From Clusters to Answers: Quantifying Topics

Neaton Jia Jun Ang, Wei Hao
Columbia University
neaton.ang@columbia.edu, wh2473@columbia.edu

Abstract

Customer support teams receive large volumes of repetitive tickets, making it essential to organize incoming queries into coherent topics and convert the most frequent ones into an effective FAQ. While BERTopic provides a scalable way to cluster ticket text, evaluating the *quality* of these clusters remains challenging, especially when human review is costly. This project develops a scalable framework for quantifying topic quality by combining traditional computed metrics with an LLM-as-a-judge approach. We design a custom LLM evaluation rubric to approximate human judgement and compare its performance against standard metrics such as cluster density, coherence, and intra–inter similarity measures. Using a sample of human-annotated labels as ground truth, we assess how well each metric reflects true cluster quality. Our results show that LLM-based scoring aligns most closely with human evaluations and provides a practical, automated method for selecting high-quality clusters for downstream FAQ generation.

1 Data Understanding & Preparation

1.1 Objectives

The primary objective of this phase is to develop a comprehensive understanding of the customer support ticket dataset. Specifically, we aim to:

- Identify the nature and composition of the data, including ticket types, sources, and coverage.
- Assess available information such as ticket text, metadata, and resolution notes for downstream use.
- Understand the dataset structure and the relationships between fields.
- Quantify distributions of key variables such as ticket categories and channels to detect imbalances.

1.2 Schema

The main fields used in this project are summarized in Table 1.

Table 1: Key fields used in the analysis.

Field	**Role in analysis**
Number	Unique identifier for each ticket; used as a primary key.
Created / Updated	Timestamps for creation and last update; used for time-series and activity analysis.
Incident State	Current ticket status (e.g., New, Active, Resolved); used for status tracking.
Assignment Group	Team or queue responsible for the ticket (e.g., CU Marketplace, AP Payments); used to examine load distribution.
Source	Channel through which the ticket was raised (e.g., Phone, Self-Service); not directly used in clustering.
Type / Subtype	Platform and more granular category (e.g., Jaggaer, Voucher Tactical Questions); used for segmentation and analysis.
Short Description	Concise summary written by the user; one of the key text fields for clustering.
Description	Detailed ticket body text; main text input to BERTopic.
Customer Communication, Work Notes, Resolution Notes	Conversation history and internal notes; not used in the current clustering pipeline.
User, Created By, Updated By	User and agent identifiers; not directly used in the analysis.

1.3 Data Investigation

1.3.1 Number of tickets per month

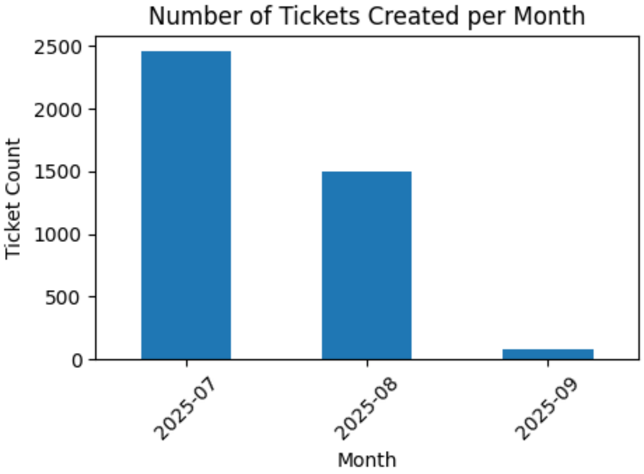

Figure 1: Number of tickets per month.

The monthly distribution of tickets is imbalanced, with September showing lower volume due to the dataset cutoff.

1.3.2 Tickets per assignment group

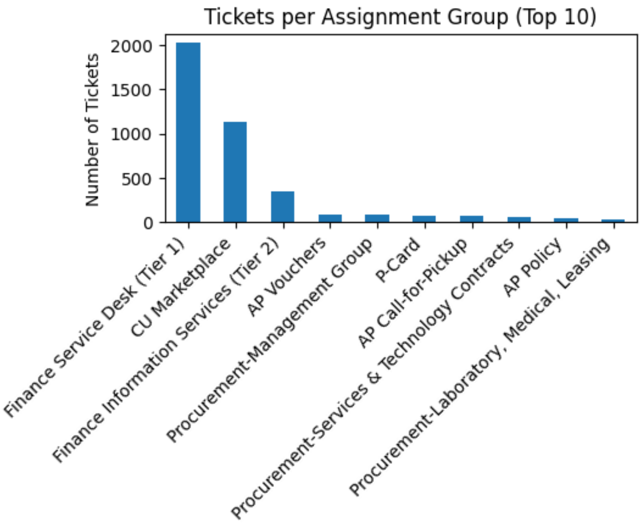

Figure 2: Tickets per assignment group.

Finance Team and CU Marketplace receive the most tickets.

1.3.3 Tickets per type

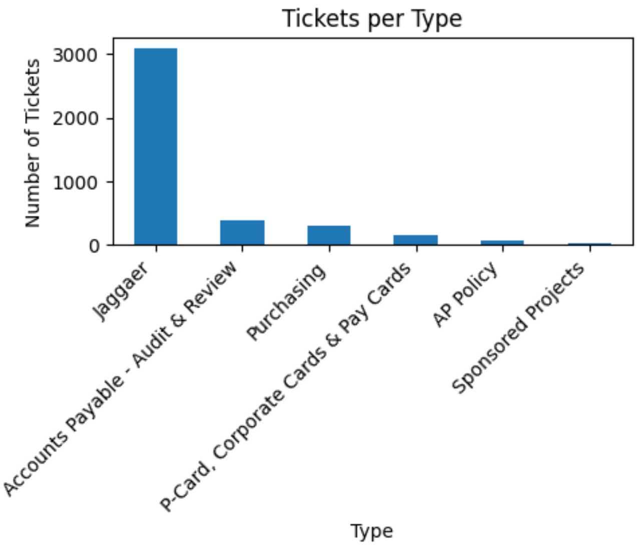

Figure 3: Tickets per type.

Jaggaer accounts for approximately 90% of all tickets.

1.3.4 Tickets by subtype

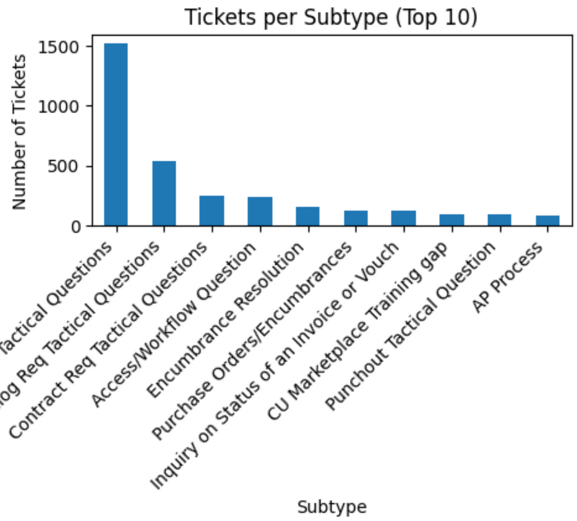

Figure 4: Tickets by subtype.

Voucher Tactical Questions dominate the dataset, followed by Non-Catalog and Contract Request Tactical Questions.

1.3.5 Word cloud of short descriptions

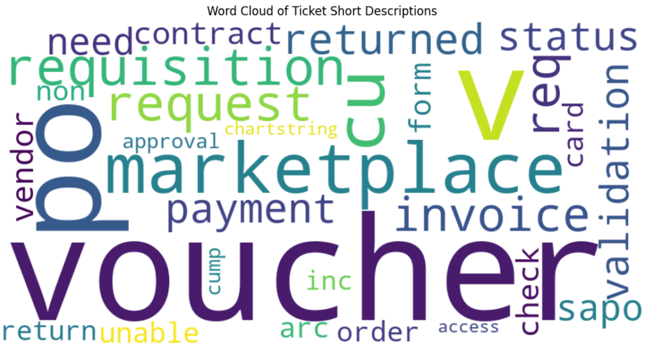

Figure 5: Word cloud of short descriptions.

Frequent terms include “status”, “voucher”, “payment”, and “invoice”, aligning with the subtype distribution.

1.4 Light Cleaning and Normalization

A light cleaning step was applied before clustering to standardize text fields while preserving meaning.

Light Cleaning

- Convert non-string inputs (e.g., NaN) into empty strings
- Trim leading and trailing whitespace
- Remove non-informative characters

Normalization

- Lowercase all text.

- Filter characters outside letters, numbers, and a small set of symbols.
- Collapse multiple spaces into a single space.

Sample Effects

Table 2: Sample effects of cleaning and normalization.

Original	Cleaned	Explanation
"Unable to reset Password!!!"	unable to reset password	Lowercased; punctuation removed
"EMAIL link doesn't work / expired??"	email link doesn t work / expired	Apostrophe normalized; spacing fixed

2 BERTopic Run

2.1 Model Setup

We applied BERTopic on the full dataset of customer support tickets using the concatenated text input:

$$\text{Document} = \text{Short Description} + \text{Description}$$

No manual rewriting, cleaning beyond light normalization, or domain-specific post-processing was applied.

BERTopic was run using its standard end-to-end pipeline:

- Transformer-based document embeddings
- UMAP for dimensionality reduction
- HDBSCAN for density-based clustering and topic discovery

This configuration allows the model to automatically infer cluster structure without specifying the number of topics in advance.

2.2 Topic Summarisation for LLM Judging

To enable reliable LLM-based evaluation, we first transform each BERTopic topic into a compact, structured summary that can be judged consistently. While BERTopic provides topic descriptors such as `Name`, `Top_n_words`, and `Representation`, these fields are often noisy and not directly human-readable as a decision basis for document-level judging.

For each topic, we construct a **tags distribution** string by aggregating the most common non-empty values of `Representation`, `Top_n_words`, and `Name` across documents assigned to that topic. These are concatenated into a single descriptor of the form:

`tags_distribution` = `Representation` | `Top_n_words` | `Name`

We then prompt an LLM to convert this descriptor into a structured topic summary in JSON with three fields: (i) `label` (a concise help-center-style topic name), (ii) `description` (the core tasks/problems covered), and (iii) `variability` (common edge cases or subtypes). This summarisation step produces a stable semantic representation (`topic_summary`) that is used as the reference text for downstream judging (intra-cluster fit, inter-cluster distinctiveness, and hybrid agreement).

To ensure consistent formatting and reduce parsing failures, the summary is requested under a fixed JSON schema and persisted per topic (with checkpointing) into a minimal FAQ sheet containing `topic_id`, `tags_distribution`, `topic_summary`, and associated computed metrics.

2.3 Evaluation Metrics

To assess topic quality, we computed several commonly used intrinsic metrics:

- **Coherence** (c_v) - measures semantic consistency of the top terms
- **Topic Diversity** - proportion of unique terms across topic keywords
- **Silhouette Score** - measures cluster separation based on embedding distance

These metrics provide initial signals about topic quality, though they each have known limitations when applied to short, noisy text such as support tickets.

2.4 Initial Results

The first comparison focused on the coherence score (c_v). However, the coherence values across discovered topics were extremely close to one another, making it difficult to distinguish good topics from weak ones using this metric alone. In practice, this is expected when:

- Many topics share overlapping vocabulary
- Documents are very short
- Clusters contain mixed or ambiguous ticket language

As a result, coherence was insufficient as a primary signal for topic quality. This motivated the move toward alternative evaluation approaches such as topic diversity, silhouette scores, and ultimately an LLM-based evaluation framework.

3 Deduplication with SemDeDup

3.1 Motivation

During the initial BERTopic run, the subtype *Voucher Tactical Questions* dominated the dataset. This suggested that many tickets may contain repeated or near-duplicate content. Excessive repetition can distort clustering by:

- Creating disproportionately large clusters

- Artificially boosting coherence (since duplicates increase term consistency)
- Reducing topic diversity due to repeated phrasing
- Lowering overall cluster quality by overwhelming the embedding space

To address this, we introduced a semantic deduplication step using SemDeDup.[1]

3.2 Method

SemDeDup identifies redundant documents using embedding similarity. Tickets whose embeddings fall within a small neighbourhood of another document are considered semantic duplicates and removed. This retains unique content while filtering out highly similar or repeated tickets.

From the original dataset, SemDeDup removed:

$$\mathrm{SemDeDup\ removed}\ 488\ \text{tickets}\ (12.10\%)\ \text{out of}\ 3545\ \text{kept.}$$

3.3 Quantitative Results

We evaluated topic quality both before and after deduplication using coherence, diversity, silhouette, and HDBSCAN outlier percentage. Results are summarised in Table 3.

Table 3: Topic quality before and after SemDeDup.

	c_v	Diversity	Silhouette	Outlier %
Before Dedup	0.49	0.62	0.059	25.49
After Dedup	0.51	0.58	0.052	32.02

3.4 Interpretation

Deduplication produced a mix of improvements and expected trade-offs:

Higher coherence. Removing duplicates increases semantic quality within each topic, resulting in a slight coherence boost (0.49 → 0.51). Topics become cleaner because repeated phrases no longer dominate the term distributions.

Lower diversity. Diversity dropped (0.62 → 0.58) due to fewer unique terms in the remaining documents. This is expected when removing large volumes of repeated templates or duplicate reports.

Lower silhouette score. Silhouette decreased slightly (0.059 → 0.052). Without dense groups of duplicate tickets, clusters appear less well-separated in embedding space. This reflects a more realistic view of the dataset rather than inflated separation.

Higher outlier percentage. Outliers increased from 25.49% to 32.02%. Deduplication reduces artificial cluster density created by duplicates, naturally causing more points to be classified as noise.

3.5 Summary

Deduplication removes redundant tickets that disproportionately influence clustering. The result is fewer large, homogeneous clusters and more balanced, semantically meaningful topics. While density-based metrics decrease slightly, the overall topic quality improves, indicating a trade-off of **quantity for quality** that better reflects the true structure of support-ticket text.

4 Additional Computed Metrics for Topic Quality

To complement intrinsic BERTopic scores, we introduced several computed metrics designed to capture different aspects of topic quality. These metrics were evaluated against human-annotated judgements to determine whether they serve as reliable proxies for human assessment.

4.1 Gini Coefficient (Topic Balance)

We define a Gini index over the distribution of document counts across topics. Let p_i be the proportion of documents assigned to topic i (using the POST-assignment topic frequencies). The Gini coefficient is:

$$G = \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} |p_i - p_j|}{2K \sum_{i=1}^{K} p_i}.$$

A higher Gini score indicates that a small number of topics dominate the corpus (i.e., cluster imbalance). Imbalanced topic distributions often signal that the model has failed to separate coherent subtopics and is collapsing content into oversized clusters.

4.2 Keyword / N-gram Frequency Score

To assess representativeness, we compute a keyword frequency score by summing the global corpus frequencies of a topic's top N words and normalising by the total token mass. Let $f(v)$ be the corpus frequency of token v, and $\mathrm{TopN}(t)$ the set of top N words for topic t. In code, this is stored as `kw_freq_score`:

$$\mathrm{kw\_freq\_score}(t) = \frac{1}{\sum_v f(v)} \sum_{w \in \mathrm{TopN}(t)} f(w).$$

Higher values indicate that a topic's top words cover a larger share of the overall corpus tokens, suggesting that the topic captures common and representative terminology rather than rare or noisy phrases.

4.3 Uniqueness Rate

We also measure how distinctive each topic's keywords are relative to other topics. After case-folding, we track in how

[1] https://github.com/facebookresearch/SemDeDup

many topics each word appears. For topic t, let $\text{TopN}(t)$ be its top N words and define:

$$\text{unique\_rate}(t) = \frac{\#\{w \in \text{TopN}(t) : w \text{ appears in topic } t \text{ only}\}}{|\text{TopN}(t)|}.$$

In the implementation, this value is stored as `unique_rate`. Higher values correspond to more distinctive topics with less keyword overlap; lower values suggest that topics share many of the same surface forms and may not be well-separated.

4.4 Fuzzy Topic Stability / Name Matching Score

To evaluate topic stability across runs, we compute a fuzzy matching score between topic names from a reference BERTopic run and those from a new run. Each topic name (constructed from top keywords) is compared using a string-distance measure such as Word Error Rate (WER) or Levenshtein distance. For a topic t in the new run, its stability score can be expressed as:

$$\text{Stability}(t) = 1 - \min_{s \in \mathcal{S}} \text{dist}(t, s),$$

where $\mathcal{S}$ is the set of topic names from the reference run and $\text{dist}(\cdot, \cdot)$ is a normalised string distance. Stable topics yield similar keyword-based labels across runs; unstable ones show high variance in naming.

4.5 Correlation with Human Annotation

To determine whether these metrics reflect human interpretation of topic quality, we manually annotated a subset of topics and computed correlations between each metric and the corresponding human quality ratings. The goal is to evaluate whether:

- Balanced topics (low Gini)
- Representative topics (high `kw_freq_score`)
- Distinctive topics (high `unique_rate`)
- Stable topics (high fuzzy name-matching score)

align with human-judged coherence and interpretability. This analysis indicates which automated metrics best approximate human evaluation and can be trusted for large-scale topic quality assessment.

4.6 Motivation for a Scalable Evaluation Framework

Ultimately, the goal of this work is to identify a reliable and scalable approach for evaluating clustering quality across many BERTopic configurations. Since each choice of parameters (UMAP, HDBSCAN, embedding model, deduplication settings) produces a different set of clusters, we require an evaluation framework that can compare dozens, or even hundreds, of topic models without manual intervention.

A practical challenge in topic modeling is the absence of ground truth labels. As a result, evaluation traditionally relies on:

- **Human annotation**: which is accurate but slow and costly
- **Computed metrics**: which are scalable but often weak proxies for true semantic quality
- **LLM-as-a-judge scoring**, which offers scalable semantic evaluation but requires careful prompt design and validation

To accelerate human annotation, topic summaries must be informative and representative; this allows annotators to quickly assess whether a cluster is coherent, meaningful, and distinct. However, relying exclusively on human ratings does not scale. Thus, our aim is to identify which automated metrics most strongly correlate with human judgement.

The workflow is:

1. Generate many clustering results using different parameter settings.
2. Compute intrinsic metrics (e.g., Gini, keyword frequency, uniqueness rate, silhouette, coherence).
3. Obtain human annotations for a subset of topics to serve as a reference.
4. Design an LLM-based evaluation rubric and optimise prompts for consistent scoring.
5. Measure correlations between:

$$\begin{aligned}&\text{(human scores)} \quad \text{vs} \\ &\{\text{computed metrics, LLM-judge scores}\}\end{aligned}$$

6. Select the metrics that best approximate human evaluations.

A scalable evaluation pipeline is essential because our goal is not merely to evaluate a single BERTopic run, but to compare a large number of clustering outputs produced under different configurations.

5 LLM-as-a-Judge Evaluation

While computed metrics capture structural or lexical properties of topics, they cannot reliably assess semantic coherence or distinctiveness. To address this, we designed an LLM-based evaluation framework that scores clusters according to their semantic fit and boundary clarity. The goal is to approximate human annotation at scale while ensuring consistency across many BERTopic runs.

We introduce three complementary LLM-derived metrics: (1) intra-cluster fit, (2) inter-cluster distinctiveness, and (3) hybrid agreement with BERTopic assignments.

5.1 Intra-Cluster Fit Score

For each topic, we generate a topic summary (based on its representative words and documents) and ask the LLM to judge whether each document in the cluster *fits* the proposed description. Two scores are produced:

- **Mean Fit Score**: Average LLM judgement of fit over all documents in the topic.

- **Fit Rate**: Proportion of documents that exceed a specified "fit" threshold.

Higher scores indicate that the cluster is semantically coherent and that its summary meaningfully reflects its contents. Because the LLM evaluates each document individually, the reliability of this metric increases with cluster size: larger clusters provide more evidence of internal consistency.

5.2 Inter-Cluster Distinctiveness Score

Semantic coherence alone is insufficient, topics must also be distinguishable from neighbouring clusters. For each topic, we identify its two nearest neighbouring topics in embedding space and ask the LLM to judge whether the topic's summary is meaningfully distinct from the summaries of these neighbours.

The rationale for focusing on the two closest topics is that semantic boundary errors typically arise from local neighbourhood overlap rather than distant, unrelated clusters. Evaluating against the two nearest neighbours captures the majority of meaningful distinctiveness issues while keeping the evaluation computationally tractable.

The LLM produces a **distinctiveness score**, reflecting how well the topic is separated from its neighbours in terms of meaning.

5.3 Hybrid Agreement Rate

The hybrid agreement metric evaluates whether a document is correctly assigned to its BERTopic cluster by comparing its semantic fit to the assigned topic versus the fit to the nearest neighbouring topics. This metric blends both intra-cluster coherence (fit to assigned topic) and inter-cluster separation (fit to competing topics), offering a holistic measure of assignment quality.

For each topic, we consider its k nearest neighbour topics in embedding space ($k = 2$ in our experiments). For each document assigned to topic t, the LLM produces:

- $F_t(d)$: Document-topic fit score for topic t,

- $F_{n_1}(d), F_{n_2}(d)$: Fit scores for the nearest neighbour topics n_1 and n_2.

A document is considered *correctly assigned* if:

1. $F_t(d)$ is defined
2. $F_t(d) \geq \max\{F_{n_1}(d), F_{n_2}(d)\}$
3. LLM judged the document as a semantic "fit" for topic t

If no neighbour fit is available (e.g., missing evaluation), correctness reduces to whether the document fits its assigned topic.

We also compute a decision margin:

$$\text{margin}(d) = F_t(d) - \max\{F_{n_1}(d), F_{n_2}(d)\}$$

A positive margin indicates strong agreement with the assigned topic; a negative margin indicates that a neighbour topic better explains the document.

Per-topic aggregation. For each topic t, we derive the following summary metrics:

- **Hybrid Agreement Rate**:

$$\text{HAR}(t) = \frac{\#\{d \in t : d \text{ agrees with its assignment}\}}{\#\{d \in t \text{ judged}\}}$$

- **Average Margin**:

$$\text{AvgMargin}(t) = \frac{1}{N_t} \sum_{d \in t} \text{margin}(d)$$

  computed over valid, finite margins.

- **Disagreement Count**: Number of documents where the LLM favoured a neighbour topic over the assigned topic.

Interpretation. The hybrid agreement metric is explicitly a blend of:

- **Intra-Cluster Fit** ("Does the document match its assigned topic?")

- **Inter-Cluster Competition** ("Is the document more semantically aligned with a nearby competing topic?")

A high hybrid agreement rate indicates that:

1. Documents are semantically consistent with their assigned topic

2. They are *more* consistent with their assigned topic than with any neighbouring topic

This provides a strong proxy for human judgement of topic boundaries, capturing both coherence *and* distinctiveness in a single scoring framework.

5.4 Summary

Together, the three LLM-based metrics provide:

- Semantic Coherence (Intra-Fluster Fit)

- Semantic Separation (Inter-Cluster Distinctiveness)

- Document-to-Topic assignment correctness (Hybrid Agreement)

These measurements enable scalable semantic evaluation across many BERTopic runs.

6 Optimising the LLM Judge with GEPA

Although LLM-based metrics provide scalable semantic evaluation, their accuracy depends heavily on the quality of the judging prompt. Since we possess a human-annotated subset of topics, we use these labels as a validation set to optimise the LLM judge. We adopt a GEPA-style (Generate-Evaluate-Perturb-Aggregate) optimisation loop, which iteratively improves the prompt based on cases where the LLM disagrees with human judgement.

6.1 Initial Evaluation of the Judge

We begin with a baseline judging prompt and run it across the human-annotated validation set. For each sample, we store:

- Input context (document, topic summary, neighbour summaries)
- LLM's judgement (`judge_label`)
- Human label (`human_label`)
- LLM's reasoning trace

We compute **human_fit**, defined as the agreement rate between the LLM's predictions and the human annotations.

6.2 GEPA Step 1: Generate (Identify Failure Cases)

GEPA begins by collecting all instances where the judge made an incorrect decision. We define the failure set as $\mathcal{F} = \{(x, y, \hat{y}) : \hat{y} \neq y\}$, where y is the human label and $\hat{y}$ is the LLM judge's label. These failure cases form the basis for reflective prompt revision.

6.3 GEPA Step 2: Evaluate (LLM Self-Reflection)

The LLM is then asked to examine only the failure cases and reflect on its own mistakes. The reflection prompt follows the GEPA framework:

Reflection instruction: "Here are examples where your judgment disagreed with human labels. Explain the causes of the disagreement. Rewrite your judging instructions to avoid these errors and better align with human decisions."

The output is a revised prompt variant π', which represents a self-improved judging strategy derived directly from analysing its previous errors.

6.4 GEPA Step 3: Perturb (Prompt Mutation)

GEPA treats the revised prompt π' as a *mutated* candidate. Different types of perturbations may be proposed by the LLM, including:

- Clarifying semantic boundaries
- Adding explicit rules for ambiguous cases
- Specifying how to treat documents with overlapping meaning
- Tightening examples or decision criteria.

Each mutation is intended to address a specific pattern observed in the error set $\mathcal{F}$.

6.5 GEPA Step 4: Aggregate (Re-evaluate and Select)

The revised prompt π' is re-evaluated on the same human-annotated validation set, yielding a new score human_fit(π'). If the revised prompt demonstrates better alignment with human labels, i.e., human_fit$(\pi') >$ human_fit(π), then π' replaces the previous prompt. Otherwise, the mutation is discarded.

This iterative loop continues until:

- No further improvement is observed OR
- A pre-set optimisation limit is reached

6.6 Outcome

Through GEPA optimisation, the judge prompt becomes progressively more aligned with human interpretation of cluster quality. The refined prompt is then used to score all clusters across all BERTopic configurations, enabling scalable semantic evaluation. The next section presents the correlation analysis between human annotations, computed metrics, and LLM-derived metrics using the optimised GEPA prompt.

7 Correlation Analysis Across All Metrics

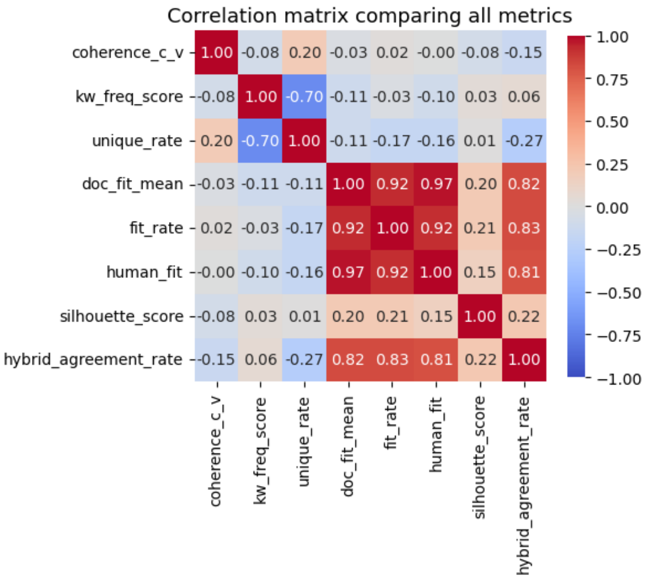

Figure 6: Correlation matrix comparing computed metrics, LLM-based evaluations, and human annotations.

Figure 6 shows the correlation matrix comparing all computed metrics, LLM-based evaluations, and human annotations. Although the matrix shown corresponds to one clustering configuration, we repeated this analysis across multiple BERTopic runs and observed highly consistent patterns. Importantly, after applying GEPA optimisation to the LLM judge, we observed a substantial improvement in alignment with human labels: on a representative random cluster, the correlation between LLM scores and human fit increased by 23.1%. This demonstrates the effectiveness of GEPA in refining the semantic decision boundaries used by the judge.

7.1 LLM-based metrics strongly correlate with human judgements

The LLM intra-cluster metrics (`doc_fit_mean` and `fit_rate`) show exceptionally high correlation with human labels:

$$r \approx 0.97 \text{ for doc\_fit\_mean}, \qquad r \approx 0.92 \text{ for fit\_rate}$$

This indicates that the LLM judge reliably mirrors human assessments of cluster quality. The hybrid agreement rate also shows strong correlation with human labels ($r = 0.81$), confirming that combining intra- and inter-cluster judgement captures semantic assignment correctness in a way that aligns with human evaluation.

7.2 Computed metrics correlate weakly with human ratings

Traditional intrinsic metrics such as coherence (c_v), silhouette score, and keyword frequency score exhibit little or no correlation with human labels ($|r| < 0.10$). This supports the observation that these metrics capture structural or lexical properties but fail to reflect semantic cluster quality.

The only partially meaningful computed metric is the uniqueness rate ($r \approx -0.16$), which still shows weak alignment and inconsistent behaviour across runs.

Several additional computed metrics were explored during preliminary experiments but are not reported here, as they demonstrated similarly weak or unstable correlations with human-fit scores on annotated samples. We therefore restrict our analysis to metrics that exhibited at least minimal signal during initial validation.

7.3 Hybrid agreement captures both coherence and distinctiveness

The hybrid agreement rate correlates:

- Strongly with intra-cluster fit ($r \approx 0.82$)
- Moderately with silhouette ($r \approx 0.22$)
- Weakly with topic distinctiveness proxies such as uniqueness rate

This confirms that the hybrid metric successfully blends intra-cluster and inter-cluster information, providing a more holistic measure of cluster quality than either component alone.

7.4 Summary

Across multiple BERTopic configurations, the same trends consistently emerge: LLM-based metrics show substantially stronger alignment with human judgement than traditional intrinsic metrics, which exhibit weak or inconsistent correlations. This robustness across runs reinforces the conclusion that the LLM-as-a-judge framework-especially after GEPA prompt optimisation-provides a scalable and semantically reliable method for evaluating topic models under a wide range of clustering parameters.

8 Trusting Hybrid Agreement as a Quality Signal

To evaluate whether the hybrid agreement rate provides a reliable and human-aligned measure of clustering quality, we performed a controlled experiment across several BERTopic runs. We selected four clusterings whose coherence scores (c_v) were intentionally similar (around 0.50), making them difficult to distinguish using traditional intrinsic metrics. These clusterings also varied substantially in the number of discovered topics, mirroring realistic differences produced by parameter changes.

Table 4 summarizes the hybrid-agreement rates and coherence scores for the four clusterings.

Table 4: Comparison of clusterings with similar coherence but differing hybrid agreement.

Cluster	# Topics	Hybrid-Agreement	Coherence
Cluster 1	35	0.604	∼0.54
Cluster 2	27	0.573	∼0.51
Cluster 3	52	0.51	∼0.50
Cluster 4	40	0.50	∼0.51

8.1 Hybrid Agreement Provides Clear Differentiation

Although the coherence values are nearly indistinguishable, the hybrid-agreement metric exhibits meaningful separation across clusterings, producing a ranking:

$$\text{Cluster 1} > \text{Cluster 2} > \text{Cluster 3} \approx \text{Cluster 4}.$$

This demonstrates that hybrid agreement is more sensitive than coherence to differences in both intra-cluster semantic fit and inter-cluster boundary clarity. Clustering quality that is invisible to coherence becomes visible under the hybrid metric.

8.2 Human Annotation Confirms the Ranking

To validate whether hybrid agreement reflects human judgement, each clustering was evaluated by human annotators. Because Clusters 1 and 2 were initially tied in hybrid agreement and exhibited very similar characteristics, they underwent a more rigorous assessment: annotators compared them independently across a one-week period, with each comparison repeated three times.

The majority vote identified **Cluster 1** as the superior clustering,exactly matching the hybrid-agreement ranking.

This one-to-one alignment between human preference and hybrid-agreement scores provides strong evidence that hybrid agreement captures meaningful, semantic cluster quality that humans can perceive but coherence cannot.

8.3 Conclusion

The experiment confirms that:

- Coherence (c_v) is not sensitive enough to differentiate clusterings with similar lexical structure
- Hybrid agreement produces a clear and stable ranking
- Hybrid agreement rate *correlates with human judgement*

Therefore, hybrid agreement serves as a trustworthy, scalable, and semantically aligned metric for evaluating topic-model cluster quality, and can be used to compare large numbers of BERTopic configurations when human annotation is not feasible.

9 Limitations and Future Work

Although the proposed evaluation framework demonstrates strong alignment with human judgement and provides a scalable alternative to manual annotation, several limitations remain.

9.1 Limitations

Small annotated dataset. Our human-labelled dataset is relatively small, which restricts the statistical power of correlation analyses and limits the diversity of linguistic patterns exposed to the LLM judge. Larger and more varied annotated sets would enable more reliable prompt optimisation under GEPA and better generalisation across domains.

Cost of LLM-based evaluation. While LLM-as-a-judge provides semantically rich scoring, it introduces non-trivial computational and monetary costs. Running thousands of document-topic evaluations becomes expensive for large-scale datasets, and prompt optimisation (GEPA) amplifies this cost due to repeated judge calls.

Dependence on model behaviour. LLM judgements can be sensitive to prompt formulation and model versioning. Although GEPA ameliorates this by adapting the prompt to human-labelled failures, the evaluation pipeline still depends on the stability and consistency of the underlying LLM.

Cluster size imbalance. Topics with extremely small or large document counts may yield unstable LLM scores or distorted cluster summaries. Further work is needed to normalise or adjust metrics across clusters of varying sizes.

9.2 Future Work

Expanding human annotation. A natural next step is to collect a larger, more diverse human-annotated set, allowing more robust comparisons and improved GEPA optimisation. This would also enable more sophisticated human-LLM calibration, such as conditional scoring or hierarchical annotation protocols.

Cost reduction strategies. Several avenues exist for reducing LLM evaluation cost:

- Using smaller specialist models or distilled judges
- Caching and reusing doc-summary comparisons across runs
- Active-selection strategies that judge only “borderline” documents
- Adaptive sampling of documents within large clusters

These approaches could significantly lower the computation required to evaluate many clustering configurations.

Extending to other clustering methods. Although we focus on BERTopic, the evaluation framework is model-agnostic. Future work could apply hybrid agreement scoring to alternative topic models, sentence-embedding clusters, hierarchical topic structures, or dynamic topic tracking over time.

Improving robustness of LLM judges. Further improvements may include:

- Designing more rigorous prompt templates informed by error taxonomies
- Training small adapter models to mimic human decisions (LLM-to-LLM distillation),
- Incorporating self-consistency or debate-style evaluation.

Automated selection of optimal clustering parameters. With a validated evaluation metric, a promising direction is to integrate hybrid-agreement scoring into parameter search loops, enabling automated selection of BERTopic hyperparameters via Bayesian optimisation or evolutionary search.

Overall, while the current framework proves effective even at small scale, future extensions will focus on improving robustness, reducing cost, and broadening applicability to more datasets and clustering regimes.

References

[1] Maarten Grootendorst. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794, 2022. `https://arxiv.org/abs/2203.05794`.

[2] Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, Omar Khattab. *GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning*. arXiv:2507.19457, 2025. `https://arxiv.org/abs/2507.19457`.

[3] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, Ari S. Morcos. *SemDeDup: Data-efficient learning at web-scale through semantic deduplication*. arXiv:2303.09540, 2023. `https://arxiv.org/abs/2303.09540`.

[4] Thomas Compton. *Holistic Evaluations of Topic Models*. arXiv:2507.23364, 2025. `https://arxiv.org/abs/2507.23364`.

[5] Qipeng Zhu, Yanzhe Chen, Huasong Zhong, Yan Li, Jie Chen, Zhixin Zhang, Junping Zhang, Zhenheng Yang. *UniAPO: Unified Multimodal Automated Prompt Optimization*. arXiv:2508.17890, 2025. `https://arxiv.org/abs/2508.17890`.

[6] Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, Michael Zeng. *Automatic Prompt Optimization with "Gradient Descent" and Beam Search*. arXiv:2305.03495, 2023. EMNLP 2023. `https://arxiv.org/abs/2305.03495`.

[7] Ernest A. Dyagin, Nikita I. Kulin, Artur R. Khairullin, Viktor N. Zhuravlev, Alena N. Sitkina. *Automatic Prompt Optimization with Prompt Distillation*. arXiv:2508.18992, 2025. `https://arxiv.org/abs/2508.18992`.

10 Software and Reproducibility

The code used for this paper is publicly available in the GitHub repository linked below. The repository contains end-to-end scripts for:

- Data preprocessing and light normalization
- BERTopic configuration and topic clustering
- Semantic deduplication using SemDeDup
- Computation of cluster-level evaluation metrics
- Topic summarisation via large language models
- LLM-as-a-judge evaluation (intra-cluster, inter-cluster, hybrid agreement)
- GEPA-based prompt optimisation

All intermediate artifacts, including topic summaries, per-topic document dumps, embeddings, and evaluation outputs, are produced deterministically given a fixed random seed. Detailed instructions for reproducing each experiment are provided in the repository documentation.

Code repository: `https://github.com/neatonang/hybrid_agreement_clustering.git`

Acknowledgements

We thank the authors and maintainers of BERTopic, SemDeDup, and related open-source libraries for making their tools publicly available. We also acknowledge the developers of large language models, and APIs used in this work. Human annotations were performed manually and independently to ensure unbiased evaluation.