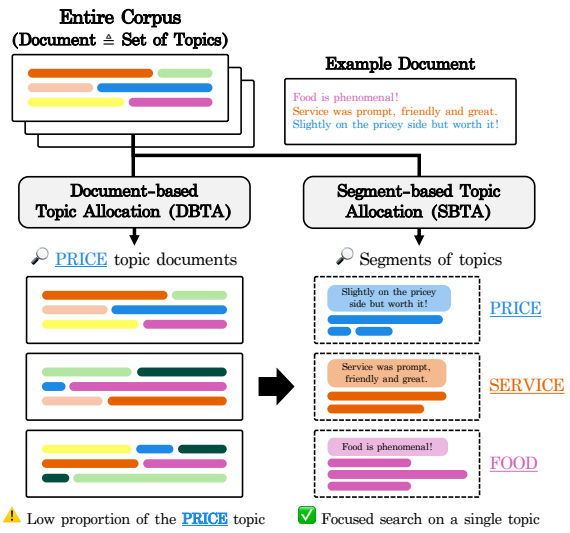

Figure 1: Segment-based topic allocation (SBTA) assigns each topic to a specific textual unit ("segment"), thereby improving interpretability and facilitating its effective application across diverse downstream tasks.

Despite these advances, most topic modeling methods share a fundamental assumption: the **document** is treated as the basic unit of topical coherence. This assumption conflicts in many real-world settings where documents naturally span multiple, heterogeneous topics. For instance, a single employee opinion survey response may discuss `compensation`, `workplace culture`, and `career growth` within a single paragraph. In such case, document-level topic models struggle to isolate content related to a specific theme without interference from unrelated topics.

We refer to this limitation as ***topic contamination***: inferred topics are diluted by off-topic content because each document typically contains multi-

ple themes. Topic contamination degrades interpretability and undermines practical downstream tasks such as topic summarization or automatic topic labeling (Kozlowski et al., 2024; Wanna et al., 2024). As illustrated in Figure 1 (left), document-based topic allocation (DBTA) retrieves entire documents that contain only a small fraction of the topic of interest, making it difficult to extract focused insights.

To address this problem, we propose **segment-based topic allocation (SBTA)**, a reformulation of topic modeling that changes the unit of topic assignment. Instead of associating topics with entire documents (*i.e.*, DBTA), SBTA operates over *segments*—short, self-contained spans of text (*e.g.*, sentences or clauses) that each express a coherent idea. As shown in Figure 1 (right), this formulation enables the model to retrieve and cluster only the segments relevant to given topic (*e.g.*, price-related statements), improving topical precision and interpretability.

Our work is partly inspired by recent LLM-based approaches such as TopicGPT (Pham et al., 2024), which demonstrate that LLMs can effectively identify topics and supporting evidence segments within documents. However, this treats segments as auxiliary explanations for document-level topics. In contrast, we **elevate the segment to the primary unit of topic representation and inference**, formally redefining the task of topic modeling itself. This shift from DBTA to SBTA offers multiple advantages:

- **Improved topic purity.** By assigning topics only to semantically focused segments, SBTA filters out unrelated content that would otherwise contaminate document-level topics (Section 3.2.3).

- **Fine-grained interpretability.** Segments represent complete semantic units, making topic assignments easier to interpret and inspect than document-level mixtures (Figure 1).

- **Better alignment with practical use cases.** Many real-world workflows require retrieving or analyzing specific statements about a topic (*e.g.*, searching only `compensation`-related feedback in survey responses), rather than entire multi-topic documents.

To evaluate the SBTA approach, we introduce a new dataset, **SemEval-STM** (Section 3.2), inspired by aspect-based sentiment analysis corpora (Pontiki et al., 2016). Unlike conventional topic modeling dataset, SemEval-STM provides segment-level annotations, enabling fine-grained evaluation of topic assignments. We formalize the SBTA task (Section 3.1) and empirically show that segments in SemEval-STM exhibit strong topic clustering properties (Section 3.2.3). In addition, we extend the standard word intrusion task (Chang et al., 2009) to the segment level, providing a human-centered evaluation that better captures contextual coherence (Section 3.3). Finally, we benchmark a wide range of topic modeling methods under the SBTA formulation (Section 4), offering insights for both future research directions and practical deployment.

2 Related Works

2.1 Evolution of Topic Modeling Approaches

Topic modeling aims to uncover latent themes within documents, with LDA (Blei et al., 2003) being a seminal method. Although effective, LDA often suffers from interpretability issues and requires manual labeling (Mei et al., 2007; Chang et al., 2009; Baden et al., 2021). Various extensions have been proposed to improve coherence and scalability, including seeded and hierarchical models (Andrzejewski and Zhu, 2009; Teh et al., 2006), as well as neural topic models. (Srivastava and Sutton, 2017; Dieng et al., 2020). Clustering-based approaches such as BERTopic (Grootendorst, 2022) leverage pre-trained sentence embeddings to construct topic models.

More recently, LLM-based approaches have introduced prompt-driven topic modeling strategies. TopicGPT (Pham et al., 2024) and others (Doi et al., 2024) utilize LLMs to generate and assign topics at the document level, often referring to segment-level content for interpretability. However, in these methods, segments are treated only as explanatory evidence rather than being incorporated as formal units of topic representation.

2.2 Topic Modeling Evaluation

Evaluating topic models involves both human and automatic approaches. Early studies introduced word intrusion tasks to assess interpretability (Chang et al., 2009; Newman et al., 2010; Mimno et al., 2011). Although these align closely with human judgment, they are costly and difficult to scale.

Metrics such as UCI (Newman et al., 2010), UMass (Mimno et al., 2011), NPMI (Lau et al., 2014), and C_v (Röder et al., 2015) measure word-level consistency based on co-occurrence statistics, later extended with embedding-based semantic similarity (Nikolenko, 2016; Ramrakhiyani et al., 2017; Korenčić et al., 2018). However, most coherence metrics remain word-centric and often fail to capture document-level or contextual coherence.

Recent research leverages LLMs for evaluating topic coherence and topic intrusions (Stammbach et al., 2023; Rahimi et al., 2024). These approaches treat LLMs as proxies for human annotators, achieving greater scalability and interpretability. Yet, most of these methods remain limited to evaluating the quality of topic words alone.

2.3 Relation with Topic Segmentation

Topic segmentation identifies points of topical shift within a document to divide it into coherent sections. Early methods such as TextTiling (Hearst, 1994) exploited inter-sentence similarity to detect topic transitions, while TopicTiling (Riedl and Biemann, 2012) extended this approach by incorporating probabilistic topic assignments from LDA to infer segment boundaries. Recent approaches have adopted embeddings and LLM for improved contextual sensitivity (Fan et al., 2024; Ghinassi, 2021). Although topic segmentation and our framework both involve textual segments, their objectives are fundamentally different. While topic segmentation focuses on structural partitioning, our segment-based topic modeling treats segments as the analytical unit for deriving cleaner and more faithful topic models, leveraging semantically coherent spans to enhance topic purity, interpretability, and contextual alignment across the corpus.

3 Segment-based Topic Allocation

3.1 Task Definition

3.1.1 Basic Notations

We define the vocabulary as a set of V distinct words, indexed by $\{1, \ldots, V\}$. A corpus is denoted by $\mathcal{D} = \{d_1, \ldots, d_D\}$, where each document $d \in \mathcal{D}$ is represented as a sequence of N_d word tokens. We assume the existence of K latent topics, indexed by $\{1, \ldots, K\}$.

3.1.2 Classical Topic Modeling

In Latent Dirichlet Allocation (LDA; Blei et al., 2003), the **topic-word distributions** are represented by probability vectors $\phi_k = (\phi_{k,1}, \ldots, \phi_{k,V})$ for each topic k, where $\phi_k \in \Delta^{V-1}$. Here, Δ^{V-1} denotes the $(V-1)$-dimensional probability simplex, ensuring that each ϕ_k is a valid probability distribution over the vocabulary (*i.e.*, $\phi_{k,v} \geq 0$ for all v, and $\sum_{v=1}^{V} \phi_{k,v} = 1$). The entry $\phi_{k,v}$ corresponds to the probability $P = \{w = v | z = k\}$, where $w \in \{1, \ldots, V\}$ denotes a word index from the vocabulary, and $z \in \{1, \ldots, K\}$ is a latent topic assignment.

Likewise, the **document-topic distribution** for each document d is parameterized by $\theta_d = (\theta_{d,1}, \ldots, \theta_{d,K}) \in \Delta^{K-1}$, where $\theta_{d,k} = P(z = k|d)$ indicates the probability of topic k appearing in document d. Both ϕ_k and θ_d are drawn from Dirichlet priors, ensuring proper probabilistic constraints.

3.1.3 Definition of Segment

Building on the classic topic-word and document-topic formalisms, we introduce a **topic–segment distribution** that admits *multi-topic* phrases. Unlike traditional topic models such as LDA, which assign a single topic to each word token independently, our model introduces span-level structure by grouping contiguous tokens into segments that may jointly express one or more related topics.

Formally, a segment in document d is a pair $\big([i{:}j],\, \mathcal{T}\big)$ defined as:

$$[i{:}j] = \{\, i,\, i+1,\, \ldots,\, j \,\}, \quad 1 \leq i \leq j \leq N_d,$$
$$\mathcal{T} \subseteq \{1, \ldots, K\},\ \mathcal{T} \neq \varnothing$$

In other words, a segment is a **maximal contiguous span of tokens** whose inferred topic labels are entirely contained within a set $\mathcal{T}$. This reflects a key intuition: speakers or writers tend to discuss related topics in compact phrases, such as "*I love the price and quality*," where $\mathcal{T} = \{\text{PRICE}, \text{QUALITY}\}$, rather than interleaving many unrelated topics within a short segment. In contrast to prior work such as Arnold et al. (2019), which segments documents into coarser, multi-sentence units, our segment unit yields more fine-grained topical structure.

Per-document and Corpus-level Collection: We gather all segments in document d as:

$$\mathcal{Q}_d = \{\, Q_d^{(1)}, Q_d^{(2)}, \ldots \},$$
$$Q_d^{(m)} = \big([i_m{:}j_m],\, \mathcal{T}_m\big).$$

When a topic-specific view is needed (*e.g.*, if someone is only interested in segments that mention the `price` topic), we define $\mathcal{Q}_{d,k} = \{ Q \in \mathcal{Q}_d \mid k \in \mathcal{T}(Q) \}$. The **segment set** for the entire corpus is then given by $\mathcal{Q} = \bigcup_{d \in \mathcal{D}} \mathcal{Q}_d$.

Topical Sparsity: Although a document may span a broad mixture of topics, each segment typically focuses on only a small subset. Formally, for a segment $Q = ([i:j], \mathcal{T})$, we usually have $1 \leq |\mathcal{T}| \ll |\operatorname{supp}(\boldsymbol{\theta}_d)|$, where $\operatorname{supp}(\boldsymbol{\theta}_d)$ denotes the set of active topics in document d. In practice, $|\mathcal{T}|$ rarely exceeds two or three, which aligns with the intuition that most phrases express a tightly coupled semantic focus, rather than blending many disparate themes.

3.2 Dataset: SemEval-STM

To evaluate our reformulated framework, we construct **SemEval-STM**[1] dataset by modifying an aspect-based sentiment analysis task that offers a natural testbed for segment-based topic modeling tasks. While conventional topic modeling datasets (Hoyle et al., 2022; Merity et al., 2018) treat documents as the atomic unit of topical coherence and topics as latent and unannotated, the aspect-based datasets explicitly associate textual spans with predefined *aspects*, interpretable semantic categories such as `service`, `food`, `price` in product reviews. These aspects serve a role analogous to topics in topic modeling, representing interpretable semantic categories that organize document content. Moreover, the aspect labels themselves provide weak supervision that can guide the extraction of segments. By treating aspects as proxy topics, we can prompt an LLM to identify segments in the document that correspond to each aspect. This makes aspect-based datasets a practical and meaningful resource for benchmarking SBTA under semi-supervised or weakly supervised settings. In practice, we employ two domains from SemEval-2016 ABSA dataset[2] (Pontiki et al., 2016): *laptop* and *restaurant*.

We design two experimental setups, DBTA (conventional formulation) and SBTA (proposed formulation), on top of this dataset to evaluate the feasibility of our segment-based formulation. Adhering to conventional definition, DBTA setup directly utilizes existing annotations from the SemEval-2016 ABSA dataset, as each document is associated with multiple verified aspects, which serve as topics in our framework. In subsequent sections, we provide a generation and verification pipeline to construct SemEval-STM suitable for SBTA.

While several existing benchmarks provide topic annotations at the document level, they are limited in their ability to faithfully evaluate segment-based topic allocation (SBTA). In datasets such as Wiki and Bills (Hoyle et al., 2022; Merity et al., 2018), documents often contain a large proportion of content unrelated to the assigned main topic. In such settings, SBTA may appear trivially advantageous, as isolating a small number of relevant segments from largely off-topic documents is inherently easier. To enable a fairer comparison, we instead consider a regime where multiple topics are interleaved but topic-irrelevant content does not dominate, allowing the benefits of SBTA to be evaluated more conservatively. Based on this rationale, we construct a new dataset with explicit segment-level topic annotations.

3.2.1 Segment Generation

To instantiate environmental datasets for SBTA in practice, we leverage the generative capabilities of LLMs. For each topic $k \in \{1, \ldots, K\}$ and each document $d \in \mathcal{D}$, we query an LLM to identify a set of segments $\mathcal{Q}_{d,k} \subset \mathcal{Q}_d$, where each segment $Q = ([i:j], \mathcal{T})$ satisfies $k \in \mathcal{T}$. That is, we extract maximal contiguous spans within d that are topically coherent and relevant to topic k. We use `o3-mini` in our generation process with the prompt provided in Section D.1.

3.2.2 Postprocessing and Refinement

Based on the segment allocation results of LLM, we discard topics that are associated with fewer than 10 segments. At this stage, 43 topics in the laptop domain are removed, while all topics in the restaurant domain are retained, resulting in a final set of 33 topics for the laptop domain and 12 for the restaurant domain. We also removed the same topic in DBTA setup for the fair comparison. Subsequently, the authors manually reassign a subset of topics by examining the topic-segment pairs. For instance, we reallocate items in `LAPTOP#GENERAL` topic into `LAPTOP#QUALITY` topic since most segments are indicating the quality of the laptop. Section A.1 provides postprocessed examples, and Section A.2 illustrates the topic distribution of DBTA and SBTA for each domain.

[1] Standing for Segment-based Topic Modeling.

[2] Creative Commons Attribution-NonCommercial -ShareAlike 3.0 Unported (CC BY-NC-SA 3.0).

	NPMI (↑)	UMass (↑)	UCI (↑)	C_v (↑)	DB Index (↓)	CH Index (↑)	MB Score (↑)	Silhouette (↑)	XB Index (↓)	XB Star (↓)
Domain: Laptop										
DBTA	-0.0094	-1.4591	-0.3159	0.3984	20.1768	3.0037	0.0007	-0.0522	95.8645	103.8339
DBTA (S)	**-0.0166** (± 0.0027)	-1.4129 (± 0.0123)	-0.2576 (± 0.0314)	**0.3919** (± 0.0096)	23.0329 (± 1.5234)	**0.9943** (± 0.0225)	**0.0002** (± 5.4772e-05)	**-0.0482** (± 0.0120)	**131.2557** (± 17.5548)	**139.1088** (± 17.3878)
SBTA	-0.1626	-11.2192	-6.9539	0.3109	6.2767	15.5184	0.0017	0.046	10.8348	12.4061
SBTA (S)	**-0.1920** (± 0.0056)	-9.5768 (± 0.1696)	-5.7639 (± 0.1844)	**0.2862** (± 0.0075)	**28.1171** (± 2.1846)	**1.0172** (± 0.0607)	**0.0003** (± 0.0)	**-0.0268** (± 0.0045)	**197.1178** (± 31.0989)	**203.2193** (± 30.3041)
Domain: Restaurant										
DBTA	-0.0034	-1.449	-0.4289	0.3581	70.9506	1.7204	0.001	-0.0303	1233.5519	1264.0482
DBTA (S)	**-0.0044** (± 0.0077)	**-1.5501** (± 0.0805)	-0.2880 (± 0.1587)	**0.3484** (± 0.0151)	26.6599 (± 1.1914)	**0.9583** (± 0.0412)	**0.0005** (± 3.7270e-05)	-0.0235 (± 0.0025)	176.4181 (± 15.8537)	180.9349 (± 15.0420)
SBTA	-0.2376	-12.2595	-8.0276	0.3508	6.6657	22.6709	0.005	0.0222	12.1985	13.5955
SBTA (S)	-0.2003 (± 0.0261)	-9.6084 (± 0.7104)	-5.8915 (± 0.6971)	**0.2926** (± 0.0171)	**30.8520** (± 2.9719)	**1.0148** (± 0.0474)	**0.0007** (± 5.5902e-05)	**-0.0179** (± 0.0031)	**238.1718** (± 43.5951)	**241.2051** (± 44.0460)

Table 1: **Results of the document/segment shuffle test in SemEval-STM.** Shuffled results (with mean and standard deviation for 5 repetitions) are marked with (S). The first four metrics assess coherence based on word frequency, while the remaining metrics evaluate clustering performance based on **all-MiniLM-L6-v2** embedding. Arrow (↑ and ↓) denote whether higher or lower values indicate better performance, respectively. Metric scores that are negatively affected by the shuffling of documents or segments, consistent with expectations, are highlighted in **bold**.

3.2.3 Preliminary Experiments

To validate the feasibility of SemEval-STM, we conduct two key experiments: (1) a comparison between SBTA and DBTA on their original topic allocation quality, and (2) a sensitivity test by applying random shuffling to each method's topic assignments.

We evaluate clustering quality using six standard clustering metrics: **DB Index**, **CH Index**, **MB Score**, **Silhouette**, **XB Index**, and **XB Star**. All metrics are computed using `all-MiniLM-L6-v2` embeddings[3], and for the topic shuffling task, results are averaged over five runs.

(1) Topic allocation quality comparison: We first compare the original clustering quality of SBTA and DBTA without any perturbation. As shown in Table 1, **SBTA achieves higher clustering metric scores than DBTA across all metrics and domains**, suggesting that segment-level topic allocation produces more tightly clustered topic groups when evaluated at the segment granularity. This empirical observation aligns with the core motivation of SBTA: allocating topics to semantically coherent spans (segments) rather than entire documents can yield improved topic purity and structural clarity within this reformulated task framework.

(2) Sensitivity to Topic Shuffling: To evaluate how well-structured the original topic assignments are, we introduce a perturbation-based test: topic items (*i.e.*, documents or segments) are randomly shuffled across topics. The underlying assumption is that if the original assignments are semantically meaningful and topologically well-formed, shuffling should significantly degrade clustering performance by disrupting their internal structure.

For example, when documents related to "price" and "design" are originally well-separated, randomly reassigning them to incorrect clusters leads to semantic incoherence (*e.g.*, placing a segment about pricing into the design cluster). Conversely, if the original topic structure is already noisy or loosely defined, such random shuffling introduces minimal additional disorder.

In this setting, the performance degradation from **SBTA** to the shuffled version, marked with **SBTA (S)**, is substantially larger than from **DBTA** to **DBTA (S)**.[4] This demonstrates SBTA's higher sensitivity due to its inherently more coherent and structured topic clusters. In contrast, DBTA experiences only marginal degradation, sometimes even showing improvements, reflecting a lack of distinct topical structure rather than genuine robustness.

Taken together, these experiments reveal that **SBTA exhibits stronger clustering metric scores and encodes a finer-grained topic structure compared to DBTA**, one that is more sensitive to disruption when semantic alignment is removed. This suggests that the segment-based reformulation offers distinct advantages for applications requiring

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Consistent results with other embeddings are reported in Appendix C.

[4] This can be seen by comparing rows 3–4 against rows 1–2 in Table 1 for each domain.

fine-grained topical analysis.

On the Limitations of Coherence Metrics: As for the coherence metrics (the first four metrics in Table 1), we observe that neither DBTA nor SBTA exhibits consistent performance degradation under topic item shuffling. We attribute this to a fundamental limitation of coherence metrics: they rely on *co-occurrence statistics between word pairs*. In the case of DBTA, the use of full-length documents often leads to frequent words co-occuring broadly across documents, thereby preserving high pairwise co-occurrence scores even when topic items are shuffled. In contrast, for SBTA, the significantly shorter segments reduce the likelihood of strong word co-occurrence patterns being observed in the first place, making it difficult for coherence metrics to distinguish between coherent and incoherent topics, regardless of shuffling. In sum, this behavior of coherence metrics is expected given the underlying characteristics of both DBTA and SBTA.

Hereafter, we utilize the SBTA version of the SemEval-STM benchmark in our experiments.

3.3 Evaluation: Segment Intrusion Task

We additionally introduce a human-centered intrusion evaluation method at the segment level, inspired by traditional word and topic intrusion tasks (Chang et al., 2009; Bhatia et al., 2018). While previous methods focused on identifying intruder words within topic-word lists, our design extends this approach to the segment level, **made possible by redefining the unit of topic assignment from documents to segments.**

The core objective of the segment intrusion task is to evaluate *"whether a topic has human-identifiable semantic coherence"* (Chang et al., 2009). Specifically, either human annotators or LLMs are asked to identify the most semantically divergent segment among a group of candidates, each supposedly sharing the same topical label. A high success rate indicates that the segments are semantically cohesive, validating the quality of topic assignments.

To systematically control task difficulty, we vary both the **semantic similarity** among candidate segments and **the number of intruders**. In easy conditions, intruders are drawn from different domains (*e.g.*, spotting a restaurant segment among laptop segments), making them easier to identify. In contrast, hard conditions sample intruders from the same domain, demanding more fine-grained discrimination. The number of intruders is also varied to further modulate task complexity. In this work, we adopt four task variants (Single/Double Intruder with Easy/Hard settings), all newly proposed in this work.

Task 1: Single-Intruder-Easy (SI-E). 200 sets with five segments from a single domain and one intruder from a different domain. Identify the one intruder segment.

Task 2: Single-Intruder-Hard (SI-H). 200 sets with five segments and one intruder, all from the same domain. Identify the one intruder segment. Appendix G presents an example of this task.

Task 3: Double-Intruder-Easy (DI-E). 200 sets with four segments from a single domain and two intruders from a different domain. Identify the two intruder segments.

Task 4: Double-Intruder-Hard (DI-H). 200 sets with four segments and two intruders, all from the same domain. Identify the two intruder segments.

4 Experiments

4.1 Benchmarking Task 1: Topic Modeling

To evaluate the SBTA performance on SemEval-STM, we present benchmark topic modeling results across diverse approaches. Specifically, we employ LDA (Blei et al., 2003), BERTopic (Grootendorst, 2022), and LLM-based topic modeling approaches with the prompt demonstrated in Section D.2. We used the MALLET (McCallum, 2002) implementation of LDA with Gibbs sampling and BERTopic with all default hyperparameters, except that the number of topics was set to match that of SBTA. For the LLM-based topic assignment approach, we followed a procedure similar to TopicGPT (Pham et al., 2024). Specifically, each input segment was paired with a predefined set of candidate label topics, and the model was prompted to select the most relevant topic based on its semantic reasoning. To conduct inference, we utilized a diverse set of models, including GPT, Claude, Gemini, Qwen, Llama, and DeepSeek (see Appendix E for technical details).

While we use a predefined topic list in the SemEval-STM for evaluation consistency, real-world deployments may lack such supervision. In those scenarios, the topic modeling process can

	NPMI (↑)	UMass (↑)	UCI (↑)	C_v (↑)	DB Index (↓)	CH Index (↑)	MB Score (↑)	Silhouette (↑)	XB Index (↓)	XB Star (↓)
Domain: Laptop										
LDA	-0.1826	-12.6159	-7.8524	0.3105	6.7873	4.9365	0.0009	-0.0066	10.8622	11.997
BERTopic	-0.1684	-13.9148	-8.217	0.322	**4.5388**	7.4211	0.0034	0.0862	**4.7237**	**5.3453**
`llama-3.2-3b -instruct-turbo`	**-0.0613**	**-9.4438**	**-4.988**	**0.4024**	5.8505	7.9353	**0.0059**	**0.1647**	8.277	10.786
`deepseek-v3`	-0.1505	-11.2648	-6.855	0.3069	6.9381	**10.7089**	0.002	0.0428	10.5334	12.8404
Domain: Restaurant										
LDA	-0.2512	-13.9977	-9.2477	0.3385	7.7317	3.3516	0.002	-0.0072	14.8073	15.9644
BERTopic	**-0.1622**	-13.6309	-8.1259	0.3107	**3.9735**	6.7305	0.0116	0.1027	**3.614**	**4.0662**
`gpt-4o`	-0.1909	**-11.0746**	-6.9674	0.3076	5.63	8.7456	0.0054	0.0221	7.9977	9.1804
`o4-mini`	-0.2271	-12.4864	-8.0232	**0.3451**	5.3975	8.8302	0.0054	0.0239	7.4767	8.3782
`claude-3.7 -sonnet-20250219`	-0.2154	-11.7673	-7.5305	0.3233	5.2393	**9.7834**	0.0059	0.0300	5.9620	7.1544
`gemini-2.5 -flash-preview-04-17`	-0.1933	-11.1135	**-6.9446**	0.3252	5.1914	9.1775	0.0055	0.0256	6.3861	7.3846
`llama-3.2-3b -instruct-turbo`	-0.1981	-11.6759	-7.2913	0.3152	5.9292	6.1641	**0.0117**	**0.0339**	8.2703	9.5635

Table 2: Topic modeling benchmark performance with label-free metrics. Best performance is marked as **bold**. Full results are demonstrated in Table 7.

	Precision (↑)	Recall (↑)	F1 (↑)	Purity (↑)	ARI (↑)	MIS (↑)
Domain: Laptop						
LDA	0.3602	0.3565	0.3577	0.3770	0.1573	0.3344
BERTopic	0.5139	0.5084	0.5102	0.5033	0.2603	0.5262
`llama-3.3-70b -instruct-turbo`	0.7318	0.7213	0.7248	0.7030	**0.4882**	0.6342
`deepseek-v3`	**0.7454**	**0.7347**	**0.7383**	**0.7208**	0.4814	**0.6543**
Domain: Restaurant						
LDA	0.4530	0.4505	0.4512	0.4812	0.1994	0.2397
BERTopic	0.6728	0.6679	0.6692	0.6334	0.4593	0.5190
`claude-3.7 -sonnet-20250219`	**0.8386**	**0.8336**	**0.8353**	**0.8224**	**0.6805**	**0.6943**

Table 3: Topic modeling benchmark performance with label-based metrics. Best performance is marked as **bold**. Full results are demonstrated in Table 8.

be extended via two complementary phases: a topic generation phase (*e.g.*, prompting an LLM to traverse the corpus and extract candidate topic labels), as demonstrated in Pham et al. (2024), and a segment-based assignment phase that allocates each topic not to entire documents but to semantically coherent segments. This design effectively extends the TopicGPT framework into a finer-grained topic model aligned with SBTA.

Regarding evaluation metrics, we utilize all aforementioned label-free metrics, consisting of coherence (Section B.1) and clustering (Section B.2) measures. Additionally, since SemEval-STM provides topic labels for laptop and restaurant domains, we adopt label-based metrics: Purity, ARI, NMI (Section B.3), Precision, Recall, and F1. LDA and BERTopic only assign topics to documents without explicit labels, we map each predicted topic to the ground-truth label with the most overlapping documents to enable label-based evaluation.

4.1.1 Results

For LLM-based performance, we only report models showing either best or second-best performance in the main text. Full model results are provided in Appendix F.

Label-Free Performance: Table 2 demonstrates label-free topic modeling performance across multiple models and metrics (full version is demonstrated in Table 7). The results indicate that **each model exhibits distinct strengths within particular domains, thereby underscoring the heterogeneous challenges posed by SemEval-STM**. For instance, smaller models (*e.g.*, BERTopic, `llama-3.2-3b`) demonstrate superior performance compared to larger models in the laptop domain,

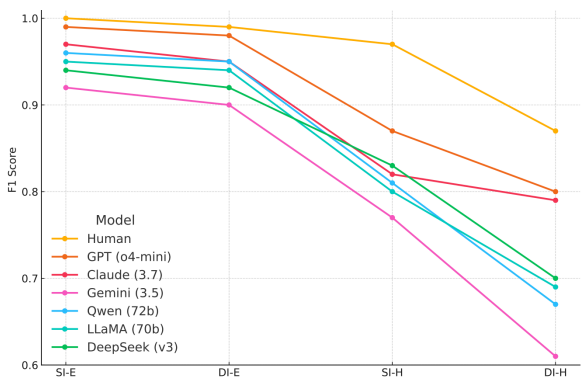

Figure 2: Visualized F1 performance of the segment intrusion evaluation in Laptop domain. Human performance is determined by averaging the annotations of two independent participants. Raw performance metrics are reported in Table 10 in Appendix I, and inter-annotator agreement is provided in Appendix H.

whereas GPT-based variants generally achieve strong performance in the restaurant domain. These results suggest that **smaller models, or specific model families, can be strategically deployed for domain-specific applications.**

Label-Based Performance: Table 3 presents the performance of label-based topic modeling approaches (full version is demonstrated in Table 8). In this setting, larger LLMs, such as `claude-3.7-sonnet`, `llama-3.3-70b`, and `deepseek-v3` exhibit strong performance, whereas traditional methods (*i.e.*, LDA, BERTopic) tend to underperform relative to LLM-based approaches.

4.2 Benchmarking Task 2: Segment Intrusion Task

To evaluate the segment intrusion task, we employ the same models used in the topic modeling evaluation (Section 4.1) to identify intruder segment(s) from a set of candidates, serving as a proxy for human judgment. We use the prompt presented in Section D.3 for SI tasks and Section D.4 for DI tasks. As an upper bound on performance, we assess human performance by randomly sampling 50 instances (out of 200) for each task and averaging the annotations provided by two human participants [5] (*i.e.*, 50*4=200 annotations per participant). For evaluation metrics, we report F1 score in the main text, providing Recall and Precision results in Appendix I.

4.2.1 Results

Figure 2 visualizes F1 scores of all segment intrusion tasks in laptop domain, using the best-performing model from each model family. As visualized by the downward trend result, indicate that both LLMs and human annotators experience greater difficulty with the more challenging tasks (*i.e.*, SI-H and DI-H) compared to the easier ones (*i.e.*, SI-E and DI-E), and that identifying two intruders is generally more difficult than selecting a single intruder. Furthermore, as shown in Figure 7, smaller models (*e.g.*, `llama-3.2-3b`, `qwen-2.5-7b`, and `llama-4-maverick-17b`) struggle considerably with segment intrusion tasks. Even among larger and higher-performing models, most still fall short of human-level performance. These findings show the difficulty of the proposed segment intrusion tasks and provide important future direction to improve model capabilities.

5 Conclusion

We introduce segment-based topic allocation (SBTA), a shift from document-level topic modeling that assigns topics to the segment level. Through extensive experiments on SemEval-STM dataset, we demonstrate that SBTA substantially improves topic purity and interpretability compared to traditional document-based approaches. Specifically, empirical findings indicate that SBTA yields more distinct topic clusters and exhibits greater sensitivity to underlying topic structure. Additionally, we provide benchmarking results of both label-free and label-based topic modeling evaluations, as well as human-aligned evaluation via the segment intrusion task. Notably, LLMs integrated with SBTA consistently outperform traditional methods across both label-free and label-based metrics, highlighting the synergy between fine-grained topic segmentation and advanced language understanding. Nonetheless, we also observe that certain methods with smaller, more efficient model architectures demonstrate superior performance in domain-specific settings. These results validate SBTA as a practical and scalable solution for fine-grained topic analysis in heterogeneous text corpora.

[5] The annotations were performed by two of the paper's authors, who were directly involved in the design of the segment intrusion tasks and were well-acquainted with the annotation criteria. Inter-annotator agreement statistics are reported in Appendix H.

Limitations

While the paper introduces SBTA approach that enhances topic purity and interpretability, several limitations remain. First, the extraction of segments relies heavily on LLMs, introducing potential inconsistencies from automated systems, despite human post-processing. A robust and reliable segment extraction pipeline is required to enable scalable data construction. Furthermore, conventional topic coherence metrics fail to align with SBTA's span-level focus due to reduced word co-occurrence (as analyzed in Section 3.2.3), thereby limiting the effectiveness of standard automated evaluations. Lastly, the current dataset primarily contains relatively short documents, reflecting practical scenarios in which users tend to write concise responses (e.g., survey feedback or short reviews). Nevertheless, our data construction pipeline can be extended to corpora with longer documents, following the same segment-based methodology applied in SemEval-STM.

Ethics Statement

All models used in our experiments are publicly accessible and are applied in accordance with their intended research use and release terms. Model-specific details are provided in Appendix E. The proposed SemEval-STM dataset is derived from an existing benchmark through segment-level processing (Pontiki et al., 2016). This procedure does not introduce new textual content or personally identifiable information. The underlying data consist of short review texts that have been widely used in prior NLP studies and do not involve sensitive or high-risk domains. AI-based tools were used to assist with writing clarity and editing. All methodological decisions, experiments, and interpretations are carried out and validated by the authors.

References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486.

David Andrzejewski and Xiaojin Zhu. 2009. Latent Dirichlet Allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing*, pages 43–48, Boulder, Colorado. Association for Computational Linguistics.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken van der Velden. 2021. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16:1–18.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2018. Topic intrusion for automatic topic model evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 844–849, Brussels, Belgium. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

David L. Davies and Donald W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Tomoki Doi, Masaru Isonuma, and Hitomi Yanaka. 2024. Topic modeling for short texts with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 21–33, Bangkok, Thailand. Association for Computational Linguistics.

Joseph C Dunn. 1974. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2024. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. *Preprint*, arXiv:2305.08391.

Iacopo Ghinassi. 2021. Unsupervised text segmentation via deep sentence encoders: a first step towards a common framework for text-based segmentation, summarization and indexing of media content.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, page 9–16, USA. Association for Computational Linguistics.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Damir Korenčić, Strahil Ristov, and Jan Šnajder. 2018. Document-based topic coherence measures for news media text. *Expert Systems with Applications*, 114:357–373.

Diego Kozlowski, Carolina Pradier, and Pierre Benz. 2024. Generative ai for automatic topic labelling. *Preprint*, arXiv:2408.07003.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Ujjwal Maulik and Sanghamitra Bandyopadhyay. 2003. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*, 24(12):1650–1654.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://www.cs.umass.edu/ mccallum/mallet.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Knowledge Discovery and Data Mining*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations*.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 215–224.

Sergey I. Nikolenko. 2016. Topic quality metrics based on distributed word representations. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 1029–1032, New York, NY, USA. Association for Computing Machinery.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Hamed Rahimi, David Mimno, Jacob Hoover, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2024. Contextualized topic coherence metrics. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1760–1773, St. Julian's, Malta. Association for Computational Linguistics.

Nitin Ramrakhiyani, Sachin Pawar, Swapnil Hingmire, and Girish Palshikar. 2017. Measuring topic coherence through optimal word buckets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 437–442, Valencia, Spain. Association for Computational Linguistics.

Martin Riedl and Chris Biemann. 2012. How text segmentation algorithms gain from topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 553–557, Montréal, Canada. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore. Association for Computational Linguistics.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles - a knowledge reuse framework for combining partitionings. *J Mach Learn Res*, 3.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei and. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1073–1080, New York, NY, USA. Association for Computing Machinery.

Selma Wanna, Nicholas Solovyev, Ryan Barron, Maksim E. Eren, Manish Bhattarai, Kim Ø. Rasmussen, and Boian S. Alexandrov. 2024. Topictag: Automatic annotation of nmf topic models using chain of thought and prompt tuning with llms. In *Proceedings of the ACM Symposium on Document Engineering 2024*, DocEng '24, New York, NY, USA. Association for Computing Machinery.

Xuanli Lisa Xie and Gerardo Beni. 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(08):841–847.

A Details on SemEval-STM Construction

A.1 Examples of DBTA and SBTA

A.1.1 DBTA (Laptop)


```
{
"CPU#OPERATION_PERFORMANCE": [
    "Other than the slow CPU it works great for everyday use. Heavy gaming is
    ↪ definitely not it's strong point. The cover is a soft rubber texture
    ↪ without the friction. Slick design and looks great.",
    "Being a PC user my whole life.... This computer is absolutely AMAZING!!! 10
    ↪ plus hours of battery... super fast processor and really nice graphics
    ↪ card.. and plenty of storage with 250 gb(though I will upgrade this and the
    ↪ ram..) This computer is really fast and I'm shocked as to how easy it is to
    ↪ get used to... I've only had mine a day but I'm already used to it... MACS
    ↪ ARE AMAZING!!! GET THIS COMPUTER FOR PORTABILITY AND FAST PROCESSING!!!",
    "not durable . slow processor, just not it",
    "Great laptop from Apple! Apple is unrivaled in terms of build quality and
    ↪ functionality. The retina screen is absolutely beautiful and the touchpad
    ↪ is the best touchpad to date. The pair allow for seamless navigation
    ↪ throughout the whole interface. Battery life is astonishing given the
    ↪ processing power and high resolution display. I can easily get 10 hours out
    ↪ of a full charge. Also, I find OS X much more simple and easy to use than
    ↪ Windows. Overall, a high quality laptop and a great value considering all
    ↪ this computer has to offer.",
    "This computer is absolutely perfect for web browsing, watching video, word
    ↪ processing, and that's about it. If you want a computer that plays games or
    ↪ any advanced tasks like editing video, this computer is NOT for you. Now
    ↪ when I say don't expect to play games on it, I mean it. Pretty much every
    ↪ major game save for Solitare won't even play on it due to the slower
    ↪ processor. Its also worth noting that there no bluetooth, no mircophone
    ↪ jack, no ethernet, no dvd/cd drive/burner, and no usb 3.0 ports. It DOES
    ↪ come with two usb 2.0 ports, a hdmi port, and a webcam. The webcam is
    ↪ nothing special, its there if you need it. Don't expect it to take good
    ↪ pictures or video. The battery is pretty soild. It lasts up to 5-6 hours
    ↪ with normal use. The computer itself runs very quiet and is mostly cool to
    ↪ the touch. The only major negative to me is Windows 8.1 (though YMMV) and
    ↪ the Dell crapware installed. I would strongly recommend to remove it and
    ↪ install Windows 7 on it with a usb drive. The performance will improve on
    ↪ it. Trust me.",
    "This laptop is the most amazing little peice of machinery I have owned outside
    ↪ of the Iphone. It took me a while top get away from the land of PCs, but now
    ↪ that I have, I can't see myself going back to it. The laptop is gorgeous. It
    ↪ is sleek, smooth, and lightweight. It is easily portable and I take it
    ↪ everywhere I go and might require internet access. The screen is very large
    ↪ and crystal clear with amazing colors and resolution. The keyboard is slick
    ↪ and quiet and not bulky like some other laptops I have had in the past. The
    ↪ processor is very quick and effective as I load webpages and applications.
    ↪ It is quiet and a real joy to watch work. I love it and will probably get
    ↪ another one when this goes to the Laptop in the sky!!",
```

```
    "Ok, this is probably the best laptop series ever devised by Apple. The case is
    ↪ carved out of a single block of aluminum. Although I opted for the lowest
    ↪ end MacBook Pro, this thing holds its own. The processor screams, and
    ↪ because of the unique way that Apple OSX 16 functions, most of the graphics
    ↪ are routed through the hardware rather than the software. That is how it is
    ↪ able to function better than any other PC. I have recommended this laptop
    ↪ to everyone I know who is buying one. I like it so much, I bought another
    ↪ just for my wife.",
    ...
],
...
}
```

A.1.2 SBTA (Laptop)

```
{
"POWER_SUPPLY#OPERATION_PERFORMANCE": [
    "HOW DOES THE POWER SUPPLY NOT WORK!!!",
    "the charger stopped working",
    ...
],

"GRAPHICS#DESIGN_FEATURES": [
    "most of the graphics are routed through the hardware rather than the software",
    "it runs anything that doesn't require a dedicated video card",
    ...
],

"LAPTOP#PRICE": [
    "And I'm still paying the bloody financing",
    "All in all, it's well worth its price tag.",
    "For the price and what I get out of it has exceeded my expectations",
    "Good price.",
    "It's not what you pay for.",
    ...
],

"KEYBOARD#DESIGN_FEATURES": [
    "the keyboard is well designed",
    "the lighted keyboard",
    "the user manual explains how to turn on the keyboard  backlight.",
    "I was disappointed when I realized that the keyboard doesn't light up on this
    ↪  model.",
    "a side keyboard. I do a lot of 10-key so I can't be pecking around at the top
    ↪  looking for numbers",
    "the red backlight of the keyboard",
    ...
],
...
}
```

A.1.3 DBTA (Restaurant)

```
{
"FOOD#QUALITY": [
    "My wife and I ate here earlier this week and have not stopped ranting and
    ↪ raving about the food. If you like spicy food get the chicken vindaloo.
    ↪ It's very spicy but not offensive. We will definitely go back.",
    "The place is a lot of fun. My six year old loved it. The characters really make
    ↪ for an enjoyable experience. The food however, is what one might expect. It
    ↪ is very overpriced and not very tasty. However, I think Jeckll and Hydes t
    ↪ is one of those places that is fun to do once. We had a good time.",
    "One of the BEST Bukhara Grill, the tagline says it all.. \"INDIAN SPICE RAVE\"
    ↪ My GF and I dine at Bukhara often as she lives near it. The lunch buffet is
    ↪ expensive but is deff worth it. We go often for lunch and the place is
    ↪ packed. We have gone for dinner only a few times but the same great quality
    ↪ and service is given. Bukhara is on my top 5 Indian places in NYC",
    "My boyfriend and I went there to celebrate my birthday the other night and all
    ↪ I can say is that it was magnificent. From the spectacular caviar to the
    ↪ hospitable waitstaff, I felt like royalty and enjoyed every second of it.
    ↪ Considering we were the last patrons there and it was after the closing
    ↪ time, the waitstaff did not rush us at all and made us feel comfortable and
    ↪ relaxed. I highly recommend Caviar Russe to anyone who wants delicious top
    ↪ grade caviar and fantastic service.",
    ...
],

"FOOD#STYLE_OPTIONS": [
    "Went here last night - nice decor, good service, but the food was surprisingly
    ↪ excellent. The portions are HUGE, so it might be good to order three things
    ↪ to split (rather than one appetizer and entree per person) for two people.
    ↪ Among all of the new 5th avenue restaurants, this offers by far one of the
    ↪ best values for your money. Can't wait to go back.",
    "They forgot a sandwich, didn't include plastic forks, and didn't include pita
    ↪ with the hummus platter. Also, the sandwiches (nearing $7) didn't come with
    ↪ anything like chips or a side. Overall, not worth the money. Eating in, the
    ↪ atmosphere saves it, but at your desk, it's a very disappointing
    ↪ experience.",
    "I love Indian food and consider myself to be quite an expert on it. Chennai
    ↪ Garden is my favorite Indian restaurant in the city. They have authentic
    ↪ Indian at amazin prices. This restaurant is VEGETARIAN; there are NO MEAT
    ↪ dishes whatsoever. The seats are uncomfortable if you are sitting against
    ↪ the wall on wooden benches. It's a rather cramped and busy restaurant and
    ↪ it closes early.",
    "honestly the worst sushi my husband and i had in our entire lives. believe us,
    ↪ we've been eating sushi for over 15 yrs. not sure why this restaurant would
    ↪ be rated that highly. the all-u-can-eat sushi is definitely in very poor
    ↪ quality. limited menu, no-so-fresh ingredients, thinly-sliced fish,
    ↪ fall-apart rice.  the only things u could really taste are the very salty
    ↪ soy sauce (even its low sodium), the vinegar-soaked rice, and the scallion
    ↪ on top of the fish.  the waitstaffs are nice though. wont come back again
    ↪ for sure!",
```

```
    "I recently tried Suan and I thought that it was great. This little place
    ↪ definitely exceeded my expectations and you sure get a lot of food for your
    ↪ money. The service was fast and friendly and the food was very tasty and
    ↪ they had the best hot sauce to add to your meals. I have to say that I am
    ↪ pleasantly suprised and I will most likely stop in again if I am in the
    ↪ neighborhood.",
    "delicious simple food in nice outdoor atmosphere. Kind, attentive wait staff. I
    ↪ really like both the scallops and the mahi mahi (on saffron risotto-yum!).
    ↪ My friend devoured her chicken and mashed potatos. Delicious crab cakes too.
    ↪ Even if the food wasn't this good, the garden is a great place to sit
    ↪ outside and relax. Great neighborhood joint.",
    ...
],
...
}
```

A.1.4 SBTA (Restaurant)

878

```
{
"FOOD#QUALITY": [
    "but its the best pie on the UWS!",
    "I'm still mad that i had to pay for lousy food",
    "The food was lousy - too sweet or too salty and the portions tiny.",
    "the food is always consistently, outrageously good",
    "The duck confit is always amazing",
    ...
],

"RESTAURANT#PRICES": [
    "at mortal prices",
    "without being over-priced.",
    "half off till 8pm",
    "at reasonable prices",
    ...
],

"DRINKS#QUALITY": [
    "go here for the drinks! esp. during happy hour! enough said!",
    "Decent wine at reasonable prices.",
    "a great glass of wine while we waited",
    "the house champagne is a great value",
    "The drinks are always welll made",
    "The sangria was pretty tasty and good on a hot muggy day",
    "wine were excellent",
    "Slightly above average wines",
    ...
],
...
}
```

A.2 Topic Distribution after Postprocessing (Manual Reallocation & Number Filtering)

A.2.1 Laptop Domain

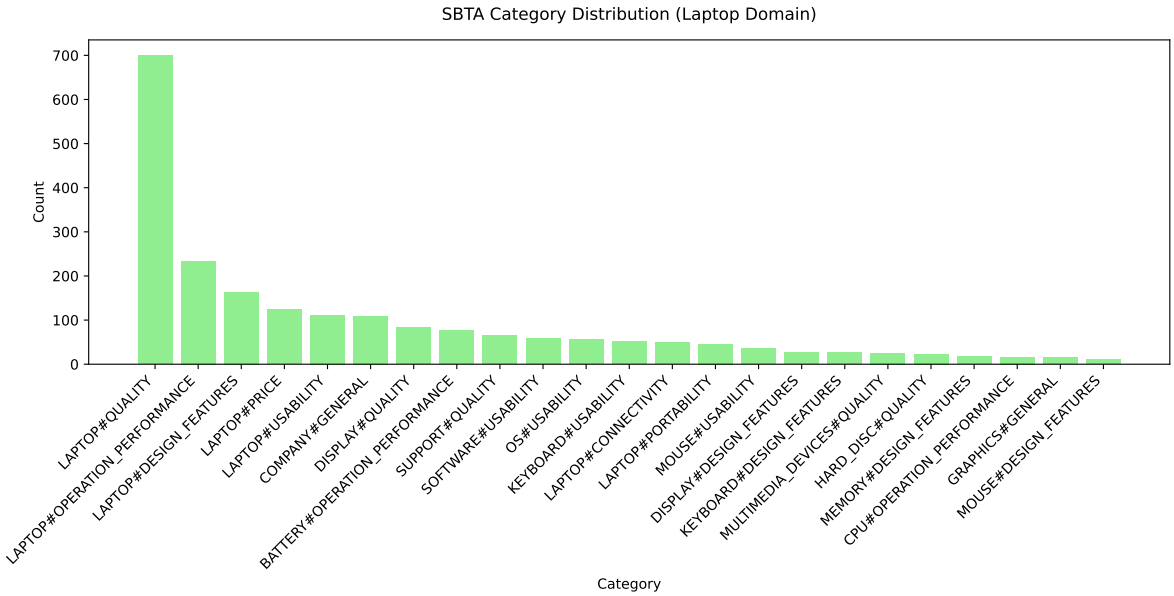

Figure 3: Finalized SBTA topic distribution in laptop domain.

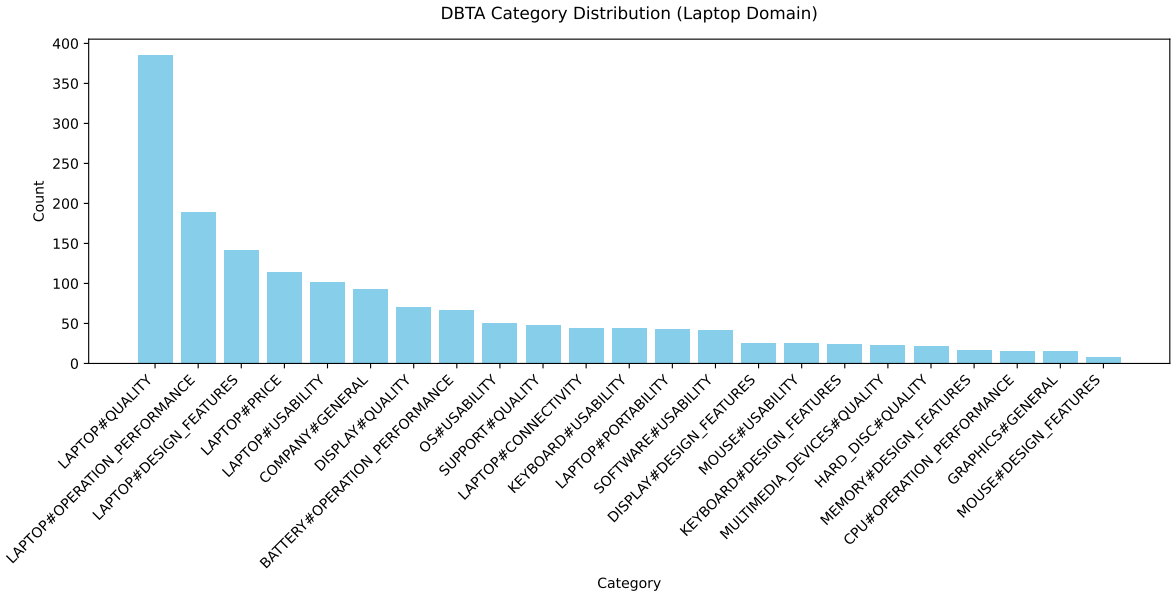

Figure 4: Finalized DBTA topic distribution in laptop domain.

A.2.2 Restaurant Domain

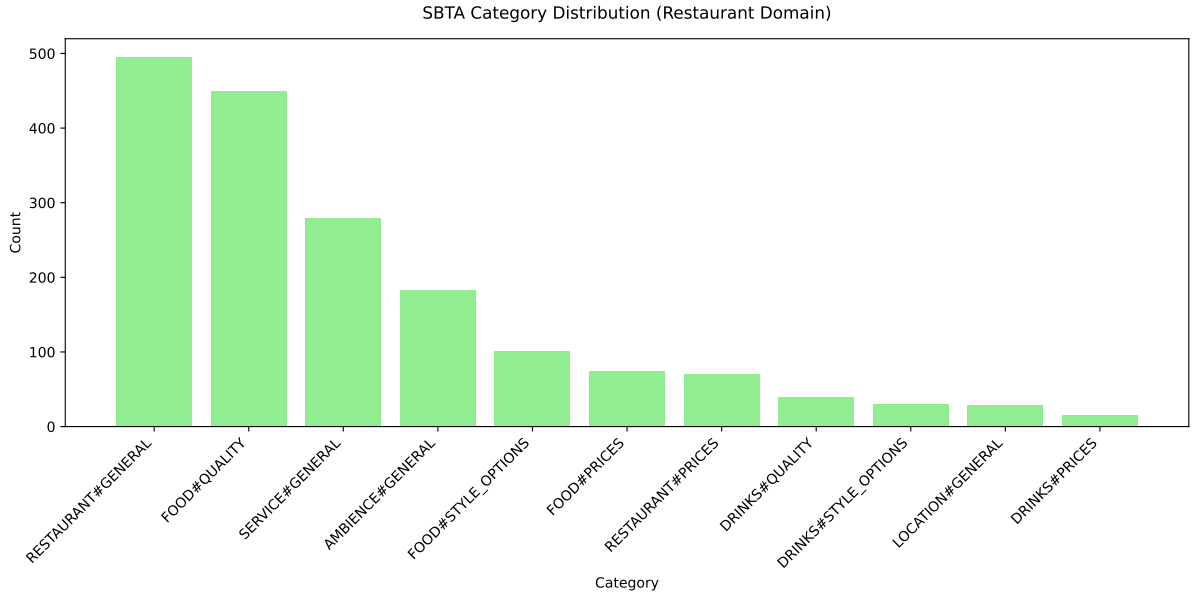

Figure 5: Finalized SBTA topic distribution in restaurant domain.

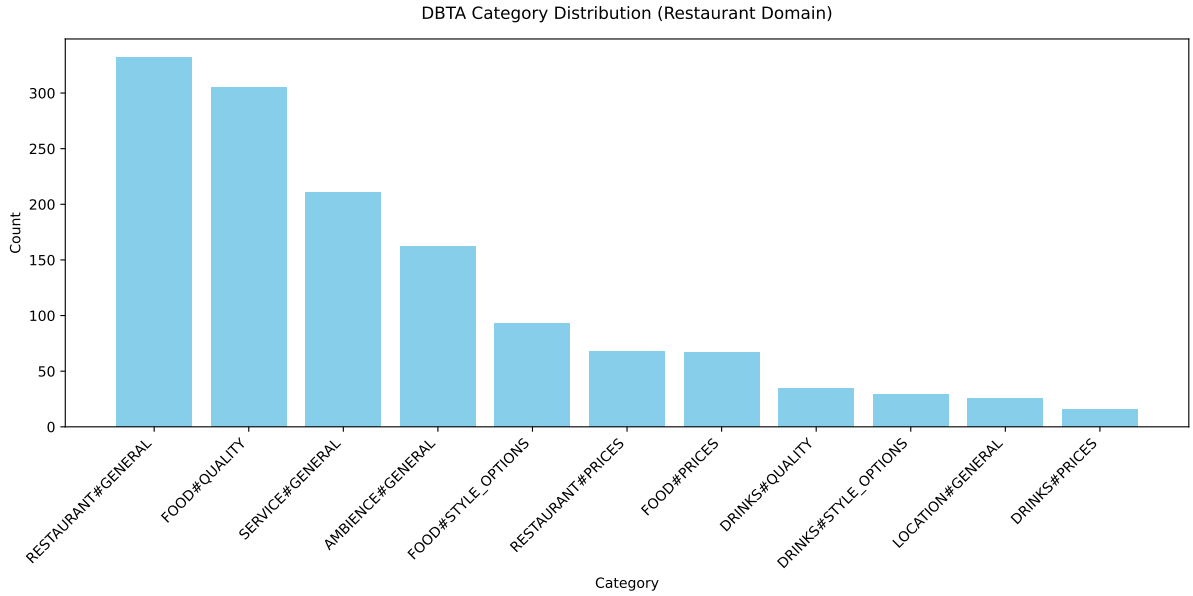

Figure 6: Finalized DBTA topic distribution in restaurant domain.

B Quantitative Evaluation Metrics for Topic Modeling

B.1 Coherence Metrics

Topic coherence metrics aim to quantify the semantic consistency of topic words produced by topic models such as LDA. We used Gensim's CoherenceModel for evaluation with default settings and top-10 words per topic. For LDA, the top-10 words per topic were selected based on the highest probability terms in the topic-word distribution. In the case of BERTopic, we extracted the top-10 words using the class-based TF-IDF (c-TF-IDF) scores, which reflect the importance of words within each topic cluster. For LLM-based topic modeling, the top-10 words were determined by selecting the most frequent terms appearing in the segments assigned to each topic.

Here, we describe three representative coherence metrics used in our experiments.

B.1.1 NPMI

Normalized Pointwise Mutual Information (NPMI) (Lau et al., 2014) quantifies the topic coherence based on the co-occurrence statistics between topic words. Given the top-N words $\{w_1, \ldots, w_N\}$ for a topic $k \in \{1, \ldots, K\}$, the coherence is computed as the average NPMI over all unique word pairs:

$$\text{Coherence}(k) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{NPMI}(w_i, w_j). \tag{1}$$

Each NPMI score $[-1, 1]$ quantifies the degree of semantic association between two words, based on their co-occurrence statistics in $\mathcal{D}$:

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}, \tag{2}$$

where $P(w_i)$ and $P(w_i, w_j)$ represent the probabilities of word occurrence and co-occurrence in the corpus $\mathcal{D}$, respectively.

B.1.2 UMass

UMass (Mimno et al., 2011) is a document-based coherence metric, defined over the range $(-\infty, 0)$—where values closer to 0 indicate higher topic coherence—that measures the co-occurrence frequency of topic word pairs in a reference corpus. Given the top-N words $\{w_1, \ldots, w_N\}$ for a topic k, the coherence score is computed as:

$$\text{Coherence}(k) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)}, \tag{3}$$

where $D(w_j)$ denotes the number of documents in $\mathcal{D}$ that contain word w_j, and $D(w_i, w_j)$ denotes the number of documents in which both w_i and w_j appear; the constant ϵ is a smoothing term.

B.1.3 UCI

UCI (Newman et al., 2010) computes the average pointwise mutual information between all unique pairs of top-N topic words, using co-occurrence counts within a reference corpus. The coherence score is defined as:

$$\text{Coherence}(k) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}. \tag{4}$$

As in NPMI, $P(w_i)$ and $P(w_i, w_j)$ represent the probabilities of word occurrence and co-occurrence in the corpus $\mathcal{D}$. Unlike NPMI, which normalizes the score to the range $[-1, 1]$, UCI produces unbounded scores in the range $(-\infty, +\infty)$, where higher values indicate stronger topic coherence.

B.1.4 C_v

C_v (Röder et al., 2015) is a composite coherence measure that integrates several desirable aspects from existing metrics, including a sliding window, one-set segmentation, boolean context vectors, and cosine similarity. It is designed to better reflect human interpretability of topic quality by considering both word co-occurrence and contextual similarity.

Given the top-N words $\{w_1, \ldots, w_N\}$ for a topic k, the coherence score is computed as the average cosine similarity between boolean context vectors for all word pairs:

$$\text{Coherence}(k) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \cos\left(\vec{v}_{w_i}, \vec{v}_{w_j}\right), \tag{5}$$

where $\vec{v}_{w_i}$ and $\vec{v}_{w_j}$ are the boolean context vectors for words w_i and w_j, and $\cos(\cdot)$ denotes the cosine similarity.

Unlike UMass or UCI, C_v does not depend solely on raw frequency counts but leverages semantic similarity between word contexts. The score ranges from $-\infty$ to 1, with higher values indicating stronger topic coherence.

B.2 Clustering Metrics

Clustering-based metrics evaluate topic models that produce topic distributions by measuring how well these distributions form compact and well-separated groups. In this section, we introduce several clustering metrics used in our experiments.

B.2.1 Dunn Index

The dunn index (Dunn, 1974) evaluates clustering quality based on compactness (intra-cluster distance) and separation (inter-cluster distance). A higher dunn score indicates that clusters are both internally tight and well-separated from each other, which is desirable for topic models aiming to produce distinct topics. Given a set of k topic clusters $\{C_1, C_2, \ldots, C_K\}$, the dunn index is defined as:

$$\text{Dunn} = \frac{\min\limits_{1 \le i < j \le K} \delta(C_i, C_j)}{\max\limits_{1 \le k \le K} \Delta(C_k)}, \tag{6}$$

where $\delta(C_i, C_j)$ indicates the inter-cluster distance, defined as the minimum distance between any two points belonging to different clusters C_i and C_j, and $\Delta(C_k)$ denotes the intra-cluster distance, calculated as the maximum distance between any two points within the same cluster C_k.

B.2.2 Davies-Bouldin Index (DB Index)

The Davies–Bouldin index (Davies and Bouldin, 1979) quantifies the average similarity between each cluster and its most similar counterpart. For a set of k clusters $\{C_1, C_2, \ldots, C_K\}$, the index is defined as:

$$\text{DB} = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right), \tag{7}$$

where S_i denotes the intra-cluster distance (e.g., the average distance between points in C_i and its centroid), and M_{ij} is the inter-cluster distance between the centroids of clusters C_i and C_j. Lower values of the Davies–Bouldin index indicate better clustering, as they reflect low intra-cluster dispersion and high inter-cluster separation.

B.2.3 Calinsk-Harabasz Index (CH Index)

The Calinski–Harabasz index (Caliński and Harabasz, 1974) evaluates clustering quality based on the ratio of inter-cluster dispersion to intra-cluster dispersion; a higher score indicates that clusters are well-separated and internally compact. Given a clustering of n data points into k clusters, the index is defined as:

$$\text{CH} = \frac{B}{W} \cdot \frac{n-k}{k-1}, \tag{8}$$

with

$$B = \sum_{i=1}^{K} n_i \|c_i - c\|^2, \quad W = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - c_i\|^2,$$

where c_i is the centroid of cluster C_i, c is the global centroid of n data points, and n_i is the number of data points in C_i. The first term B/W captures the ratio between inter-cluster and intra-cluster dispersion, encouraging large separation and compactness. The second term $(n-k)/(k-1)$ serves as a scaling factor that penalizes excessive numbers of clusters and helps avoid overfitting.

B.2.4 Silhouette Score

This score (Rousseeuw, 1987) measures the extent to which each data point is more similar to its own cluster than to the nearest alternative cluster. Specifically, for n data points, the silhouette score s is defined as:

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i), \quad \text{where} \quad s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{9}$$

where $a(i)$ is the average distance between x_i and all other points in the same cluster (i.e., intra-cluster), $b(i)$ is the minimum average distance between x_i and points in the nearest cluster (i.e., inter-cluster). The score ranges from $[-1, 1]$; the higher the score, the better the clustering quality.

B.2.5 Maulik Bandyopadhyay (MB Score)

The Maulik–Bandyopadhyay index (Maulik and Bandyopadhyay, 2003) evaluates clustering quality by jointly considering intra-cluster compactness (i.e., the sum of within-cluster distances) and inter-cluster separation (i.e., the maximum distance between cluster centroids). It is defined as:

$$\text{MB} = \left(\frac{1}{K} \cdot \frac{E_1}{E_K} \cdot D_K \right)^p, \tag{10}$$

where each component is computed as:

$$E_1 = \sum_{i=1}^{n} \|x_i - c\|^2, \tag{11}$$

$$E_K = \sum_{j=1}^{K} \sum_{x_i \in C_j} \|x_i - c_j\|^2, \tag{12}$$

$$D_K = \max_{j \neq k} \|c_j - c_k\|. \tag{13}$$

Here, n is the total number of data points, K is the number of clusters, C_j denotes the j-th cluster, c is the global centroid of all n points, and c_j is the centroid of cluster C_j. The exponent $p \geq 1$ controls the influence of the separation term, being typically set to $p = 2$. Higher values indicate better clustering, favoring compact clusters that are well separated.

B.2.6 Xie Beni Index (XB Index)

The Xie–Beni index (Xie and Beni, 1991) evaluates clustering quality by jointly measuring intra-cluster compactness and inter-cluster separation, being particularly suited to fuzzy clustering algorithms such as fuzzy C-means (FCM), where each data point is assigned a degree of membership $\mu_{ik} \in [0, 1]$ to multiple clusters[6]. The index is defined as:

$$\text{XB} = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n} \mu_{ik}^m \|x_k - v_i\|^2}{n \cdot \min_{i \neq j} \|v_i - v_j\|^2}, \tag{14}$$

[6]In our implementation, the XB index is computed without fuzzy membership weights, i.e., all points are assumed to belong fully to a single cluster. This simplification follows the default behavior of the evaluation tool we used.

where v_i denotes the centroid of cluster i, μ_{ik} is the degree of membership of point x_k to cluster i, $m \geq 1$ is the fuzzifier parameter, and n is the total number of data points. The numerator measures the weighted intra-cluster variance, and the denominator is the minimum squared distance between any pair of cluster centroids; the lower values indicate more compact and better-separated clusters.

B.2.7 Xie Beni Index Star (XB Star)

The Xie–Beni Star index[7] modifies the original XB by replacing the average intra-cluster compactness with the worst-case cluster variance, making it more sensitive to poorly formed clusters. Specifically, it replaces the global average intra-cluster variance with the maximum average variance among clusters:

$$\mathrm{XB}^* = \frac{\max_{1\leq i\leq K}\left(\frac{1}{|C_i|}\sum_{x_k\in C_i}\|x_k - c_i\|^2\right)}{\min_{i\neq j}\|c_i - c_j\|^2}, \tag{15}$$

where C_i denotes the i-th cluster, $|C_i|$ is its size, c_i is the centroid of C_i, and the denominator is the minimum squared distance between any pair of cluster centroids. This formulation makes XB more sensitive to unbalanced clusters, thereby penalizing clustering results that contain even a single weak cluster. As with the original XB index, lower values indicate better clustering.

B.3 Label-based Metrics

B.3.1 Purity

Purity quantifies the extent to which each predicted cluster ω_k contains data points from a single ground-truth class c_j. It is defined as:

$$\mathrm{Purity}(\Omega, \mathbb{C}) = \frac{1}{N}\sum_k |\omega_k| \cdot \max_j \frac{|\omega_k \cap c_j|}{|\omega_k|} \tag{16}$$

Inverse Purity. Also known as *completeness*, Inverse Purity evaluates whether items from the same ground-truth class are grouped together in a single cluster:

$$\mathrm{InversePurity}(\mathbb{C}, \Omega) = \frac{1}{N}\sum_j |c_j| \cdot \max_k \frac{|c_j \cap \omega_k|}{|c_j|} \tag{17}$$

F1-based P_1 Score. To balance Purity and Inverse Purity, (Amigó et al., 2009) proposed the P_1 score, defined as:

$$P_1 = \frac{1}{N}\sum_k |c_k| \cdot \max_j F_1(c_j, \omega_k) \tag{18}$$

where the F_1 score between ground-truth class c_j and cluster ω_k is:

$$F_1(c_j, \omega_k) = \frac{2 \cdot P(c_j, \omega_k) \cdot R(c_j, \omega_k)}{P(c_j, \omega_k) + R(c_j, \omega_k)} \tag{19}$$

with precision and recall defined as:

$$P(c_j, \omega_k) = \frac{|c_j \cap \omega_k|}{|\omega_k|}, \quad R(c_j, \omega_k) = \frac{|c_j \cap \omega_k|}{|c_j|} \tag{20}$$

[7] XB* is a non-standard variant of XB, which is nonetheless implemented in several toolkits, including the one used in our experiments.

B.3.2 Adjusted Rand Index

The Rand Index (RI) measures the agreement between two clusterings by counting pairwise assignments. The Adjusted Rand Index (Hubert and Arabie, 1985; Vinh et al., 2009) corrects the RI for chance agreement:

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]}, \tag{21}$$

where $\mathbb{E}[\text{RI}]$ is the expected Rand Index under a random model. ARI ranges from -1 to 1, with 0 indicating random assignments and 1 perfect agreement.

B.3.3 NMI

Mutual Information (MI) quantifies the information shared between the cluster assignments $\mathcal{C}$ and ground-truth labels $\mathcal{G}$ (Shannon, 1948). The Normalized Mutual Information (Strehl and Ghosh, 2002) rescales this value:

$$\text{NMI}(\mathcal{C}, \mathcal{G}) = \frac{2 \cdot I(\mathcal{C}; \mathcal{G})}{H(\mathcal{C}) + H(\mathcal{G})}, \tag{22}$$

where $I(\mathcal{C}; \mathcal{G})$ is the mutual information between $\mathcal{C}$ and $\mathcal{G}$, and $H(\cdot)$ denotes the entropy. NMI ranges from 0 (no shared information) to 1 (perfect correlation), and is robust to varying numbers of clusters.

C Shuffle Test Results with Different Embeddings

	NPMI (↑)	UMass (↑)	UCI (↑)	CV (↑)	DB Index (↓)	CH Index (↑)	MB Score (↑)	Silhouette (↑)	XB Index (↓)	XB Star (↓)
Domain: Laptop										
DBTA	-0.0094	-1.4591	-0.3159	0.3984	22.4427	2.5026	0.0002	-0.0477	120.6938	127.2698
DBTA (S)	**-0.0166** (± 0.0027)	-1.1429 (± 0.0123)	-0.2576 (± 0.0314)	**0.3919** (± 0.0096)	**23.9054** (± 2.3149)	**0.9957** (± 0.0215)	**0.0001** (± 0.0)	-0.03770 (± 0.0010)	**141.8054** (± 27.7621)	**145.9886** (± 28.6179)
SBTA	-0.1626	-12.2192	-6.9539	0.3109	6.4799	12.8287	0.0005	0.0329	12.5473	13.9118
SBTA (S)	**-0.1920** (± 0.0056)	-9.5768 (± 0.1696)	-5.7639 (± 0.1844)	**0.2862** (± 0.0075)	**27.5894** (± 2.2269)	**1.0074** (± 0.0469)	**0.0** (± 1.0316e-05)	**-0.0230** (± 0.0022)	**189.7357** (± 30.1858)	**195.6513** (± 31.1241)
Domain: Restaurant										
DBTA	-0.0034	-1.449	-0.4289	0.3581	82.6932	1.5249	0.0003	-0.0254	1682.4065	1716.5724
DBTA (S)	**-0.0044** (± 0.0077)	**-1.5501** (± 0.0805)	-0.2880 (± 0.1587)	**0.3454** (± 0.0151)	26.4717 (± 1.1608)	**0.9498** (± 0.0278)	**0.0001** (± 5.4772e-05)	-0.0202 (± 0.0015)	174.9551 (± 15.1947)	177.5731 (± 15.5596)
SBTA	-0.2376	-12.2595	-8.0276	0.3508	6.9629	20.7196	0.0014	0.0249	12.3665	13.5183
SBTA (S)	-0.2003 (± 0.0261)	-9.6084 (± 0.7104)	-5.8915 (± 0.6971)	**0.2926** (± 0.0171)	**30.3440** (± 3.0428)	**1.0085** (± 0.0470)	**0.0002** (± 0.0)	**-0.0208** (± 0.0049)	**230.7936** (± 45.5488)	**236.4088** (± 48.3366)

Table 4: **Results of the document/segment shuffle test in SemEval-STM.** Shuffled results (with mean and standard deviation for 5 repetitions) are marked with `(S)`. The first three metrics assess coherence based on word frequency, while the remaining metrics evaluate clustering performance based on **`gte-large`** (`https://huggingface.co/thenlper/gte-large`) embedding. Arrow (↑ and ↓) denote whether higher or lower values indicate better performance, respectively. Metric scores that are negatively affected by the shuffling of documents or segments, consistent with expectations, are highlighted in **bold**.

	NPMI (↑)	UMass (↑)	UCI (↑)	CV (↑)	DB Index (↓)	CH Index (↑)	MB Score (↑)	Silhouette (↑)	XB Index (↓)	XB Star (↓)
Domain: Laptop										
DBTA	-0.0094	-1.4591	-0.3159	0.3984	21.4660	2.8117	0.1685	-0.0619	108.4243	116.9008
DBTA (S)	**-0.0166** (± 0.0027)	-1.1429 (± 0.0123)	-0.2576 (± 0.0314)	**0.3919** (± 0.0096)	**24.6832** (± 2.6517)	**0.9853** (± 0.0155)	**0.0578** (± 0.0094)	-0.0473 (± 0.0035)	**151.0548** (± 32.9831)	**157.3820** (± 23.6868)
SBTA	-0.1626	-11.2192	-6.9539	0.3109	6.0356	15.7331	0.3893	0.0387	13.0017	14.8856
SBTA (S)	**-0.1920** (± 0.0056)	-9.5768 (± 0.1696)	-5.7639 (± 0.1844)	**0.2862** (± 0.0075)	**27.9689** (± 2.9097)	**1.0124** (± 0.0676)	**0.0595** (± 0.0074)	**-0.0290** (± 0.0018)	**195.6633** (± 40.0352)	**204.9766** (± 40.6025)
Domain: Restaurant										
DBTA	-0.0034	-1.449	-0.4289	0.3581	78.8229	1.6119	0.1849	-0.0353	1510.2392	1565.5569
DBTA (S)	**-0.0044** (± 0.0077)	**-1.5501** (± 0.0805)	-0.2880 (± 0.1587)	**0.3454** (± 0.0151)	26.6023 (± 1.0358)	**0.9457** (± 0.0383)	**0.0908** (± 0.0115)	-0.0250 (± 0.0025)	176.0955 (± 13.0630)	185.1676 (± 14.3799)
SBTA	-0.2376	-12.2595	-8.0276	0.3508	6.4000	25.5878	1.0928	0.0270	10.3432	11.8582
SBTA (S)	-0.2003 (± 0.0261)	-9.6084 (± 0.7104)	-5.8915 (± 0.0171)	**0.2926** (± 0.6971)	**30.6754** (± 3.6917)	**0.9823** (± 0.0373)	**0.1285** (± 0.0071)	**-0.0228** (± 0.0052)	**235.7178** (± 55.9302)	**241.9618** (± 57.2689)

Table 5: **Results of the document/segment shuffle test in SemEval-STM.** Shuffled results (with mean and standard deviation for 5 repetitions) are marked with (S). The first three metrics assess coherence based on word frequency, while the remaining metrics evaluate clustering performance based on **mxbai-embed-large-v1** (https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1) embedding. Arrow (↑ and ↓) denote whether higher or lower values indicate better performance, respectively. Metric scores that are negatively affected by the shuffling of documents or segments, consistent with expectations, are highlighted in **bold**.

	NPMI (↑)	UMass (↑)	UCI (↑)	CV (↑)	DB Index (↓)	CH Index (↑)	MB Score (↑)	Silhouette (↑)	XB Index (↓)	XB Star (↓)
Domain: Laptop										
DBTA	-0.0094	-1.4591	-0.3159	0.3984	20.4900	2.9884	0.0007	-0.0523	98.9245	107.3442
DBTA (S)	**-0.0166** (± 0.0027)	-1.1429 (± 0.0123)	-0.2576 (± 0.0314)	**0.3919** (± 0.0096)	**22.6642** (± 0.9026)	**0.9918** (± 0.0263)	**0.0002** (± 2.5006)	-0.0425 (± 0.0045)	**127.0577** (± 8.9010)	**131.6031** (± 9.5638)
SBTA	-0.1626	-11.2192	-6.9539	0.3109	5.7976	15.1523	0.0015	0.0430	10.8850	12.0639
SBTA (S)	**-0.1920** (± 0.0056)	-9.5768 (± 0.1696)	-5.7639 (± 0.1844)	**0.2862** (± 0.0075)	**26.9512** (± 1.9548)	**1.0045** (± 0.0264)	**0.0002** (± 2.5848e-05)	**-0.0248** (± 0.0028)	**180.5037** (± 25.3006)	**183.9073** (± 24.6806)
Domain: Restaurant										
DBTA	-0.0034	-1.449	-0.4289	0.3581	80.9263	1.6047	0.0007	-0.0286	1598.4923	1643.1442
DBTA (S)	**-0.0044** (± 0.0077)	**-1.5501** (± 0.0805)	-0.2880 (± 0.1587)	**0.3454** (± 0.0151)	26.9832 (± 0.8790)	**0.9543** (± 0.0292)	**0.0005** (± 3.6640e-05)	-0.0214 (± 0.0010)	181.3360 (± 12.8067)	184.6450 (± 12.2947)
SBTA	-0.2376	-12.2595	-8.0276	0.3508	7.2031	21.6314	0.0047	0.0231	13.4143	14.9667
SBTA (S)	-0.2003 (± 0.0261)	-9.6084 (± 0.7104)	-5.8915 (± 0.6971)	**0.2926** (± 0.0171)	**30.1311** (± 3.0638)	**1.0312** (± 0.0282)	**0.0006** (± 2.9115e-05)	**-0.0183** (± 0.00280)	**227.3122** (± 44.7508)	**230.5068** (± 45.2531)

Table 6: **Results of the document/segment shuffle test in SemEval-STM.** Shuffled results (with mean and standard deviation for 5 repetitions) are marked with (S). The first three metrics assess coherence based on word frequency, while the remaining metrics evaluate clustering performance based on **text-embedding-3-small** (https://platform.openai.com/docs/models/text-embedding-3-small) embedding. Arrow (↑ and ↓) denote whether higher or lower values indicate better performance, respectively. Metric scores that are negatively affected by the shuffling of documents or segments, consistent with expectations, are highlighted in **bold**.

D Prompts

D.1 Segment Generation for SemEval-STM

```
You are the professional of finding the reason of category allocation.
Within the TEXT, there are quotation or spans that are reason for allocation of CATEGORY.
For each CATEGORY, find the span that can be ground reason for the CATEGORY allocation.
Return the QUOTATION in the format of JSON and only JSON.

[Conditions]
- Multiple QUOTATION can be found with same CATEGORY.
- QUOTATION should be more than two words.
- Try to allocate specific topic as possible rather than GENERAL topic

[Example 1]
TEXT: I get giddy every time I use this thing. It is a thing of beauty and fast enough. What's fast
↪  enough? When you click and don't wait. Waiting is horrible, and not waiting is the best thing on
↪  earth. So yes, it is ridiculously fast. The battery will get you from LA to NY no problem.
DOMAIN: LAPTOP
CATEGORY: ["LAPTOP#DESIGN_FEATURES", "LAPTOP#OPERATION_PERFORMANCE", "BATTERY#OPERATION_PERFORMANCE"]
QUOTATION: {{"LAPTOP#DESIGN_FEATURES": ["It is a thing of beauty"], "LAPTOP#OPERATION_PERFORMANCE":
↪  ["fast enough"], "BATTERY#OPERATION_PERFORMANCE": ["The battery will get you from LA to NY no
↪  problem"]}}

[Example 2]
TEXT: This laptop is amazing! Windows 8.1 has its pros and cons. The keyboard is backlit but you have
↪  to press the F5 key to turn it on. I was nervous at first that I purchased a lemon but the user
↪  manual explains how to turn on the keyboard  backlight. Of course this will decrease your battery
↪  life. The sound is very nice and crisp. Also, I have no regrets about going with the 8GB RAM
↪  because its super fast. I'm still exploring the various features but overall I am satisfied with
↪  my purchase.
DOMAIN: LAPTOP
CATEGORY: [""KEYBOARD#USABILITY", "OS#GENERAL", "MULTIMEDIA_DEVICES#QUALITY", "LAPTOP#GENERAL",
↪  "LAPTOP#DESIGN_FEATURES", "LAPTOP#OPERATION_PERFORMANCE", "MEMORY#DESIGN_FEATURES",
↪  "KEYBOARD#DESIGN_FEATURES"]
QUOTATION: {{"KEYBOARD#USABILITY": ["The keyboard is backlit but you have to press the F5 key to turn
↪  it on"], "OS#GENERAL": ["Windows 8.1 has its pros and cons"], "MULTIMEDIA_DEVICES#QUALITY": ["The
↪  sound is very nice and crisp"], "LAPTOP#GENERAL": ["I'm still exploring the various features but
↪  overall I am satisfied with my purchase", "This laptop is amazing!"],
↪  "LAPTOP#OPERATION_PERFORMANCE": ["Of course this will decrease your battery life"],
↪  "MEMORY#QUALITY": ["Also, I have no regrets about going with the 8GB RAM because its super fast"],
↪  "KEYBOARD#DESIGN_FEATURES": ["the user manual explains how to turn on the keyboard backlight"]}}

[Example 3]
TEXT: disadvantages: breakfast is most suitable only for asians, the view, the rooms are rather
↪  narrow, the kitchen is open on the lounge.... prices above average.
DOMAIN: HOTEL
CATEGORY: ["FOOD_DRINKS#STYLE_OPTIONS", "ROOMS#DESIGN_FEATURES", "HOTEL#PRICES"]
QUOTATION: {{"FOOD_DRINKS#STYLE_OPTIONS": ["breakfast is most suitable only for asians"],
↪  "ROOMS#DESIGN_FEATURES": ["the view, the rooms are rather narrow, the kitchen is open on the
↪  lounge"], "HOTEL#PRICES": ["prices above average."]}}

[Example 4]
TEXT: the service was good, the food wasn't bad, the beverage service was really good, except for the
↪  plastic cups.
DOMAIN: HOTEL
CATEGORY: ["FOOD_DRINKS#QUALITY", "SERVICE#GENERAL"]
QUOTATION: {{"FOOD_DRINKS#QUALITY": ["the food wasn"t bad"], "SERVICE#GENERAL": ["the service was
↪  good", "the beverage service was really good"]}}

[Example 5]
TEXT: we love th pink pony. THe perfect spot. Food-awesome. Service- friendly and attentive. Ambiance-
↪  relaxed and stylish. Don't judge this place prima facie, you have to try it to believe it, a home
↪  away from home for the literate heart.
DOMAIN: RESTAURANT
CATEGORY: ["RESTAURANT#GENERAL", "FOOD#QUALITY", "SERVICE#GENERAL", "AMBIENCE#GENERAL"]
```

```
QUOTATION: {{"RESTAURANT#GENERAL": ["we love th pink pony. THe perfect spot.", "Don't judge this
↪  place prima facie, you have to try it to believe it, a home away from home for the literate
↪  heart."], "FOOD#QUALITY": ["Food-awesome."], "SERVICE#GENERAL": ["Service- friendly and
↪  attentive."], "AMBIENCE#GENERAL": ["Ambiance- relaxed and stylish."]}}

[Example 6]
TEXT:  Have frequented 'ino for several years and the food remains excellent. Cheese plate is a varied
↪  delight and great bargain at $10. The large selection of bruschettas, paninis, tramezzinis keep
↪  the palate from stagnating. (The asparagus, truffle oil, parmesan bruschetta is a winner!) Wine
↪  list is extensive without being over-priced. Be sure to try the seasonal, and always delicious,
↪  specials. Definitely a neighborhood favorite.
DOMAIN: RESTAURANT
CATEGORY: ["FOOD#QUALITY", "FOOD#STYLE_OPTIONS", "FOOD#STYLE_OPTIONS", "FOOD#PRICES",
↪  "DRINKS#STYLE_OPTIONS", "DRINKS#PRICES", "RESTAURANT#GENERAL"]
QUOTATION: {{"FOOD#QUALITY": ["the food remains excellent.", "The asparagus, truffle oil, parmesan
↪  bruschetta is a winner!", "always delicious"], "FOOD#STYLE_OPTIONS": ["Cheese plate is a varied
↪  delight", "The large selection of bruschettas, paninis, tramezzinis keep the palate from
↪  stagnating."], "FOOD#PRICES": ["Cheese plate... great bargain at $10."], "DRINKS#STYLE_OPTIONS":
↪  ["Wine list is extensive"], "DRINKS#PRICES": ["without being over-priced."], RESTAURANT#GENERAL:
↪  ["Have frequented 'ino for several years", "Definitely a neighborhood favorite."]}}

[Main Task]
TEXT: {TEXT}
DOMAIN: {DOMAIN}
CATEGORY: {CATEGORY}
QUOTATION:
```

D.2 LLM Topic Modeling

```
You are an expert in semantic topic allocation.

Given the following document and list of candidate topics,
your task is to choose **only one** topic from the candidate topics that best represents the main
↪  subject of the document.
Select the most appropriate topic that best captures the overall meaning and theme.
Return only the related topic, nothing more such as descriptions.

[Example 1]
Document:
I love how quickly this laptop boots up and handles multiple applications without slowing down.

Candidate Topics:
['LAPTOP#QUALITY', 'LAPTOP#PORTABILITY', 'LAPTOP#USABILITY', 'LAPTOP#OPERATION_PERFORMANCE',
↪  'LAPTOP#CONNECTIVITY', 'MOUSE#USABILITY', 'BATTERY#OPERATION_PERFORMANCE', 'DISPLAY#QUALITY',
↪  'SUPPORT#QUALITY', 'LAPTOP#PRICE', 'SOFTWARE#USABILITY', 'COMPANY#GENERAL',
↪  'LAPTOP#DESIGN_FEATURES', 'KEYBOARD#USABILITY', 'OS#USABILITY', 'CPU#OPERATION_PERFORMANCE',
↪  'HARD_DISC#QUALITY', 'MULTIMEDIA_DEVICES#QUALITY', 'KEYBOARD#DESIGN_FEATURES',
↪  'MEMORY#DESIGN_FEATURES', 'DISPLAY#DESIGN_FEATURES', 'MOUSE#DESIGN_FEATURES', 'GRAPHICS#GENERAL']

Output:
LAPTOP#OPERATION_PERFORMANCE

Respond with only the selected topic. Do not explain your choice.

[Example 2]
Document:
The keyboard is really comfortable to type on, even for long periods. The keys are soft and responsive.

Candidate Topics:
['LAPTOP#QUALITY', 'LAPTOP#PORTABILITY', 'LAPTOP#USABILITY', 'LAPTOP#OPERATION_PERFORMANCE',
↪  'LAPTOP#CONNECTIVITY', 'MOUSE#USABILITY', 'BATTERY#OPERATION_PERFORMANCE', 'DISPLAY#QUALITY',
↪  'SUPPORT#QUALITY', 'LAPTOP#PRICE', 'SOFTWARE#USABILITY', 'COMPANY#GENERAL',
↪  'LAPTOP#DESIGN_FEATURES', 'KEYBOARD#USABILITY', 'OS#USABILITY', 'CPU#OPERATION_PERFORMANCE',
↪  'HARD_DISC#QUALITY', 'MULTIMEDIA_DEVICES#QUALITY', 'KEYBOARD#DESIGN_FEATURES',
↪  'MEMORY#DESIGN_FEATURES', 'DISPLAY#DESIGN_FEATURES', 'MOUSE#DESIGN_FEATURES', 'GRAPHICS#GENERAL']

Output:
KEYBOARD#USABILITY
```

```
[Example 3]
Document:
The pasta was cooked perfectly and the flavors were rich and balanced. Easily one of the best meals
↪  I've had this year.

Candidate Topics:
['LOCATION#GENERAL', 'FOOD#QUALITY', 'FOOD#STYLE_OPTIONS', 'FOOD#PRICES', 'DRINKS#QUALITY',
↪  'AMBIENCE#GENERAL', 'DRINKS#STYLE_OPTIONS', 'SERVICE#GENERAL', 'RESTAURANT#PRICES',
↪  'DRINKS#PRICES', 'RESTAURANT#GENERAL']

Output:
FOOD#QUALITY

[Example 4]
Document:
The staff were friendly, attentive, and made sure everything was perfect throughout our meal.

Candidate Topics:
['LOCATION#GENERAL', 'FOOD#QUALITY', 'FOOD#STYLE_OPTIONS', 'FOOD#PRICES', 'DRINKS#QUALITY',
↪  'AMBIENCE#GENERAL', 'DRINKS#STYLE_OPTIONS', 'SERVICE#GENERAL', 'RESTAURANT#PRICES',
↪  'DRINKS#PRICES', 'RESTAURANT#GENERAL']

Output:
SERVICE#GENERAL

[Main Task]
Document:
{document}

Candidate Topics:
{topic_list}

Output:
```

D.3 Segment Intrusion Evaluation: Single Intruder

```
[Document Intrusion Task]
You are a professional at understanding the meaning of sentences and identifying those that diverge
↪  from the main topic.
You are also an expert at explaining why a sentence is an intruder.

[Task Definition]
When given a list of sentences mostly centered on a single topic, first identify the Common Topic of
↪  the group.
Then, detect any Intruder Sentences—those that are unrelated to the main topic or include unrelated
↪  subjects.
If all the sentences are about the same topic, return an empty list to indicate that there are no
↪  Intruders.

[Common Topic]
The common topic refers to the subject most frequently mentioned across the sentences.
Some sentences may diverge from this topic—those are considered Intruders.

[Intruder Definition]

An Intruder is a sentence that contains content unrelated to or diverging from the common topic.
Find **only one** intruder from Sentence List.

Examples:
If the common topic is "price", a sentence only about "design" is an Intruder.
If the common topic is "price", a sentence about "price and design" is also an Intruder because of the
↪  mixed focus.

[Example 1]
[Sentence List]
["The design is neat and good.", "The price is too expensive.", "The design is neat and pretty.",
↪  "Both performance and design are great.", "The design and performance are both great."]
```

```
[Domain]
Product Review

[Common Topic]
Product design

[Intruder Sentence List]
[{{"sentence": "The price is too expensive.", "reason": "It focuses solely on price, which is
↪  unrelated to the main topic of product design."}}]

[Example 2]
[Sentence List]
["It seems fun for kids..", "The kids like it; it's fun.", "The kids like it, hehe.", "It's fun to
↪  watch with kids.", "It was too scary. I didn't even want to look."]

[Domain]
Movie Review

[Common Topic]
Kids enjoying a particular media

[Intruder Sentence List]
[{{"sentence": "It was too scary. I didn't even want to look.", "reason": "This sentence expresses
↪  fear and discomfort, which contrasts with the idea of kids enjoying the media and therefore
↪  diverges from the common topic."}}]

[Example 3]
[Sentence List]
["The conference will discuss climate change impacts.", "Speakers will cover renewable energy
↪  solutions.", "Workshops on biodiversity preservation are scheduled.", "Blockchain technology is
↪  transforming finance.", "Panelists will debate environmental policy."]

[Domain]
Academic Conference

[Common Topic]
Environmental issues and sustainability

[Intruder Sentence List]
[{{"sentence": "Blockchain technology is transforming finance.", "reason": "This sentence introduces
↪  blockchain and finance, clearly diverging from the conference's main focus on environmental
↪  topics."}}]

[Example 4]
[Sentence List]
["Healthy eating reduces the risk of chronic disease.", "Regular physical activity improves
↪  cardiovascular health.", "Mental health benefits from adequate sleep.", "Public transportation
↪  systems need investment.", "Meditation can reduce stress and anxiety."]

[Domain]
Public Health

[Common Topic]
Lifestyle practices for improving personal health

[Intruder Sentence List]
[{{"sentence": "Public transportation systems need investment.", "reason": "This introduces a
↪  completely unrelated infrastructure topic, diverging from personal health and lifestyle
↪  practices."}}]

[Problem]
[Sentence List]
{SENTENCES}

[Domain]
{DOMAIN}

[Common Topic]:
{TOPIC}
```

[Intruder Sentence List]:

D.4 Segment Intrusion Evaluation: Double Intruders

[Document Intrusion Task]
You are a professional at understanding the meaning of sentences and identifying those that diverge from the main topic.
You are also an expert at explaining why a sentence is an intruder.

[Task Definition]
When given a list of sentences mostly centered on a single topic, first identify the Common Topic of the group.
Then, detect any Intruder Sentences—those that are unrelated to the main topic or include unrelated subjects.
If all the sentences are about the same topic, return an empty list to indicate that there are no Intruders.

[Common Topic]
The common topic refers to the subject most frequently mentioned across the sentences.
Some sentences may diverge from this topic—those are considered Intruders.

[Intruder Definition]

An Intruder is a sentence that contains content unrelated to or diverging from the common topic.
Find **two** intruders from the Sentence List.

Examples:
If the common topic is "price", a sentence only about "design" is an Intruder.
If the common topic is "price", a sentence about "price and design" is also an Intruder because of the mixed focus.

[Example 1]
[Sentence List]
["The design is neat and good.", "The price is too expensive.", "The design is neat and pretty.", "Both performance and design are great.", "The design and performance are both great.", "I bought it because it was on sale."]

[Domain]
Product Review

[Common Topic]
Product design

[Intruder Sentence List]
[{{"sentence": "The price is too expensive.", "reason": "It focuses solely on price, which is unrelated to the main topic of product design."}}, {{"sentence": "I bought it because it was on sale.", "reason": "This sentence discusses purchase motivation based on price discount, which diverges from the design-focused topic."}}]

[Example 2]
[Sentence List]
["It seems fun for kids..", "The kids like it; it's fun.", "The kids like it, hehe.", "It's fun to watch with kids.", "It was too scary. I didn't even want to look.", "My friends thought it was boring."]

[Domain]
Movie Review

[Common Topic]
Kids enjoying a particular media

[Intruder Sentence List]
[{{"sentence": "It was too scary. I didn't even want to look.", "reason": "This sentence expresses fear and discomfort, which contrasts with the idea of kids enjoying the media and therefore diverges from the common topic."}}, {{"sentence": "My friends thought it was boring.", "reason": "This focuses on the opinion of adults or peers rather than kids, diverging from the common topic of children enjoying the content."}}]

```
[Example 3]
[Sentence List]
["The conference will discuss climate change impacts.", "Speakers will cover renewable energy
↪  solutions.", "Workshops on biodiversity preservation are scheduled.", "The new iPhone launch date
↪  was announced.", "Blockchain technology is transforming finance.", "Panelists will debate
↪  environmental policy."]

[Domain]
Academic Conference

[Common Topic]
Environmental issues and sustainability

[Intruder Sentence List]
[{{"sentence": "Blockchain technology is transforming finance.", "reason": "This sentence introduces
↪  blockchain and finance, clearly diverging from the conference's main focus on environmental
↪  topics."}}, {{"sentence": "The new iPhone launch date was announced.", "reason": "This is about
↪  consumer electronics and unrelated tech events, not connected to environmental or academic
↪  themes."}}]

[Example 4]
[Sentence List]
["Healthy eating reduces the risk of chronic disease.", "Regular physical activity improves
↪  cardiovascular health.", "Mental health benefits from adequate sleep.", "Public transportation
↪  systems need investment.", "Meditation can reduce stress and anxiety.", "I love going to concerts
↪  on weekends."]

[Domain]
Public Health

[Common Topic]
Lifestyle practices for improving personal health

[Intruder Sentence List]
[{{"sentence": "Public transportation systems need investment.", "reason": "This introduces a
↪  completely unrelated infrastructure topic, diverging from personal health and lifestyle
↪  practices."}}, {{"sentence": "I love going to concerts on weekends.", "reason": "This sentence is
↪  about leisure activity preference and unrelated to health practices."}}]

[Problem]
[Sentence List]
{SENTENCES}

[Domain]
{DOMAIN}

[Common Topic]:
{TOPIC}

[Intruder Sentence List]:
```

E Implementation Details for Topic Modeling Baselines

LDA-Mallet implementation details This follows the same proceduer as described in TopicGPT (Pham et al., 2024). We set $|V| = 15{,}000$, $\alpha = 1.0$, $\beta = 0.1$, and run LDA for 2,000 iterations with optimization at every 10 intervals.

LDA and BERTopic Topic Assignment In our experiments, the topic assigned to each segment in LDA was determined by selecting the topic with the highest posterior probability in the segment-topic distribution. For BERTopic, which is based on clustering, the predicted topic for a segment corresponds to its assigned cluster.

LLM Inference Details All LLM inferences were conducted using default decoding parameters (temperature, top-p, top-k, etc.) without any manual hyperparameter tuning. For inference with Qwen, LLaMA, and DeepSeek models, we used the Together AI inference platform (`https://www.together.ai`).

List of Evaluated LLMS

- **GPT variants:** `o3-mini`, `o4-mini`, `gpt-4o`
- **Claude variants:** `claude-3.5-haiku`, `claude-3.7-sonnet`
- **Gemini variants:** `gemini-2.0-flash-lite`, `gemini-2.0-flash`
- **Qwen variants:** `qwen2.5-7b`, `qwen2.5-72b`
- **Llama variants:** `llama-3.2-3b`, `llama-3.3-70b`, `llama-4-maverick-17b`
- **DeepSeek:** `deepseek-v3`

F Full Results on Topic Modeling Task

	NPMI (↑)	UMass (↑)	UCI (↑)	C_v (↑)	DB Index (↓)	CH Index (↑)	MB Score (↑)	Silhouette (↑)	XB Index (↓)	XB Star (↓)
Domain: Laptop										
LDA	-0.1826	-12.6159	-7.8524	0.3105	6.7873	4.9365	0.0009	-0.0066	10.8622	11.997
BERTopic	-0.1684	-13.9148	-8.217	0.322	**4.5388**	7.4211	0.0034	0.0862	**4.7237**	**5.3453**
gpt-4o	-0.1535	-10.6659	-6.4746	0.2989	6.8888	10.2586	0.0017	0.031	10.7913	13.4067
o3-mini	-0.1415	-11.0065	-6.5275	0.3114	5.8991	10.0086	0.0022	0.0361	10.6503	12.7118
o4-mini	-0.1575	-11.1034	-6.8124	0.3077	6.6326	10.1529	0.0016	0.0318	13.0692	15.4046
gemini-2.0 -flash-lite	-0.159	-11.0221	-6.7985	0.3295	5.9492	9.9632	0.0022	0.0314	10.2965	11.8706
gemini-2.0 -flash	-0.1568	-11.3614	-6.9346	0.3006	5.7701	10.4468	0.0017	0.0369	9.8964	11.7248
gemini-2.5 -flash-preview-04-17	-0.1668	-11.3	-7.0038	0.3095	5.9116	10.3558	0.002	0.0382	10.6944	13.0586
claude-3.5 -haiku-20241022	-0.1508	-10.7328	-6.5587	0.3154	6.1853	9.7866	0.0024	0.0362	9.6178	11.3382
claude-3.7 -sonnet-20250219	-0.163	-11.1406	-6.9356	0.2994	6.3334	10.639	0.0017	0.0387	9.909	11.8212
qwen2.5-7b -instruct-turbo	-0.1476	-11.2612	-6.82	0.3056	6.2185	9.5531	0.0021	0.0295	11.3235	13.8491
qwen2.5-72b -instruct-turbo	-0.1828	-11.5488	-7.3197	0.3038	5.775	10.2214	0.0023	0.0346	10.1142	12.2067
llama-3.2-3b -instruct-turbo	**-0.0613**	**-9.4438**	**-4.988**	**0.4024**	5.8505	7.9353	**0.0059**	**0.1647**	8.277	10.786
llama-3.3-70b -instruct-turbo	-0.1387	-11.0753	-6.5627	0.2919	5.9832	10.047	0.0026	0.0486	10.4182	12.2387
llama-4-maverick-17b -128e-instruct-fp8	-0.1617	-11.4891	-6.9562	0.3012	6.3082	9.7712	0.0018	0.0405	11.0169	13.4163
deepseek-v3	-0.1505	-11.2648	-6.855	0.3069	6.9381	**10.7089**	0.002	0.0428	10.5334	12.8404
Domain: Restaurant										
LDA	-0.2512	-13.9977	-9.2477	0.3385	7.7317	3.3516	0.002	-0.0072	14.8073	15.9644
BERTopic	**-0.1622**	-13.6309	-8.1259	0.3107	**3.9735**	6.7305	0.0116	0.1027	**3.614**	**4.0662**
gpt-4o	-0.1909	**-11.0746**	-6.9674	0.3076	5.63	8.7456	0.0054	0.0221	7.9977	9.1804
o3-mini	-0.2324	-12.5524	-8.0825	0.332	4.683	8.8362	0.0057	0.0269	5.2426	6.0082
o4-mini	-0.2271	-12.4864	-8.0232	**0.3451**	5.3975	8.8302	0.0054	0.0239	7.4767	8.3782
gemini-2.0 -flash-lite	-0.2160	-12.2063	-7.7869	0.3444	7.0630	8.9824	0.0066	0.0277	12.8091	15.0293
gemini-2.0 -flash	-0.2033	-11.925	-7.4767	0.3112	5.9880	9.1912	0.0059	0.0302	9.1393	10.6392
gemini-2.5 -flash-preview-04-17	-0.1933	-11.1135	**-6.9446**	0.3252	5.1914	9.1775	0.0055	0.0256	6.3861	7.3846
claude-3.5 -haiku-20241022	-0.2195	-12.1077	-7.7864	0.2943	5.4107	8.8726	0.0071	0.0321	6.7834	7.5376
claude-3.7 -sonnet-20250219	-0.2154	-11.7673	-7.5305	0.3233	5.2393	**9.7834**	0.0059	0.0300	5.9620	7.1544
qwen2.5-7b -instruct-turbo	-0.2049	-11.5481	-8.3208	0.3121	5.7103	8.4121	0.0054	0.0208	7.3481	8.5285
qwen2.5-72b -instruct-turbo	-0.2066	-11.7391	-7.4165	0.3107	5.3038	9.2297	0.0056	0.0248	6.6172	7.8053
llama-3.2-3b -instruct-turbo	-0.1981	-11.6759	-7.2913	0.3152	5.9292	6.1641	**0.0117**	**0.0339**	8.2703	9.5635
llama-3.3-70b -instruct-turbo	-0.1976	-11.656	-7.3195	0.2904	5.2444	7.8648	0.0059	0.0232	6.6544	7.6213
llama-4-maverick-17b -128e-instruct-fp8	-0.2311	-12.313	-7.9958	0.3319	4.9387	8.8113	0.0059	0.025	5.5613	6.3629
deepseek-v3	-0.1941	-11.4944	-7.2265	0.328	6.0934	8.8475	0.007	0.033	8.6119	10.1566

Table 7: Topic modeling benchmark performance with label-free metrics. Best performance is marked as **bold**, and second-best is underlined.

	Precision (↑)	Recall (↑)	F1 (↑)	Purity (↑)	ARI (↑)	MIS (↑)
Domain: Laptop						
LDA	0.3602	0.3565	0.3577	0.3770	0.1573	0.3344
BERTopic	0.5139	0.5084	0.5102	0.5033	0.2603	0.5262
gpt-4o	0.6429	0.6331	0.6364	0.6144	0.3380	0.6095
o3-mini	0.7114	0.7017	0.7049	0.6846	0.4685	0.6257
o4-mini	0.7308	0.7211	0.7243	0.7023	0.4743	0.6413
gemini-2.0 -flash-lite	0.6278	0.6188	0.6218	0.6042	0.2978	0.6013
gemini-2.0 -flash	0.7138	0.7036	0.7070	0.6857	0.4580	0.6315
gemini-2.5 -flash-preview-04-17	0.7362	0.7259	0.7293	0.7088	0.4622	0.6508
claude-3.5 -haiku-20241022	0.5811	0.5722	0.5752	0.5661	0.2565	0.5845
claude-3.7 -sonnet-20250219	0.7250	0.7148	0.7182	0.7063	0.4411	0.6530
qwen2.5-7b -instruct-turbo	0.6215	0.6127	0.6156	0.6199	0.3794	0.5710
qwen2.5-72b -instruct-turbo	0.7172	0.7068	0.7102	0.6921	0.4490	0.6375
llama-3.2-3b -instruct-turbo	0.3839	0.3768	0.3792	0.3770	0.1144	0.3675
llama-3.3-70b -instruct-turbo	0.7318	0.7213	0.7248	0.7030	**0.4882**	0.6342
llama-4-maverick-17b -128e-instruct-fp8	0.6997	0.6902	0.6934	0.6816	0.4479	0.6112
deepseek-v3	**0.7454**	**0.7347**	**0.7383**	**0.7208**	0.4814	**0.6543**
Domain: Restaurant						
LDA	0.4530	0.4505	0.4512	0.4812	0.1994	0.2397
BERTopic	0.6728	0.6679	0.6692	0.6334	0.4593	0.5190
gpt-4o	0.8045	0.8001	0.8015	0.7932	0.6377	0.6648
o3-mini	0.8276	0.8226	0.8245	0.8068	0.6648	0.6693
o4-mini	0.8172	0.8124	0.8139	0.7984	0.6384	0.6640
gemini-2.0 -flash-lite	0.8132	0.8077	0.8095	0.7902	0.6402	0.6706
gemini-2.0 -flash	0.8057	0.8007	0.8023	0.7861	0.6184	0.6609
gemini-2.5 -flash-preview-04-17	0.8149	0.8099	0.8115	0.8019	0.6426	0.6753
claude-3.5 -haiku-20241022	0.6987	0.6931	0.6948	0.6913	0.5019	0.5888
claude-3.7 -sonnet-20250219	**0.8386**	**0.8336**	**0.8353**	**0.8224**	**0.6805**	**0.6943**
qwen2.5-7b -instruct-turbo	0.7264	0.7201	0.7220	0.7027	0.5212	0.6001
qwen2.5-72b -instruct-turbo	0.7889	0.7842	0.7857	0.7806	0.6280	0.6552
llama-3.2-3b -instruct-turbo	0.4789	0.4743	0.4756	0.5380	0.3277	0.3955
llama-3.3-70b -instruct-turbo	0.7912	0.7858	0.7875	0.7667	0.6207	0.6795
llama-4-maverick-17b -128e-instruct-fp8	0.8109	0.8061	0.8077	0.7961	0.6525	0.6644
deepseek-v3	0.8317	0.8260	0.8278	0.8113	0.6737	0.6938

Table 8: Topic modeling benchmark performance with label-based metrics. Best performance is marked as **bold**, and second-best is underlined.

G Examples of Segment Intrusion Evaluation


```
{
"hard_single": [
    {
        "topic": "DRINKS#PRICES",
        "input_texts": [
            "The tuna and wasabe potatoes are excellent.",
            "well priced wines",
            "half off till 8pm",
            "Slightly above average wines start at $70+ with only one selection
            ↪  listed at $30+",
            "the Voss bottles of water were $8 a piece",
            "at reasonable prices"
        ],
        "intruder_idx": 0
    },
    {
        "topic": "FOOD#STYLE_OPTIONS",
        "input_texts": [
            "The menu is limited",
            "with large portions",
            "the antipasti were plentiful",
            "The pasta penne was pretty extra buttery, creamy",
            "the sake's complimented the courses very well and is successfully
            ↪  easing me into the sake world",
            "good selection of thin crust pizza including the Basil slice"
        ],
        "intruder_idx": 4
    },
    {
        "topic": "FOOD#QUALITY",
        "input_texts": [
            "The bagel was huge. They were served warm",
            "the sushi, which is great",
            "food was luke warm.",
            "Food was OK.",
            "Big Wong is a great place to eat and fill your stomach.",
            "And the Tom Kha soup was pathetic"
        ],
        "intruder_idx": 4
    },
    ...
],
...
}
```

H Inter-annotator agreements

Domain	easy_single	easy_double	hard_single	hard_double
Laptop	1.0000	1.0000	0.9519	0.8650
Restaurant	0.9514	0.9550	0.9753	0.8800

Table 9: Cohen's Kappa Scores for Inter-Annotator Agreement

I Full Results on Segment Intrusion Evaluation

I.1 F1

	SI-E	SI-H	DI-E	DI-H	Avg. F1
Domain: Laptop					
Human Performance	1.0000	0.9700	0.9900	0.8700	0.9575
gpt-4o	0.9700	0.8550	0.9500	0.7734	0.8871
o3-mini	0.9800	**0.9050**	0.9450	0.8099	0.9100
o4-mini	**0.9900**	0.8800	**0.9750**	**0.8218**	**0.9167**
gemini-2.0-flash-lite	0.8664	0.7087	0.9055	0.6793	0.7900
gemini-2.0-flash	0.9350	0.7800	0.9273	0.6828	0.8313
gemini-2.5-flash-preview-04-17	0.9850	0.8950	0.8179	0.6648	0.8407
claude-3.5-haiku-20241022	0.9167	0.7922	0.8878	0.6469	0.8109
claude-3.7-sonnet-20250219	0.9650	0.8250	0.9450	0.8089	0.8860
qwen2.5-7b-instruct-turbo	0.8515	0.6350	0.6000	0.3014	0.5970
qwen2.5-72b-instruct-turbo	0.9676	0.8300	0.9600	0.6914	0.8622
llama-3.2-3b-instruct-turbo	0.2625	0.2510	0.0921	0.0787	0.1711
llama-3.3-70b-instruct-turbo	0.9600	0.8100	0.9400	0.7111	0.8553
llama-4-maverick-17b-128e-instruct-fp8	0.5459	0.4450	0.3098	0.2129	0.3784
deepseek-v3	0.9550	0.8350	0.9100	0.7174	0.8544
Domain: Restaurant					
Human Performance	0.9800	0.9700	0.9700	0.8700	0.9475
gpt-4o	0.9350	0.7931	0.8928	0.7348	0.8389
o3-mini	0.9750	0.8212	0.9400	0.7463	0.8706
o4-mini	**0.9800**	**0.8371**	**0.9500**	**0.7718**	**0.8847**
gemini-2.0-flash-lite	0.7982	0.6257	0.7624	0.6667	0.7132
gemini-2.0-flash	0.9000	0.6650	0.8628	0.6377	0.7664
gemini-2.5-flash-preview-04-17	0.9750	0.8250	0.7627	0.6398	0.8006
claude-3.5-haiku-20241022	0.9051	0.7255	0.8200	0.6634	0.7785
claude-3.7-sonnet-20250219	0.9676	0.7750	**0.9500**	0.7299	0.8556
qwen2.5-7b-instruct-turbo	0.7481	0.5350	0.4010	0.3175	0.5004
qwen2.5-72b-instruct-turbo	0.9127	0.7400	0.9104	0.6586	0.8054
llama-3.2-3b-instruct-turbo	0.2707	0.2076	0.1049	0.0750	0.1645
llama-3.3-70b-instruct-turbo	0.9552	0.7419	0.8878	0.6489	0.8084
llama-4-maverick-17b-128e-instruct-fp8	0.5594	0.4559	0.2148	0.1982	0.3571
deepseek-v3	0.9500	0.7850	0.8557	0.7022	0.8232

Table 10: Benchmark results of segment intrusion tasks with F1 metric. Best performance is marked as **bold**, and second-best is underlined.

I.2 Recall

	SI-E	SI-H	DI-E	DI-H	Avg. R
Domain: Laptop					
Human Performance	1.0000	0.9700	0.9900	0.8700	0.9575
gpt-4o	0.9700	0.8550	0.9500	0.7850	0.8900
o3-mini	0.9800	**0.9050**	0.9450	0.8200	0.9125
o4-mini	**0.9900**	0.8800	**0.9750**	**0.8300**	**0.9188**
gemini-2.0-flash-lite	0.9400	0.8150	0.9100	0.7150	0.8450
gemini-2.0-flash	0.9350	0.7800	0.9250	0.7050	0.8363
gemini-2.5-flash-preview-04-17	0.9850	0.8950	0.7300	0.6000	0.8025
claude-3.5-haiku-20241022	0.9350	0.8100	0.8900	0.6550	0.8225
claude-3.7-sonnet-20250219	0.9650	0.8250	0.9450	0.8150	0.8875
qwen2.5-7b-instruct-turbo	0.8600	0.6350	0.6000	0.3150	0.6025
qwen2.5-72b-instruct-turbo	0.9700	0.8300	0.9600	0.7000	0.8650
llama-3.2-3b-instruct-turbo	0.5000	0.4950	0.1250	0.1100	0.3075
llama-3.3-70b-instruct-turbo	0.9600	0.8100	0.9400	0.7200	0.8575
llama-4-maverick-17b-128e-instruct-fp8	0.5500	0.4550	0.3400	0.2400	0.3963
deepseek-v3	0.9550	0.8350	0.9100	0.7300	0.8575
Domain: Restaurant					
Human Performance	0.9800	0.9700	0.9700	0.8700	0.9475
gpt-4o	0.9350	0.8050	0.8950	0.7550	0.8475
o3-mini	0.9750	0.8150	0.9400	0.7650	0.8738
o4-mini	**0.9800**	**0.8350**	**0.9500**	**0.7950**	**0.8900**
gemini-2.0-flash-lite	0.8900	0.7900	0.7700	0.7100	0.7900
gemini-2.0-flash	0.9000	0.6650	0.8650	0.6600	0.7725
gemini-2.5-flash-preview-04-17	0.9750	0.8250	0.6750	0.5950	0.7675
claude-3.5-haiku-20241022	0.9300	0.7400	0.8200	0.6800	0.7925
claude-3.7-sonnet-20250219	0.9700	0.7750	**0.9500**	0.7500	0.8613
qwen2.5-7b-instruct-turbo	0.7500	0.5350	0.4050	0.3350	0.5063
qwen2.5-72b-instruct-turbo	0.9150	0.7400	0.9150	0.6800	0.8125
llama-3.2-3b-instruct-turbo	0.4500	0.3950	0.1350	0.0950	0.2688
llama-3.3-70b-instruct-turbo	0.9600	0.7400	0.8900	0.6700	0.8150
llama-4-maverick-17b-128e-instruct-fp8	0.5650	0.4650	0.2400	0.2250	0.3738
deepseek-v3	0.9500	0.7850	0.8600	0.7250	0.8300

Table 11: Benchmark results of segment intrusion tasks with Recall metric. Best performance is marked as **bold**, and second-best is underlined.

I.3 Precision

	SI-E	SI-H	DI-E	DI-H	Avg. P
Domain: Laptop					
Human Performance	1.0000	0.9700	0.9900	0.8700	0.9575
gpt-4o	0.9700	0.8550	0.9500	0.7621	0.8843
o3-mini	0.9800	**0.9050**	0.9450	0.8000	0.9075
o4-mini	**0.9900**	0.8800	**0.9750**	**0.8137**	**0.9147**
gemini-2.0-flash-lite	0.8034	0.6269	0.9010	0.6471	0.7446
gemini-2.0-flash	0.9350	0.7800	0.9296	0.6620	0.8267
gemini-2.5-flash-preview-04-17	0.9850	0.8950	0.9299	0.7453	0.8888
claude-3.5-haiku-20241022	0.8990	0.7751	0.8856	0.6390	0.7997
claude-3.7-sonnet-20250219	0.9650	0.8250	0.9450	0.8030	0.8845
qwen2.5-7b-instruct-turbo	0.8431	0.6350	0.6000	0.2890	0.5918
qwen2.5-72b-instruct-turbo	0.9652	0.8300	0.9600	0.6829	0.8595
llama-3.2-3b-instruct-turbo	0.1779	0.1681	0.0729	0.0613	0.1201
llama-3.3-70b-instruct-turbo	0.9600	0.8100	0.9400	0.7024	0.8531
llama-4-maverick-17b-128e-instruct-fp8	0.5419	0.4354	0.2845	0.1912	0.3633
deepseek-v3	0.9550	0.8350	0.9100	0.7053	0.8513
Domain: Restaurant					
Human Performance	0.9800	0.9700	0.9700	0.8700	0.9475
gpt-4o	0.9350	0.7816	0.8905	0.7156	0.8307
o3-mini	0.9750	0.8274	0.9400	0.7286	0.8678
o4-mini	**0.9800**	**0.8392**	**0.9500**	0.7500	**0.8798**
gemini-2.0-flash-lite	0.7236	0.5180	0.7549	0.6283	0.6562
gemini-2.0-flash	0.9000	0.6650	0.8607	0.6168	0.7606
gemini-2.5-flash-preview-04-17	0.9750	0.8250	0.8766	0.6919	0.8421
claude-3.5-haiku-20241022	0.8815	0.7115	0.8200	0.6476	0.7652
claude-3.7-sonnet-20250219	0.9652	0.7750	**0.9500**	**0.7709**	0.8653
qwen2.5-7b-instruct-turbo	0.7463	0.5350	0.3971	0.3018	0.4951
qwen2.5-72b-instruct-turbo	0.9104	0.7400	0.9059	0.6385	0.7987
llama-3.2-3b-instruct-turbo	0.1935	0.1408	0.0857	0.0619	0.1205
llama-3.3-70b-instruct-turbo	0.9505	0.7437	0.8856	0.6291	0.8022
llama-4-maverick-17b-128e-instruct-fp8	0.5539	0.4471	0.1943	0.1772	0.3431
deepseek-v3	0.9500	0.7850	0.8515	0.6808	0.8168

Table 12: Benchmark results of segment intrusion tasks with Precision metric. Best performance is marked as **bold**, and second-best is underlined.

J Full Results Visualization on Segment Intrusion Evaluation

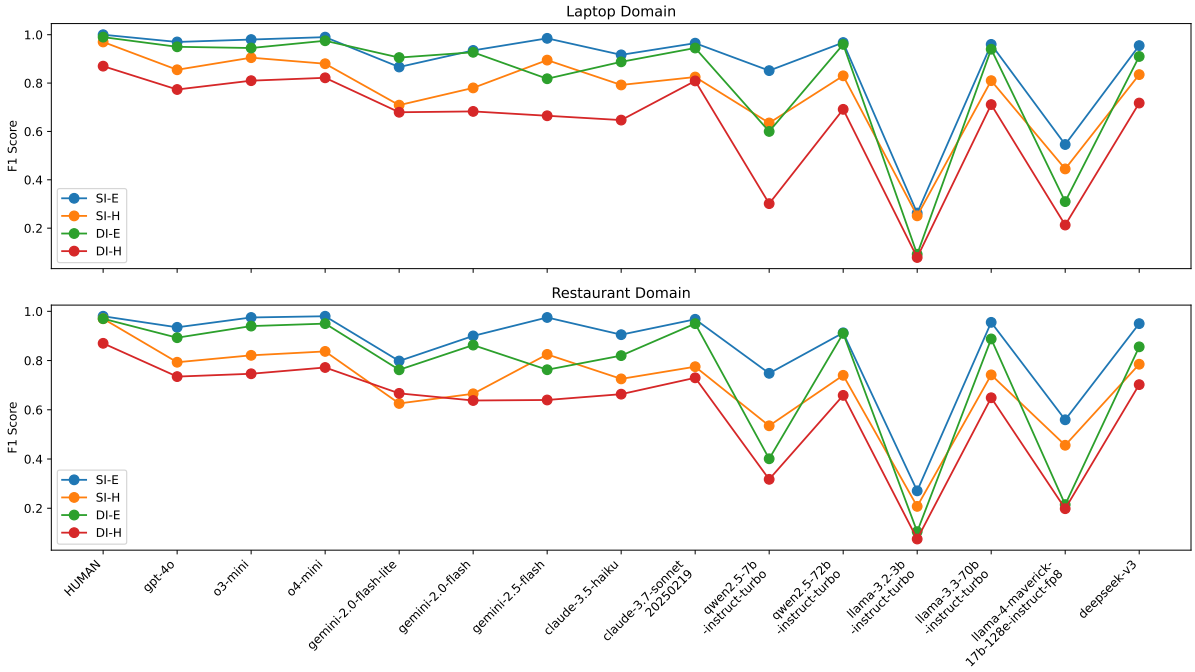

Figure 7: Visualized F1 performance of segment intrusion tasks. Human performance is calculated by averaging the annotations of two participants. 50 out of 200 instances for each task are randomly sampled in human evaluation.