From Keyterms to Context: Exploring Topic Description Generation in Scientific Corpora

Pierre Achkar[1,2] **Satiyabooshan Murugaboopathy**[1] **Anne Kreuter**[1]
Yuri Campbell[1] **Tim Gollub**[3] **Martin Potthast**[4,5,6]

[1]Fraunhofer ISI [2]Leipzig University [3]Bauhaus-Universität Weimar
[4]University of Kassel [5]hessian.AI [6]ScaDS.AI, Leipzig

Abstract

Topic models represent topics as ranked term lists, which are often hard to interpret in scientific domains. We explore **Topic Description for Scientific Corpora**, an approach to generating structured summaries for topic-specific document sets. We propose and investigate two LLM-based pipelines: *Selective Context Summarisation* (SCS), which uses maximum marginal relevance to select representative documents; and *Compressed Context Summarisation* (CCS), a hierarchical approach that compresses document sets through iterative summarisation. We evaluate both methods using SUPERT and multi-model LLM-as-a-Judge across three topic modeling backbones and three scientific corpora. Our preliminary results suggest that SCS tends to outperform CCS in quality and robustness, while CCS shows potential advantages on larger topics. Our findings highlight interesting trade-offs between selective and compressed strategies for topic-level summarisation in scientific domains. We release code and data for two of the three datasets.[1]

1 Introduction

Gaining an overview of large scientific corpora is useful for exploring research areas, identifying common methodologies, and tracking emerging developments. A common entry point is topic modeling, which reveals underlying topics and presents them as ordered lists of terms. Algorithms such as Latent Dirichlet Allocation (LDA; Blei (2012)), Contextualised Topic Models (CTM; Bianchi et al. (2021)) and BERTopic (Grootendorst, 2022) are widely used for this purpose. While effective for organising unlabelled data, these methods only provide term-based topic representations, making them difficult to interpret (Chang et al., 2009). Most topic modeling pipelines stop at this level, which limits their usefulness for knowledge-intensive tasks, particularly in scientific domains where understanding a research topic requires insight into research goals, methods, and purposes.

Recent work has sought to improve interpretability by enriching topic representations with machine-generated labels or short contextual snippets (Lau et al., 2011; Popa and Rebedea, 2021; Rosati, 2022; Azarbonyad et al., 2023). However, these approaches often rely on surface-level signals, lack domain-specific grounding and fail to incorporate document-level context. Consequently, they offer limited support for understanding the underlying content of complex domains such as science.

In this work, we explore **Topic Description for Scientific Corpora**, an approach that aims to generate structured and informative summaries for topics derived from topic models. These descriptions enrich the topic representation by incorporating document-level context while remaining aligned with the topic terms, offering a clearer view of the underlying research themes.

To this end, we propose and investigate two pipelines based on large language models (LLMs). In both, we expand on prior art for multi-step multi-document generative summarization (Zhang et al., 2024) into the extreme cases of hundreds (sometimes up to thousands) of documents, while using topic representation as guidance. The first, *Selective Context Summarisation* (SCS), uses Maximum Marginal Relevance (MMR; Carbonell and Goldstein (1998)) to select a representative subset of topic documents prior summarisation. The second, *Compressed Context Summarisation* (CCS), inspired by hierarchical summarisation approaches such as RAPTOR (Sarthi et al., 2024), applies recursive summarisation over a hierarchy constructed from the topic's documents.

We evaluate these pipelines across three topic modeling backbones—CTM (Bianchi et al., 2021), BERTopic (Grootendorst, 2022), and Top-

[1]https://github.com/pierre-achkar/newsumm-25-scs-ccs

Proceedings of The 5th New Frontiers in Summarization Workshop, pages 102–122

icGPT (Pham et al., 2024)—on three scientific corpora. Focusing on reference-free evaluation, we conduct assessment using SUPERT (Gao et al., 2020), a reference-free semantic similarity metric, and a multi-model LLM-as-a-Judge framework using open-source models. Our preliminary results suggest that the MMR-based pipeline consistently produces more focused and concise topic descriptions than the hierarchical approach. We also analyze how topic-level properties, such as size and cohesion, affect topic description quality, and complement our findings typifying topic descriptions characteristics and error sources in each pipeline.

Our contributions include:

- We explore **Topic Description for Scientific Corpora** as a promising approach for enriching topic model outputs with structured, interpretable, document-grounded summaries.
- We propose and compare two LLM-based approaches - **SCS** and **CCS** - for topic-level summarisation in scientific corpora.
- We systematically evaluate how **topic characteristics** (e.g., number of source documents, topic cohesion) influence the effectiveness of different summarisation strategies.
- We find that while SCS generally outperforms CCS, the hierarchical approach becomes **competitive for large or low-cohesion topics**, offering guidance for method selection.

2 Related Work

We review prior work on topic modeling, enhanced topic representations, and multi-document scientific summarisation. Our work explores connections between these areas by combining topic model outputs with LLM-based summarisation to enrich topic representations.

2.1 Topic Modeling

Topic modeling is widely used for uncovering thematic structure in large text collections. Latent Dirichlet Allocation (LDA; (Blei, 2012)) remains a foundational model, assuming documents are mixtures of latent topics and topics are distributions over words. Contextualized Topic Models (CTM; (Bianchi et al., 2021)) extend this framework by incorporating document embeddings from pre-trained language models such as BERT (Devlin et al., 2019) and Sentence-BERT (Reimers and Gurevych, 2019). BERTopic (Grootendorst, 2022) clusters BERT embeddings for document topic assignment, while TopicGPT (Pham et al., 2024) employs decoder-only LLMs to directly generate topics. These models are applied across various domains, including scientific literature.

2.2 Enriching Topic Representations

Beyond term lists, several methods aim to create more interpretable topic representations. Early work retrieved candidate labels from external sources such as Wikipedia and ranked them by relevance to topic terms (Lau et al., 2011; Bhatia et al., 2016). Later approaches used generative models to create more descriptive labels from topic terms (Alokaili et al., 2020). BART-TL (Popa and Rebedea, 2021) fine-tunes a BART model using weakly supervised training signals derived from heuristic labels. In the scientific domain, topic interpretation often involves producing richer textual outputs. One method clusters citation statements and summarizes them using Longformer to reflect citation intent (Rosati, 2022). Topic Pages (Azarbonyad et al., 2023) construct structured descriptions by combining definition extraction using SciBERT with contextual snippets and co-occurrence-based linking. LimTopic (Azhar et al., 2025) applies BERTopic and LLMs to generate titles and summaries for topics in scientific limitation sections. Our work investigates using LLMs to generate document-grounded structured topic descriptions reflecting research methods, purposes, and objects.

2.3 Multi-Document Scientific Summarisation

Multi-document scientific summarisation (MDSS) synthesizes coherent summaries from clusters of scientific papers. Transformer-based methods such as KGSum (Wang et al., 2022) encode documents into knowledge graphs and use two-stage decoding for improved coherence. PRIMERA (Xiao et al., 2022) applies entity-level masking during pretraining to improve salience modeling, and its effectiveness extends to domain-specific datasets such as Multi-XScience (Lu et al., 2020). Hybrid pipelines combine extractive and abstractive stages. A biomedical-focused system combines BERT-based extraction with a PEGASUS decoder for summarisation (Shinde et al., 2022), while SKT5SciSumm (To et al., 2024) uses SPECTER (Cohan et al., 2020) embeddings for clustering followed by T5-based generation, outperforming larger models like GPT-4 on some tasks. The 3A-

COT framework (Zhang et al., 2024) structures LLM prompting into Attend–Arrange–Abstract stages to improve factuality and reduce redundancy. We adapt this framework in our setting with minor adjustments to its prompt templates to generate a unified, structured output appropriate for our context. Moreover, we build on these recent LLM-based MDSS advances, adapting them to topic modeling settings, which confers guidance given by topic representations.

3 Task Formulation

We investigate **Topic Description for Scientific Corpora** as an approach to generating structured, interpretable summaries for topic model outputs. Given a topic model applied to a scientific corpus $\mathcal{D}$, each topic T_k is characterized by:

- A set of topic-specific documents $D_k = \{d_1, d_2, \cdots\}$, where each document d_i is assigned to a single dominant topic,
- A ranked list of topic terms $W_k = w_1, \ldots, w_n$, also referred to as topic representation.

Our investigation focuses on generating topic descriptions S_k that summarise the main content of D_k, remain aligned with W_k, and follow a unified structure across topics. Each description includes a brief introduction to the topic, followed by the key research objects, methods, and purposes reflected in the underlying documents. This approach facilitates systematic exploration of scientific corpora.

For our evaluation, we examine four key quality dimensions. **Relevance** assesses whether descriptions accurately reflect topic aspects by incorporating topic terms meaningfully. **Factuality** examines grounding in original documents without unsupported claims. **Coherence** considers logical flow and consistency in presenting unified topic explanations. **Fluency** evaluates linguistic quality, seeking clear, accessible language that balances readability with technical precision.

4 Methodology

Given that the amount of documents in scientific topics may vary from dozens to thousands, we choose to explore methods that are flexible and capable of circumventing LLM context window limits. For this end, we propose and investigate two LLM-based approaches for generating topic descriptions from sets of documents associated with each topic: *Selective Context Summarisation (SCS)*, which uses **Maximum Marginal Relevance** (MMR; (Carbonell and Goldstein, 1998)) to select a small, diverse subset of representative documents, and *Compressed Context Summarisation (CCS)*, which builds a hierarchical structure over all topic documents using recursive clustering and abstraction, inspired by hierarchical summarisation approaches such as RAPTOR (Sarthi et al., 2024). Both methods operate independently of the underlying topic modeling backbone.

In both pipelines, the generation process is guided by the same multi-step prompt-chain template, adapted from the 3A-COT framework (Zhang et al., 2024), with topic terms provided as guidance. The full prompts are provided in Appendix A. An overview of the pipelines is shown in Figure 1.

4.1 Selective Context Summarisation (SCS)

SCS builds on an existing integration of LLMs into topic representation, as implemented in the BERTopic library[2]. In the original implementation, representative documents for each topic are selected and passed to an LLM alongside topic terms to generate a short label. We extend this approach to generate informative topic descriptions that summarise the core content of each topic.

Given a topic, we select the ten highest-ranked terms and concatenate them to form a single string. This is then embedded using a pre-trained sentence embedding model. All documents within the topic are embedded in the same vector space and those most similar to the topic vector are retrieved.

To ensure the selected documents are both relevant and diverse, we apply Maximum Marginal Relevance (MMR; (Carbonell and Goldstein, 1998)). MMR iteratively selects documents that are similar to the topic vector while penalizing redundancy with respect to previously selected documents. This results in a representative and non-redundant subset of documents that captures the breadth of the topic and fits within the context window of the LLM.

In the generation process, we use the top 10 most representative documents and the top 10 most relevant topic terms for each topic. These are inserted into the shared prompt-chain template (see Appendix A) and passed to the LLM, which generates the description based on this context.

[2] https://maartengr.github.io/BERTopic/getting_started/representation/llm.html

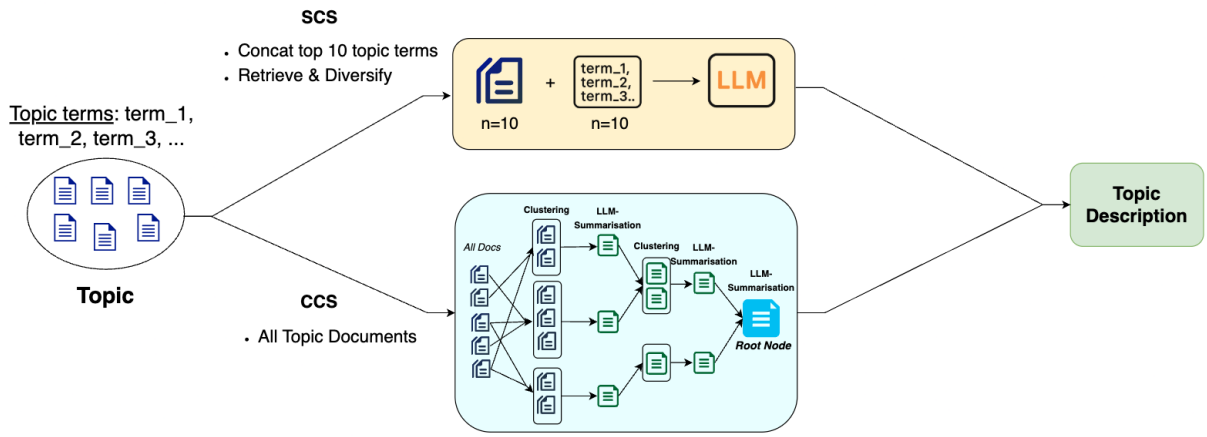


Figure 1: Overview of the two topic description pipelines. *SCS* selects a representative subset of documents using MMR and summarises them with an LLM. *CCS* summarises all topic documents via hierarchical clustering and recursive abstraction.

4.2 Compressed Context Summarisation (CCS)

The Compressed Context Summarisation method adapts hierarchical summarisation strategies to organize documents associated with each topic and generate descriptive summaries. Drawing inspiration from tree-based indexing approaches like RAPTOR (Sarthi et al., 2024), our method constructs a recursive hierarchy of summaries through iterative clustering and abstraction.

Unlike approaches that begin by segmenting long documents into smaller chunks, we start directly from short documents assigned to each topic (e.g., abstracts), without additional segmentation. These documents are embedded and projected into lower-dimensional space using UMAP (McInnes et al., 2020) to improve clustering quality.

The projected embeddings are then clustered using Gaussian Mixture Models (GMMs), which support soft assignment, allowing documents to belong to multiple clusters. Each cluster is summarised using an LLM, with the top 10 topic terms provided at each stage for additional guidance. This produces an abstract summary that captures the main content of the clustered documents. These summaries are recursively re-embedded and re-clustered, forming a tree structure in which each internal node summarises its child nodes.

This recursive summarisation continues until only one cluster remains or no further abstraction is necessary. We introduce a final root node at the top of the tree, which serves as the output of the method: a topic description generated by the LLM that summarises the top-level content in the tree.

By including all topic documents and organizing them hierarchically, CCS addresses LLM context length limitations and produces descriptions grounded in the complete topic context. The prompt-chain template used for all summarisation steps is the same one used in SCS.

5 Experimental Setup

To explore the effectiveness and generalisability of the topic description pipelines, we conduct experiments across diverse scientific domains and topic modeling backbones. This section describes the datasets, modeling configurations, and models used for generation and embedding.

5.1 Datasets

We evaluate our approaches on three domain-specific scientific corpora, using the abstracts of English-language research papers. Each dataset covers a distinct field to examine generalizability across different scientific domains.

ACL Anthology The ACL Anthology[3] contains publications in computational linguistics and NLP from conferences such as ACL, EMNLP, and NAACL. We use the official GitHub version, extracting metadata and abstracts. Non-English entries and missing abstracts are removed, resulting in 52,126 clean abstracts.

NIPS Papers The NIPS Papers Dataset[4] includes papers from the Neural Information Processing

[3] https://github.com/acl-org/acl-anthology/tree/master/python

[4] https://www.kaggle.com/datasets/benhamner/nips-papers

Systems (NIPS) between 1987 and 2016. We retain only English abstracts, removing missing entries and performing basic preprocessing. The final dataset contains 3,916 abstracts.

Quantum Computing Domain experts curated this dataset using a Boolean query on Scopus to retrieve recent papers (2010–2024) on quantum computing hardware. We retain only unique English abstracts, yielding 45,830 documents. Due to licensing restrictions, the dataset cannot be released; the full query is provided in Appendix B.

5.2 Topic Modeling

In order to provide a diverse comparison testbed of different topic modelling approaches, we have selected three backbones: CTM, BERTopic and TopicGPT. Each variant is based on a different topic modelling method: classical bag-of-words statistical estimation (CTM); clustering of vector representations of texts (BERTopic); and multi-step zero-shot topic generation (TopicGPT). Following prior art (Grootendorst, 2022; Pham et al., 2024), we assign each document to its most prominent topic to ensure comparability among the three backbones. We apply each topic modelling method to each dataset, resulting in nine topic models. Further information on the characteristics, hyper-parameter optimisation and evaluation of all topic models can be found in Appendix C.

This experimental setup comprises a diverse range of scenarios to examine the various challenges involved in creating topic descriptions.

5.3 LLM and Embedding Models

We use the `DeepSeek-V3` (DeepSeek-AI et al., 2024) model to generate topic descriptions across all pipelines. For embedding-based retrieval, we use `ModernBERT` (Warner et al., 2024a), a competitive model for sentence-level semantic similarity.

6 Evaluation Strategy

Evaluating topic descriptions presents inherent challenges due to the lack of gold-standard references and the wide variation in topics across different domains. We explore reference-free evaluation metrics that assess quality without requiring human-written summaries. We adopt two complementary strategies: **SUPERT**, a semantic similarity metric designed for multi-document summarisation, and an **LLM-as-a-Judge** framework, which uses prompting-based evaluation with LLMs.

6.1 SUPERT

SUPERT (Gao et al., 2020) is a reference-free evaluation metric developed for multi-document summarisation tasks. It creates a pseudo-reference by selecting key sentences from input documents and compares generated summaries based on their semantic similarity to this reference. The similarity is computed using contextualized embeddings and soft token alignment. SUPERT has been shown to align well with human judgments of relevance, making it well-suited for assessing how much essential content is preserved in a topic description.

6.2 LLM-as-a-Judge

We build on recent work from the Eval4NLP 2023 Shared Task (Leiter et al., 2023), which explored prompting LLMs as explainable and reference-free evaluation metrics. Our setup draws inspiration from promising systems (Kim et al., 2023), which demonstrated that zero-shot prompting, fine-grained scoring, and deterministic decoding can improve alignment with human preferences.

To align evaluation with our task formulation, we assess topic descriptions along four dimensions: **Relevance**, **Factuality**, **Coherence**, and **Fluency**. These criteria correspond to the aspects outlined in Section 3, and reflect qualities we consider important for topic descriptions. We compute the **Mean Aspect Score (MAS)**, as the average across these four evaluation dimensions.

When selecting an LLM-as-a-judge model, we prioritized open-source models with strong alignment to human judgment. To this end, we chose `Qwen2.5-7B-Instruct` (Yang et al., 2024), which achieved high alignment among open-source models in the LLMEval benchmark (Gu et al., 2024). To account for variability in model outputs, we included two additional models. Our first choice was the Orca family, as both `Orca-13B` and `OpenOrca-Platypus2-13B` have shown promising alignment in prior studies (Kim et al., 2023; Leiter and Eger, 2024). However, due to their 4k context window limitations, we selected `Mistral-7B-OpenOrca`[5], which maintains similar alignment while supporting longer contexts (32k). As a third model from a different architecture line, we added `Gemma-3-27B` (Kamath et al., 2025) to ensure diversity across the various model families.

As it is not feasible to evaluate a generated description against all documents associated with a

[5] https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca

Method	Metric	ACL			NIPS			Quantum		
		CTM	BERTopic	TopicGPT	CTM	BERTopic	TopicGPT	CTM	BERTopic	TopicGPT
SCS	SUPERT	**0.475**	**0.477**	**0.508**	**0.459**	0.465	**0.519**	**0.489**	**0.486**	**0.557**
	MAS-Qwen	**50.645**	**62.561**	**72.004**	56.703	**71.962**	**81.612**	**57.088**	**64.319**	78.650
CCS	SUPERT	0.467	0.474	0.501	0.453	**0.469**	0.515	0.458	0.484	0.552
	MAS-Qwen	49.612	62.400	65.299	**58.766**	66.295	78.108	56.647	62.839	**78.829**

Table 1: SUPERT & MAS-Qwen scores across methods, datasets, and topic modeling backbones

topic at once due to the limited context window of LLMs, we instead sample 5 random draws of 10 documents each from the full topic set. Each batch is evaluated independently, and we report the mean score across the five runs. This approach reflects a more realistic human evaluation scenario, where annotators are unlikely to read all the documents in a large collection. Moreover, it aligns with a key assumption in topic-level summarisation, where a strong topic representation should capture the central content of the topic and remain consistent and relevant across different subsets of its documents. Appendix D lists the evaluation prompts.

7 Results

In this section, we present SUPERT scores and Mean Aspect Scores from the LLM-as-a-Judge evaluation on SCS and CCS, using `Qwen-2.5-7B-Instruct` as our primary model and examining the LLM-as-a-Judge's consistency across `Mistral-7B-OpenOrca` and `Gemma-3-27B` model families. We first examine overall pipeline effectiveness, then analyze how topic size affects description quality.

7.1 Performance Across Domains and Backbones

We compare the two topic description pipelines across datasets and topic modeling backbones. The results, shown in Table 1, suggest a consistent pattern favoring SCS. It achieves higher SUPERT and MAS scores in most configurations, indicating potential advantages across domains and backbone models. CCS performs competitively, achieving strong SUPERT scores in several configurations, but tends to score lower on MAS in most settings.

To examine the consistency of the evaluation results across LLMs, we measured the correlation between the MAS scores produced by the three judge models using Kendall's tau-b. The results show strong agreement between `Qwen-2.5-7B-Instruct` and `Gemma-3-27B`, and moderate agreement across the other model pairs, as shown in Table 2.

Judge Models	τ_b	p
Qwen & Gemma	0.7255	$4.304 \cdot 10^{-6}$
Qwen & Mistral	0.5033	$2.99 \cdot 10^{-3}$
Mistral & Gemma	0.5163	$2.24 \cdot 10^{-3}$

Table 2: Kendall's τ_b correlations between MASs of judge models.

Moreover, the MASs show consistent behavior across document draws. For transparency, we include a detailed presentation of MAS on each document set draw in Appendix F.

7.2 Effect of Topic Size

To better understand how topic characteristics impact description quality, we analyze the effect of topic size on MAS distributions for SCS and CCS, cross-validated with topic cohesion (mean cosine distance among topic documents).

Figure 2 shows the distribution of winners among the probed pipelines by topic size quartile. Our findings suggest that SCS tends to perform better among the first, second, and third topic size quartiles. The pattern shifts in the Large category, where CCS matches SCS with an equal number of wins. Additionally, while the number of SCS wins tends to decline as topic size increases, CCS shows an upward trend from Small to Large categories, achieving parity with SCS in the largest quartile.

Cross-validation against topic cohesion suggests that description quality remains relatively consistent across all topic cohesion quartiles for both SCS and CCS, indicating that these approaches may be robust to variation in topical coherence and that the observed size effects above are not confounded by cohesion variations. We provide a thorough presentation against topic cohesion in Appendix G. Appendix H shows SUPERT-based results by topic size & cohesion, showing a similar trend to

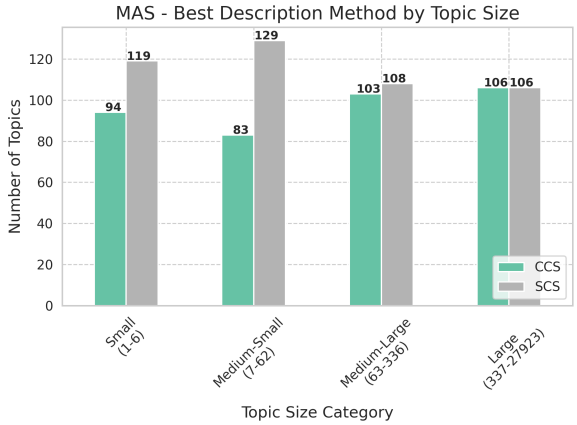


Figure 2: Winner count on LLM-Eval MAS per topic size quartile over all topic models.

MAS. Appendix E reports Kendall's τ_b correlations between LLMs on topic size preferences.

8 Discussion

This section explores both pipeline effectiveness and highlights observed trends by topic size and structure.

8.1 Selective vs. Compressed Approaches to Topic Description

Effectiveness Advantage of Selective Sampling Our results suggest a consistent effectiveness pattern favoring the Selective Context Summarisation (SCS) pipeline across multiple datasets and topic modeling backbones. The MMR selection process in SCS appears to provide a balanced set of relevant and diverse documents, creating focused yet comprehensive input for the LLM. This selective approach seems to reduce noise from peripheral documents while ensuring core topic terms remain prominent throughout the description generation process. In scientific corpora, we hypothesize that this advantage may be amplified, since documents on the same research topic often share similar objects of study, purposes, and methodologies.

Limitations of Hierarchical Compression CCS's hierarchical structure, despite its theoretical capacity to process entire document sets, appears to suffer from what we term "error propagation" and "keyword/term attrition." As abstractions build upward through the tree, inaccuracies at lower levels can amplify in subsequent steps, while important terminology may become diluted during recursive summarisation. These phenomena likely contribute to CCS's generally lower effectiveness across our evaluation metrics. From an efficiency standpoint, SCS demonstrates a superior compute-to-quality ratio, requiring only a single document set pass compared to CCS's multiple rounds of embedding, clustering, and LLM calls. The stability of SCS effectiveness across different topic modeling backbones (CTM, BERTopic, and TopicGPT) suggests its potential robustness as a general-purpose topic description method that can integrate with existing topic modeling workflows regardless of their underlying approach.

8.2 Scalability and Topic Size Effects

Size-Dependent Effectiveness Patterns Analysis of topic size effects reveals an intriguing pattern: while SCS tends to perform better for small to medium-sized topics, CCS becomes competitive and even outperforms SCS for the largest topics (4th quartile), as shown in Figure 2. This finding highlights important scalability considerations for topic description applications. For smaller topics, SCS appears to effectively identify a representative subset that captures the topic's essence. However, as topics grow larger, the fixed selection size (10 documents in our setup) may become limiting. When topics contain hundreds of documents, even carefully selected subsets may miss important sub-themes or variations. CCS shows a potentially valuable property for larger topics: its hierarchical summarisation approach scales with topic size, preserving coverage of diverse sub-themes that fixed-size selection may miss.

Effectiveness Nuances Across Size Deciles The relationship between topic size and method effectiveness shows additional nuance when examined at finer granularity. Figure 3 displays MAS per topic size decile. Notably, SCS appears to demonstrate consistent performance across the initial six deciles. CCS then shows improved performance for the seventh and eighth deciles, before SCS regains dominance for the largest topics. This suggests that while CCS may outperform SCS for some larger topics, it also appears to have a saturation point, likely due to a bottleneck in hierarchical compression of information. This scale-dependent effectiveness suggests that practical applications might benefit from exploring hybrid approaches that adaptively select between methods based on topic size. Our analysis indicates that these patterns persist when controlling for topic cohesion, suggesting that the observed effects may be genuinely related

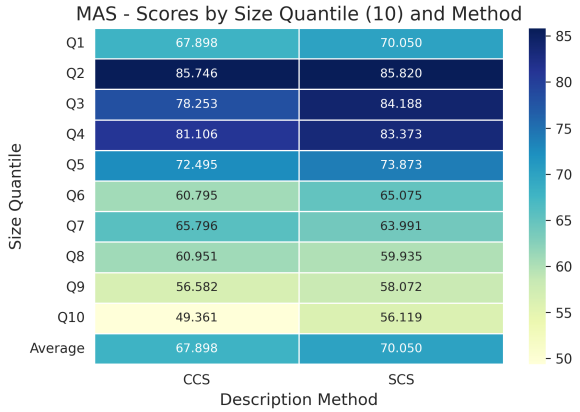


Figure 3: Mean Aspect Score per topic size decile.

to scale. This highlights topic size as a potentially important factor in designing and evaluating topic description pipelines for scientific corpora.

8.3 Qualitative Analysis

To complement our quantitative results, we conducted a targeted qualitative analysis of 45 topic descriptions. We examined 15 top-scoring, 15 low-scoring and 15 descriptions with diverging SUPERT and LLM-as-Judge scores. This enabled us to examine the behaviors of the methods beyond aggregate metrics. Examples illustrating content quality across different models and methods are provided in Appendix I.

Characteristics of Selective Context Summarisation Our analysis suggests that SCS consistently generates clear and coherent summaries in high-scoring cases, with strong alignment to the provided topic terms and good coverage of central concepts (see Example 1). It appears to demonstrate resilience to incoherence in topic terms (Example 2), as any inconsistencies in the topic terms do not compound through multiple summarisation layers. SCS descriptions tend to maintain coherence across different datasets and topic modeling backbones, indicating potential transferability. However, in low-performing cases, particularly when topic terms are overly general or lossy, the method tends to produce generic or shallow outputs. This limitation appears exacerbated when the selected representative documents contain primarily general knowledge rather than specific insights. We observe that SCS capitalize on well-selected topic terms from the underlying model, creating a synergistic effect where strong topic models may yield better descriptions.

Characteristics of Compressed Context Summarisation CCS exhibits distinctive strengths in handling complex or technical topics, often producing more detailed descriptions than SCS. However, this method shows lower alignment with the original topic terms, in several cases generating dense and nuanced content that only partially connects to the provided terms. This misalignment creates challenges in verifying how faithfully the description represents the intended topic (see Example 3). The hierarchical summarisation approach in CCS appears to struggle with effectively prioritizing the most important content, often resulting in information overflow manifested as lengthy lists or excessive detail. This limitation may stem from "document grounding distance" effects in the hierarchical summarisation process, which may not optimally distinguish central from peripheral information. Finally, CCS demonstrates greater sensitivity to topic term quality, with more frequent failures when topic terms are incoherent (see Example 4 comparatively to Example 2).

9 Conclusion & Future Work

We explored Topic Description for Scientific Corpora, an approach to creating structured, document-based summaries that go beyond term lists. To investigate this, we adapted two LLM-based pipelines: Selective Context Summarisation (SCS) and Compressed Context Summarisation (CCS). Our preliminary findings suggest that SCS tends to achieve better performance across datasets and topic modeling backbones, while CCS shows potential advantages for large topics due to its scalable, recursive structure. Our observations highlight an interesting trade-off between selective and compressed strategies: SCS appears to excel in precision and stability, while CCS may offer broader coverage for large-scale topics. The scale-dependent effectiveness patterns we observed suggest that topic size represents an important consideration for practical deployment. Together, they provide initial insights for developing interpretable topic representations in scientific domains.

This work opens several directions for further exploration, including methodological improvements and practical applications. Instead of single-vector retrieval in SCS, future work could examine more fine-grained retrieval strategies to improve coverage and adaptability for complex or broad topics.

Limitations

Despite our multi-faceted evaluation approach, several limitations of the study require further discussion. First, we do not include human assessment. Although we combine SUPERT and LLM-as-a-Judge to approximate quality, expert feedback would be valuable, especially in scientific domains where interpretability and factual accuracy benefit from domain knowledge. The use of both SUPERT and LLM-based evaluation offers complementary strengths: SUPERT captures content relevance via semantic similarity, while LLM-as-a-Judge enables structured, fine-grained evaluation. This dual setup may mitigate some metric-specific biases, though it cannot fully substitute for human judgment.

This challenge is compounded by limitations in the topic modeling stage itself. The quality of topic descriptions is directly tied to the coherence and relevance of the underlying topics and their terms. Despite optimization, CTM often produced noisy or domain-unspecific topics. Similarly, TopicGPT occasionally generated topics that were overly broad or narrowly scoped. These issues affected the resulting descriptions, even with grounded generation. This dependence on topic model quality represents a central limitation in our current investigation. In addition, our study does not include comparisons against simple or established baselines, which would help contextualize the performance of the proposed pipelines.

However, such limitations are inherent to real-world applications (academic and industrial alike) when attempting to gain an overview of large-scale scientific corpora. Our analysis engages with these challenges rather than avoiding them, which we believe is valuable for understanding practical deployment considerations.

Additionally, while the chosen LLMs are among the strongest available models, their outputs remain sensitive to prompt design and can hallucinate content. Our pipelines use a fixed 3A-COT-derived prompting strategy, but prompt wording significantly affects LLM output. No ablation or robustness analysis was conducted to assess this sensitivity in our current work. Moreover, even strong LLMs are prone to hallucination, especially when context is sparse or ambiguous. This is only partially addressed by the factuality criterion in our LLM-as-a-Judge evaluation.

Finally, our evaluation focuses on English scientific abstracts, raising questions about the generalisability of our findings to full-text documents, other genres like patents, and non-English data. The datasets themselves are closely tied to computer science, limiting insights into whether our findings generalise to other disciplines such as the social sciences or biology. These areas require further investigation.

References

Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. 2020. Automatic generation of topic labels. *CoRR*, abs/2006.00127.

Hosein Azarbonyad, Zubair Afzal, and George Tsatsaronis. 2023. Generating topic pages for scientific concepts using scientific publications. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part II*, volume 13981 of *Lecture Notes in Computer Science*, pages 341–349. Springer.

Ibrahim Al Azhar, Venkata Devesh Reddy, Hamed Alhoori, and Akhil Pandey Akella. 2025. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. *CoRR*, abs/2503.10658.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2016. Automatic labelling of topics with neural embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 953–963, Osaka, Japan. The COLING 2016 Organizing Committee.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 335–336. ACM.

Jonathan D. Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British*

Columbia, Canada, pages 288–296. Curran Associates, Inc.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2270–2282. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 81 others. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

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

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 79 others. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

JoongHoon Kim, Sangmin Lee, Seung Hun Han, Saeran Park, Jiyoon Lee, Kiyoon Jeong, and Pilsung Kang. 2023. Which is better? exploring prompting strategy for llm-based metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 164–183.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1536–1545, Portland, Oregon, USA. Association for Computational Linguistics.

Christoph Leiter and Steffen Eger. 2024. Prexme! large scale prompt exploration of open source llms for machine translation and summarization evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11481–11506.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The eval4nlp 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 117–138.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder.

Chau 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.

Cristian Popa and Traian Rebedea. 2021. BART-TL: weakly-supervised topic label generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1418–1425. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. 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*, pages 399–408.

Domenic Rosati. 2022. Moving beyond word lists: towards abstractive topic labels for human-like topics of scientific documents. *CoRR*, abs/2211.05599.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Kartik Shinde, Trinita Roy, and Tirthankar Ghosal. 2022. An extractive-abstractive approach for multi-document summarization of scientific articles for literature review. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 204–209, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021. Word embedding-based topic similarity measures. In *International conference on applications of Natural Language to information systems*, pages 33–45. Springer.

Huy Quoc To, Ming Liu, Guangyan Huang, Hung-Nghiep Tran, Andr'e Greiner-Petter, Felix Beierle, and Akiko Aizawa. 2024. Skt5scisumm – revisiting extractive-generative approach for multi-document scientific summarization. *Preprint*, arXiv:2402.17311.

Pancheng Wang, Shasha Li, Kunyuan Pang, Liangliang He, Dong Li, Jintao Tang, and Ting Wang. 2022. Multi-document scientific summarization from a knowledge graph-centric view. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6222–6233, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024a. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *CoRR*, abs/2412.13663.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024b. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Yongbing Zhang, Shengxiang Gao, Yuxin Huang, Zhengtao Yu, and Kaiwen Tan. 2024. 3a-cot: An attend-arrange-abstract chain-of-thought for multi-document summarization. *International Journal of Machine Learning and Cybernetics*.

A Summarisation Prompts

We used the `deepseek-v3` model to generate topic descriptions across all methods. To ensure consistency and structure in the outputs, we define a fixed system message and adopt a 3-step prompting framework inspired by the 3A-COT method (Zhang et al., 2024). This includes *attending* to key aspects, *arranging* extracted information, and generating the final *abstract*. The prompt templates used are provided below.

System Prompt
You are a scientific research assistant who organizes information into structured markdown documents. Your writing style sounds natural and professional. Avoid using Marketing and HR language.

Prompt 1: System prompt used for topic description generation

Attend Prompt

```
[DOCUMENTS]
```

What are the research purposes in this document?
What are the research object in this document?
What are the research methods in this document?
What are the research result in this document?
What are the main findings in this document?

Please answer the above questions:

Prompt 2: Attend prompt for extracting key information

Arrange Prompt

```
[ATTEND_OUTPUT]
```

Organize the above important information. Arrange this information in a logical order or relevance to build a coherent narrative, and consider how information from different articles can be combined to complement and connect with each other.

Prompt 3: Arrange prompt for structuring extracted content

Abstract Prompt

```
[ABSTRACTS]
—
[ARRANGE_OUTPUT]
```

Based on the above abstracts, key information and the keywords: `{topic_words}`, write a summary.

Make sure to include key information, research objectives and ideas. The summary should be structured as clean MARKDOWN with ONLY the following Headings:

Brief Introduction into the Topic, Key Research Objects, Key Research Methods, Key Research Purpose.

Each Heading should have only Keypoints listed. Avoid the use of additional MARKDOWN subsections. Avoid adding your own opinion, interpretation, or conclusions or Future Work. Use the information provided in the text only.

Prompt 4: Abstract prompt for final topic description generation

B Quantum Dataset Query

In Query 1, we present the full boolean query used for collecting the source documents for the Quantum Computing dataset. Specially, the query is specialized in the hardware part of this scientific field.

C Topic Models

This section presents implementation details and results of three topic modeling approaches used in our comparative analysis.

C.1 Overview

For all approaches involving training, we perform hyper-parameter optimization to find the best coherence and diversity metrics for each combination of topic model and dataset. For coherence, we use the Gensim implementation of the Coherence Model (Řehůřek and Sojka, 2010), specifically its default C_V metric (Röder et al., 2015). For diversity, we calculate the Inverted Rank-Biased Overlap (Webber et al., 2010; Terragni et al., 2021) of the top 10 terms representing the topics.

Table 3 shows that CTM achieves the highest coherence scores, followed by BERTopic and TopicGPT. Conversely, TopicGPT models the greatest number of topics on average, followed by BERTopic and then CTM. Moreover, while CTM achieves the most robust diversity scores on the three datasets, BERTopic and TopicGPT come on par in the ACL and the NIPS datasets, respectively. Finally, manual inspection shows that TopicGPT generally tends to construct more specific topics, with less documents per topic. While CTM has the contrary behaviour presenting broader topics with larger document sets. This observation explains in part the higher diversity scores of CTM. In turn, the low coherence scores of TopicGPT reflects the zero-shot decoupling from the underlying respective corpus.

C.2 Implementation Details

CTM In the CTM backbone, we use the GitHub implementation[6] of the original contribution (Bianchi et al., 2021). Here, we optimize

[6] https://github.com/MilaNLProc/contextualized-topic-models

Boolean Query
TITLE-ABS-KEY ("quantum comput*" OR "quantum processor" OR "quantum circuit" OR "quantum logic gate" OR "quantum gate" OR "logical qubit" OR qubit OR "quantum system" OR "quantum information processing" OR "quantum control" OR "quantum electronics" OR "quantum hardware" OR "noisy intermediate-scale quantum era" OR "NISQ" OR "multiqubit circuit" OR "quantum simulation" OR "quantum simulator") AND TITLE-ABS-KEY ((cryogen* OR "magnetic field" OR laser OR photoluminescence OR silicon OR "electric fields" OR magnetism OR fluorescence) OR ("neutral atom" OR "cold atom" OR "trap*atom" OR "atom trap" OR "rydberg" OR atoms OR "optical lattice*" OR magic OR "optical tweezer*" OR strontium OR ytterbium OR "photonic crystal fibre") OR ("ion traps" OR "trapped ions" OR "ions" OR "integrated waveguide" OR "laser induced deep etching" OR "on-chip coupling") OR (superconduct* OR "SQUIDs" OR "Josephson junction device*" OR "indium bump" OR "NbN films" OR "single flux quantum" OR "quantum flux" OR "SQUID") OR (center OR diamond OR "NV center" OR "NV centre" OR "color centre" OR "colour center" OR "silicon vacancy centre" OR "silicon vacany center") OR (photon* OR "gaussian boson sampl*" OR "squeezed light source" OR niobate OR "superconducting nanowire single-photon detector" OR "SNSPD") OR (topology OR "topological quantum computing" OR "topological insulator*") OR (semiconductor OR "molecular beam epitaxy" OR "semiconducting*" OR "crystal lattice*" OR phonons)) AND PUBYEAR > 2009 AND PUBYEAR < 2026

Query 1: Boolean query used for collecting the source documents for the Quantum Computing dataset.

over four hyper-parameters: number of topics (40—100), number of epochs (10—50), activation function ({sigmoid, relu, softplus}), number of neurons (100—500). All other hyper-parameters use standard values from the implementation.

BERTopic For this approach, we use the standard BERTopic package[7]. This standard pipeline consists of mainly three stages: Embedding (Em.) stage, Dimensionality Reduction (DR) stage, the Clustering (Cl.) stage, and Topic Representation (TR) stage. For the Em stage, we use the nomic-ai/modernbert-embed-base[8] model (Nussbaum et al., 2024), which is an embedding model trained on the ModernBERT (Warner et al., 2024b) encoder. For the DR and Cl. stages, we opt for the standard pairing with UMAP (McInnes et al., 2018) and HDBSCAN (McInnes et al., 2017). Finally, in the TR stage, we use class-TFIDF, which was introduced in (Grootendorst, 2022). Overall, we optimize four hyper-parameters: UMAP - number of neighbors (5—50), number of components (2—15) and min. distance (0.0—0.5); HDBSCAN - min. cluster size (10—50). For UMAP, we fix the metric to cosine, and euclidean for HDSCAN. All other hyper-parameters use standard values from their implementations.

TopicGPT We follow the original TopicGPT pipeline (Pham et al., 2024), using the open-source implementation available at GitHub [9] and altering only the document-assignment stage to align with BERTopic and CTM. For topic generation, we randomly sample 1,000 documents from each dataset and leverage GPT-4 to propose an initial set of top-level topics, which we then iteratively refine into subtopics to build a complete hierarchical structure. In the subsequent assignment phase—applied to the full datasets—we replace TopicGPT's default routine (which, for each document, prompts GPT-3.5-turbo with the finalized hierarchy and returns the best-matching topic with a supporting quote) with a two-part prompt to GPT-3.5-turbo: (i) assign each document to its best-matching topic in our hierarchy; and (ii) extract ten representative keywords per document. Finally, we post-process all extracted keywords for each topic by tokenizing them on whitespace, converting to lowercase, stripping punctuation, aggregating token frequencies, and selecting the ten most frequent tokens per topic—thereby exactly matching the output format of our BERTopic and CTM backbones.

C.3 Results

Table 3 presents topic modeling evaluation results across three datasets (ACL, NIPS, and Quantum) for three different topic modeling approaches: CTM, BERTopic, and TopicGPT. The evaluation metrics used in the comparison are Coherence, Di-

[7]https://maartengr.github.io/BERTopic/index.html

[8]https://huggingface.co/nomic-ai/modernbert-embed-base

[9]https://github.com/chtmp223/topicGPT

Document Assignment Prompt Template

You will receive a document and a topic hierarchy. Assign the document to the most relevant topic of the hierarchy. Then, output the topic label, and supporting keywords from the document. DO NOT make up new topics or keywords.

[Topic Hierarchy]
`{tree}`

[Instructions]
1. Topic label must be present in the provided topic hierarchy. You MUST NOT make up new topics.
2. The keywords must be taken from the document. You MUST NOT make up keywords or quotes. All keywords MUST NOT contain stop words.

[Document]
`{Document}`

Double check that your assignment exists in the hierarchy! Your response should be in the following format:
`[Topic Level] Topic Label: keyword1, keyword2, etc`

Your response:

Prompt 5: Prompt template used for document-to-topic assignment in the TopicGPT adaptation.

versity, and Number of Topics (N.Topics). Our findings show that CTM achieves the highest coherence scores across all three datasets (0.664 for ACL, 0.601 for NIPS, and 0.692 for Quantum). It also maintains high diversity scores above 0.94 for all datasets. BERTopic shows moderate coherence performance (0.504 for ACL, 0.458 for NIPS, and 0.546 for Quantum), with somewhat lower diversity metrics, particularly for the Quantum dataset (0.799). TopicGPT demonstrates coherence scores between 0.458 and 0.526 across datasets, with strong diversity in the NIPS dataset (0.963) but lower diversity for ACL (0.881) and Quantum (0.809). Regarding the number of topics identified, TopicGPT produces substantially more topics than the other approaches, particularly for NIPS (276). BERTopic identifies the fewest topics overall with just 24 for the NIPS dataset. For the ACL dataset, the number of topics is more consistent across models (CTM: 72, BERTopic: 70, TopicGPT: 66). The Quantum dataset shows moderate variation, with CTM identifying 59 topics, BERTopic 72, and TopicGPT significantly more at 169.

D Evaluation Prompts

We evaluate summaries along four dimensions: relevance, coherence, factuality, and fluency. Each is scored independently using a dedicated prompt, detailed below.

Aspect Definitions

- **Relevance:** The rating measures how well the summary captures the key points of the documents. Consider whether all and only the important aspects are contained in the summary.
- **Coherence:** This rating evaluates how seamlessly the sentences of the summary flow together, creating a unified whole. Assess how smoothly the content transitions from one point to the next, ensuring it reads as a cohesive unit.
- **Factuality:** This rating gauges the accuracy and truthfulness of the information presented in the summary compared to the original documents. Scrutinize the summary to ensure it presents facts without distortion or misrepresentation, staying true to the source content's details and intent.
- **Fluency:** This rating evaluates the clarity and grammatical integrity of each sentence in the summary. Examine each sentence for its structural soundness and linguistic clarity.

E Gemma-3-27B & Mistral-7B-OpenOrca Results

To complement the main results, we report the MAS obtained using Gemma-3-27B and Mistral-7B-OpenOrca in Table 4. These models provide additional perspectives on the quality of the generated descriptions and help examine the consistency of trends observed with Qwen-2.5-7B-Instruct.

To further examine inter-model agreement, we compute Kendall's τ_b between the rankings of

TM	Dataset	Coherence	Diversity	N.Topics
CTM	ACL	0.664	0.994	72
	NIPS	0.601	0.949	38
	Quantum	0.692	0.996	59
BERTopic	ACL	0.504	0.972	70
	NIPS	0.458	0.930	24
	Quantum	0.546	0.799	72
TopicGPT	ACL	0.458	0.881	66
	NIPS	0.472	0.963	276
	Quantum	0.526	0.809	169

Table 3: Topic modeling evaluation results across three scientific datasets.

Method	**Metric**	**ACL**			**NIPS**			**Quantum**		
		CTM	BERTopic	TopicGPT	CTM	BERTopic	TopicGPT	CTM	BERTopic	TopicGPT
SCS	MAS-Mistral	**74.913**	**74.306**	**76.307**	**75.497**	**75.109**	**77.240**	**72.077**	**73.127**	**76.120**
	MAS-Gemma	**81.537**	**85.163**	**87.411**	**82.691**	**83.865**	**89.400**	**82.006**	**85.255**	**89.395**
CCS	MAS-Mistral	73.092	71.207	74.471	66.384	70.771	75.201	65.031	72.758	74.528
	MAS-Gemma	81.483	85.155	86.799	81.194	83.370	88.549	80.204	85.120	89.125

Table 4: MAS scores across methods, datasets, and topic modeling backbones using Mistral and Gemma as judge models.

Evaluation Prompt Template

Instruction:
In this task you will evaluate the quality of a summary written for multiple documents.
To correctly solve this task, follow these steps:
1. Carefully read the document, be aware of the information it contains.
2. Read the proposed summary.
3. Rate each summary on a scale from 0 (worst) to 100 (best) by its `{aspect}`. Decimals are allowed.

 # Definition:
`{definition}`

Source documents:
`{source}`

Summary:
`{summary}`

Score:

Prompt 6: Evaluation prompt template used for scoring topic descriptions across relevance, factuality, coherence, and fluency

method–size combinations (i.e., CCS/SCS across the four topic size categories: *Small*, *Medium-Small*, *Medium-Large*, and *Large*) for each pair of judge models. We evaluate agreement across the full 8-item ranking. This provides a single τ_b score per pair, reflecting overall alignment in method preferences across topic sizes. As shown in Table 5, `Qwen2.5-7B-Instruct` shows moderate to strong alignment with both `Gemma-3-27B` and `Mistral-7B-OpenOrca`, while `Gemma-3-27B` and `Mistral-7B-OpenOrca` exhibit weaker agreement.

Model Pair	τ_b	p
Qwen & Gemma	0.6183	0.0340
Qwen & Mistral	0.6910	0.0178
Gemma & Mistral	0.2857	0.3988

Table 5: Kendall's τ_b between full method–size rankings of each model pair.

F Impact of Document Sampling in LLM-Eval

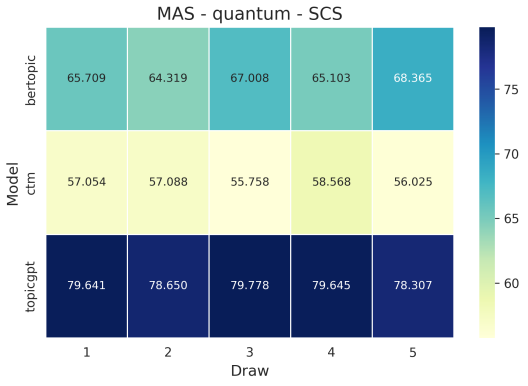
In order to analyze the impact of using subsets of documents of topics as reference documents in the LLM-Eval strategies, we present a detailed visual-

ization of the Quantum dataset results in Figure 4 across all five document draws for each TM and TD approach. Similar patterns are observed for the other datasets.

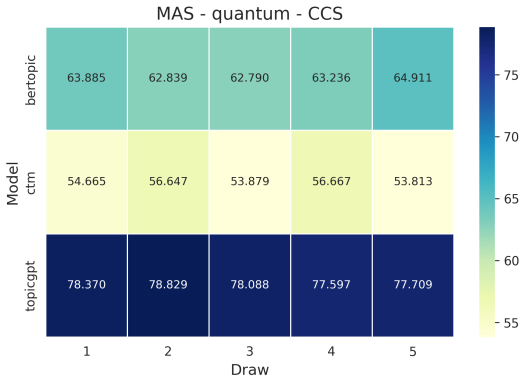
From visual inspection of Figure 4, we observe that scores remain relatively stable across different document draws for the same TM and TD method. When fixing a topic modeling approach and a topic description pipeline, the fluctuations in LLM-Eval MAS are generally small, with most variations remaining within 5 points to the mean on our 100-point scale.

While a comprehensive variance analysis across all datasets would provide further statistical rigor, the consistency observed in the Quantum dataset suggests that our sampling approach may produce reliable evaluations. The observed stability indicates that randomly sampling 10 documents five times appears to provide a reasonable approximation of how a topic description would be evaluated against the full document collection.

The observed consistency across document draws supports our decision to use this sampling approach as a practical solution to the context window limitations of LLMs. While a more exhaustive analysis would be valuable for future work, the current evidence suggests that our methodology may provide reliable evaluations of topic descriptions despite using only subsets of the complete document collections.



(a) SCS



(b) CCS

Figure 4: LLM-Eval MAS for every draw of 10 documents per topic.

G Effect of Topic Cohesion on Mean Aspect Score (MAS)

To study the impact of topic cohesion on the quality of topic descriptions, We compute the mean cosine embedding distance among all documents for each topic. We call this indicator "Topic Cohesion." Figure 5 shows the MAS distributions for all topics grouped by their topic cohesion quartile. Interestingly, topic cohesion appears to play a relatively minor role in the MAS distributions across all quartiles. There is a slight downward trend indicating some anticipated TD quality degradation towards topics of low cohesion. However, this effect appears minor among all TD approaches, only becoming more pronounced in the low cohesion quartile. Even there, the best topic descriptions of SCS and CCS are competitive with TD's best scores in the more cohesive quartiles.

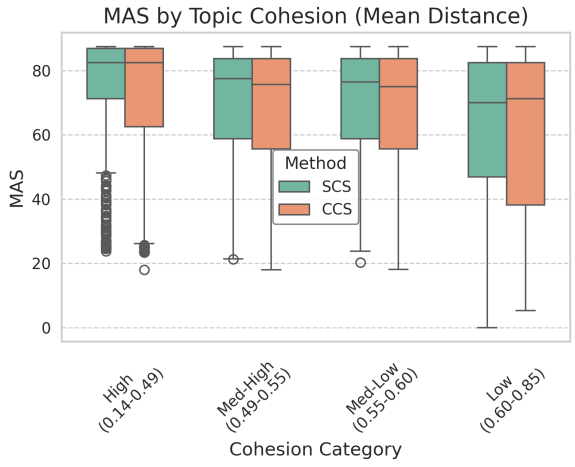


Figure 5: Box-plots of MAS distributions from Qwen conditioned on topic cohesion quartile.

H Effect of Topic Size and Cohesion on SUPERT

Figure 6 shows the distribution of winners per topic size category based on the SUPERT metric. SCS leads in the first and second quartiles, with CCS gaining a slight edge in the third. In contrast to MAS-Qwen, which shows CCS catching up in the largest category, SUPERT continues to favor SCS in the fourth quartile. This suggests that SCS may be more aligned with SUPERT's relevance-focused evaluation, even as topic size increases.

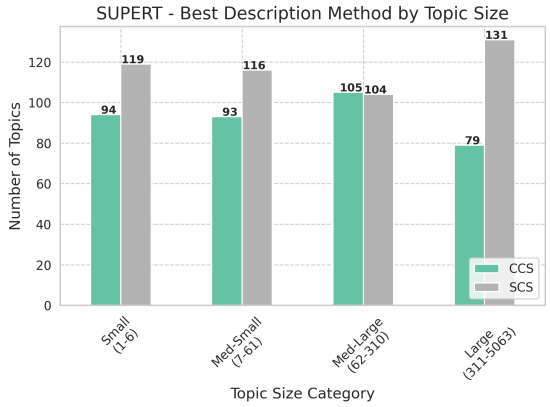


Figure 6: Winner count on SUPERT per topic size quartile over all topic models

Figure 7 shows SUPERT distributions grouped by topic cohesion (mean pairwise cosine distance). Both SCS and CCS appear relatively stable across cohesion levels, with only slight performance degradation from high to low cohesion.

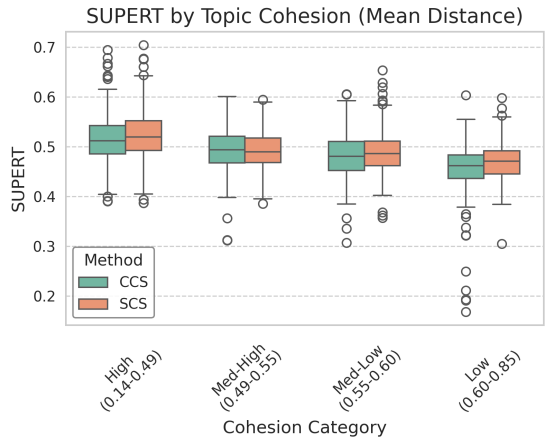


Figure 7: Box-plots of SUPERT distributions conditioned on topic cohesion quartile

The SUPERT experiments on effectiveness nuances across size deciles, to be seen in Figure 8 reveal patterns consistent with the results of LLM-Eval in Section 8.2: method effectiveness appears tied to topic size, with SCS showing advantages for smaller topics (first to sixth decile). The effect of CCS slightly outperforming SCS in LLM-Eval for the larger topics (seventh and eigth deciles) is not to be seen with SCS even achieving minially better scores with SUPERT. Still, the overall trend is replicated from LLM-Eval and this consistency further suggests that scale, besides intrinsic topic quality, drives performance differences and supports the potential viability of adaptive hybrid approaches.

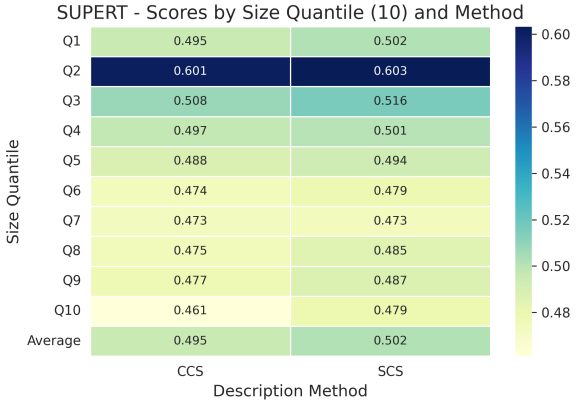


Figure 8: SUPERT Score per topic size decile.

I Examples of generated Topic Descriptions

To demonstrate the results of our topic description pipeline, we present a selection of examples covering different datasets and modelling approaches. Each example provides metadata, the generated summary, and evaluation scores. Together, these elements help contextualize the content and quality of the generated descriptions.

acl-53

Dataset: ACL
Model: TopicGPT
Method: SCS
Category: Best
Topic Words: data, LOD, digital, linguistics, studies, linked open, humanities, multilingual, modeling
SUPERT: 0.7043
LLMEval: 87.5000

Topic Summary: Linked Open Data (LOD) in the Humanities
Brief Introduction into the Topic
Linked Open Data (LOD) is increasingly being integrated into the Humanities to enhance the accessibility, interoperability, and analysis of machine-readable information. Two notable initiatives, LODinG and DigItAnt, exemplify this trend. LODinG explores broad applications across various Humanities disciplines, while DigItAnt focuses on creating, linking, and exploiting LOD lexica for historical linguistics and epigraphy data.

Key Research Objects
- LODinG: Covers digital humanities, computational linguistics, digital lexicography, philology, translatology, cultural and literary studies, media studies, information science, and law.
- DigItAnt: Manages LOD lexica and heterogeneous resources in historical linguistics and epigraphy, supporting the H2IOSC Italian infrastructures cluster project.

Key Research Methods
- LODinG: Collects, models, links, releases, and analyzes multilingual data in LOD format, developing a modular, cross-domain Humanities data model.
- DigItAnt: Designs architecture for LOD lexica and links Linguistic Linked Open Data (LLOD) with non-LLOD datasets, following FAIR principles.

Key Research Purpose
- LODinG: Improves extraction and modeling of multilingual LOD data, demonstrates LLOD applications in interdisciplinary contexts, and develops a cross-domain data model.
- DigItAnt: Aims to build and integrate LOD lexica for historical linguistics and epigraphy, support the H2IOSC project, and contribute to Italian Humanities infrastructure.

Example 1: Example topic description for topic acl-53.

quantum-49

Dataset: Quantum
Model: CTM
Method: SCS
Category: Ambiguous
Topic Words: foremost, boast, ineffective, craft, exacerbate, maintains, complication, pack, prevail, upon
SUPERT: 0.4825
LLMEval: 62.5

Summary of Quantum Computing and Related Technologies
Brief Introduction into the Topic
Quantum computing and related technologies focus on advancing fault-tolerant quantum machines, enhancing quantum communication, and exploring novel materials and methods for quantum applications. Research spans from error correction in quantum computing to the development of high-performance quantum platforms, quantum networking, and advanced optical materials.

Key Research Objects
Fault-tolerant quantum machines: Large-scale quantum computers relying on error correction using logical and physical qubits.
Quantum emitters and waveguides: Dipole quantum emitters embedded in dielectric slab waveguides.
Quantum platforms: Diamond with engineered coherent defects for quantum control.
Quantum simulations: SU(3) gauge theories in Quantum Chromodynamics (QCD) using discrete subgroup S(1080).
Quantum networking: Quantum wrapper networking with classical headers and quantum payloads.
Quantum box arrays: Self-assembled porous (metal-)organic networks modified by C60 adsorbates.
Quantum optics: High refractive index polymers for infrared (IR) optics.
Quantum communication: DWDM transmission using ultra-low loss G.654.D fiber and lumped doped fiber amplifiers.
Quantum spin manipulations: Ce³⁺ doped LiYF<inf>4</inf> crystals for ENDOR spectroscopy.
Quantum geometry: Desargues and Pappus properties in projective geometry applied to quantum systems.

Key Research Methods
Error correction: Spreading the value of a single logical qubit over multiple physical qubits.
Waveguide analysis: Studying the effective quantum yield of dipole emitters in dielectric slab waveguides.
Materials engineering: Controlling coherent defects in diamond for quantum applications.
Digitization schemes: Practical digitization of SU(3) gauge theories using discrete subgroups.
Noise characterization: Analyzing in-fiber scattering processes in quantum wrapper networking.
Adsorbate modification: Using C60 to modify electronic states in quantum box arrays.
Inverse vulcanization: Synthesizing high refractive index polymers from waste materials.
DWDM transmission: Demonstrating high-speed data transmission using advanced fiber and amplifiers.
ENDOR spectroscopy: Resolving Mims ENDOR spectra for electron-nuclear interactions.
Geometric analysis: Applying Desargues and Pappus properties to quantum systems.

Key Research Purpose To **craft** large, fault-tolerant quantum machines capable of robust computations.
To **boast** the performance of quantum emitters and waveguides by optimizing their properties.
To **maintain** long spin coherence times in diamond-based quantum platforms.
To **pack** the infinite-dimensional gluon field into a minimal qubit representation for QCD simulations.
To **exacerbate** the understanding of noise impacts in quantum networking.
To **prevail** in engineering quantum states in on-surface network architectures.
To **foremost** develop low-cost, high-performance materials for IR optics.
To **complicate** the understanding of high-speed data transmission in optical communication.
To **upon** investigate electron-nuclear interactions for spin manipulations.
To **ineffective** link projective geometry principles to quantum physics and information.

Example 2: Example SCS topic description for topic quantum-49.

acl-20

Dataset: ACL
Model: BERTopic
Method: CCS
Category: Ambiguous
Topic Words: word2vec, embeddings, vocabulary, wordnet, semantic, semantics, embedding, contextualized, word, linguistic
SUPERT: 0.4274
LLMEval: 85.4250

Brief Introduction into the Topic
The extraction and analysis of health-related information from social media platforms, particularly **Twitter**, have become critical in **pharmacovigilance**, **public health monitoring**, and **mental health studies**. Advances in **Natural Language Processing (NLP)** and **transformer-based models** (e.g., BERT, RoBERTa, BERTweet) have improved the ability to process informal and noisy text, enabling the identification of **Adverse Drug Events (ADEs)**, **disease mentions**, and **psychological conditions** such as **depression**, **anxiety**, and **mood disorders**. Social media data, especially **tweets**, provide valuable insights into public health trends and individual health experiences.

Key Research Objects
- **Psychological and Mood-Related Terms**: Keywords such as **depression**, **depressed**, **mood**, **psychological**, and **anxiety**, relevant to ADE classification and diagnosis.
- **Tweets**: Informal text from **Twitter** containing colloquial language related to health, drug effects, and psychological conditions.
- **Disease Mentions**: Identification of diseases such as **depression** and **anxiety** in **Spanish tweets** and other social media content.
- **Adverse Drug Events (ADEs)**: Mentions of adverse effects from medications and dietary supplements in tweets.
- **Social Media Platforms**: **Twitter**, **Reddit**, and **health forums** as primary sources of health-related data.

Key Research Methods
1. **Transformer-Based Models**
- **BERT**, **RoBERTa**, **BERTweet**, and **CT-BERT** are fine-tuned for tasks such as **classification**, **Named Entity Recognition (NER)**, and **information extraction**.
- **Ensemble Methods**: Combining multiple BERT variants to improve accuracy in classifying health-related tweets.
2. **Large Language Models (LLMs)**
- **Data Augmentation**: Generating synthetic data to address data imbalance in ADE extraction and disease mention detection.
- **Retrieval-Augmented Generation (RAG)**: Mapping informal tweet language to standardized terms like **MedDRA Preferred Terms**.
3. **Advanced NLP Techniques**
- **Transfer Learning**: Fine-tuning multilingual models like **mBERT** for disease mention extraction in **Spanish tweets**.
- **Knowledge Graph Integration**: Enhancing models with external evidence from medical gazetteers and ontologies.
- **Flair-NER Framework**: Evaluating embeddings for capturing complex disease mentions in social media text.
4. **Preprocessing and Optimization**
- **Domain-Specific Preprocessing**: Cleaning and preparing social media text for analysis.
- **Hyperparameter Optimization**: Enhancing model performance through tuning.

Key Research Purpose
- **Enhance ADE Discovery**: Improve the accuracy and robustness of ADE extraction and normalization from social media data.
- **Improve Disease Mention Extraction**: Develop advanced NER systems for identifying diseases in informal text, particularly in non-English languages like **Spanish**.
- **Standardize Health-Related Terms**: Map colloquial language to standardized medical terms (e.g., **MedDRA Preferred Terms**) for better pharmacovigilance and health monitoring.
- **Promote Open Research**: Provide open-source code and methodologies to foster collaboration and further innovation in the field.
- **Monitor Public Health Trends**: Leverage social media data for real-time monitoring of public health issues, including **COVID-19** and **psychological conditions**.

Example 3: Example topic description for topic `acl-20`.

quantum-49

Dataset: Quantum
Model: CTM
Method: CCS
Category: Worst
Topic Words: foremost, boast, ineffective, craft, exacerbate, maintains, complication, pack, prevail, upon
SUPERT: 0.1685
LLMEval: 7.1150

Brief Introduction into the Topic
The topic revolves around the analysis of certain systems or strategies that are deemed **foremost** in their field but may **boast** features that are ultimately **ineffective**. These systems often **craft** solutions that inadvertently **exacerbate** existing issues, leading to further **complication**. Despite these challenges, the systems **maintain** their prominence, as they are designed to **pack** significant functionality. The research aims to explore why such systems **prevail** and the factors that drive their continued use **upon** further scrutiny.

Key Research Objects
The primary objects of research include systems or strategies that are considered leading in their domain. These objects are characterized by their advanced features, which are often highlighted as strengths but may contribute to inefficiencies or unintended consequences. The study focuses on understanding the dynamics of these systems and their impact on the broader context in which they operate.

Key Research Methods
The research employs a combination of qualitative and quantitative methods to evaluate the effectiveness of the systems under study. This includes analyzing case studies, conducting surveys, and performing comparative assessments to identify patterns and outcomes. The methods aim to uncover the reasons behind the systems' continued use despite their potential drawbacks.

Key Research Purpose
The purpose of the research is to critically examine the systems that are widely regarded as top-tier in their field. It seeks to identify the factors that contribute to their perceived success, as well as the unintended consequences that may arise from their implementation. The study aims to provide a comprehensive understanding of why these systems **prevail** and how they impact their respective domains.

Example 4: Example CCS topic description for topic quantum-49.