

# An Answer is just the Start: Related Insight Generation for Open-Ended Document-Grounded QA

Saransh Sharma<sup>1</sup>, Pritika Ramu<sup>2‡</sup>, Aparna Garimella<sup>1</sup>, Koyel Mukherjee<sup>1</sup>

<sup>1</sup>Adobe Research, India

<sup>2</sup>University of Maryland, College Park

{sarsharma, garimell, komukher}@adobe.com, pramu@umd.edu

## Abstract

Answering open-ended questions remains challenging for AI systems because it requires synthesis, judgment, and exploration beyond factual retrieval, and users often refine answers through multiple iterations rather than accepting a single response. Existing QA benchmarks do not explicitly support this refinement process. To address this gap, we introduce a new task, *document-grounded related insight generation*, where the goal is to generate additional insights from a document collection that help improve, extend, or rethink an initial answer to an open-ended question, ultimately supporting richer user interaction and a better overall question answering experience. We curate and release **SCOpE-QA** (Scientific Collections for **Open-Ended QA**), a dataset of 3,000 open-ended questions across 20 research collections. We present **INSIGHTGEN**, a two-stage approach that first constructs a thematic representation of the document collection using clustering, and then selects related context based on neighborhood selection from the thematic graph to generate diverse and relevant insights using LLMs. Extensive evaluation on 3,000 questions using two generation models and two evaluation settings shows that **INSIGHTGEN** consistently produces useful, relevant, and actionable insights, establishing a strong baseline for this new task.

## 1 Introduction

Recent advances in generative language modeling (OpenAI, 2023; DeepMind, 2023; Bai et al., 2022; Touvron et al., 2023) have led to a variety of AI-powered document interaction tools, including NotebookLM,<sup>1</sup> Acrobat’s AI Assistant,<sup>2</sup> and ChatPDF.<sup>3</sup> These applications allow users to interact

<sup>‡</sup>Work done while at Adobe Research

<sup>1</sup><https://notebooklm.google>

<sup>2</sup><https://www.adobe.com/acrobat/generative-ai-pdf.html>

<sup>3</sup><https://www.chatpdf.com>

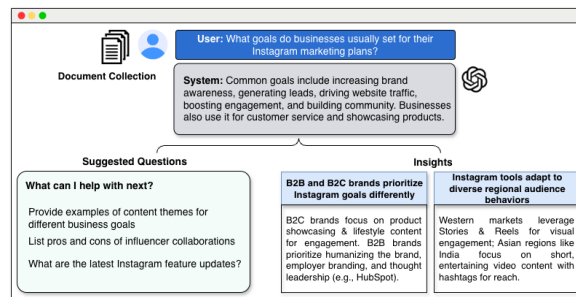


Figure 1: Diagram showing the key difference between insights and traditional follow up or suggested questions. Follow up questions aim to guide the next turn of the conversation, while insights are designed to support the current turn, especially for open ended questions.

with documents through question answering or to generate new content grounded in the documents. Document-grounded question answering (QA) is largely powered by retrieval-augmented generation (RAG) techniques (Sharma, 2025), which enhance LLM performance in diverse scenarios, including single-hop and multi-hop questions (Wu et al., 2025; Li et al., 2025; Zhang et al., 2025), and multimodal questions (Mei et al., 2025; Abootorabi et al., 2025).

Most existing document-grounded QA efforts have primarily focused on answering fact-based questions (e.g., *Who is the father of the current president of the United States?*), where there exists a single, objectively correct answer (Zhang et al., 2025; Wu et al., 2025; Zhang et al., 2024). These systems typically rely on retrieving relevant passages, either in a single-hop or multi-hop manner, and generating concise, factual responses. However, in more real-world scenarios, users interact with their documents with more open-ended queries for various purposes, such as creating reports, formulating strategies, learning, etc. In such settings, it is not straightforward to pinpoint on one “good” answer, as the task demands longer answer statements, more nuanced reasoning processes, and diverse expressions. For example, for a question

such as *Create a draft for a 3-page paper on real world applications of Piaget’s theory*, there can be multiple appropriate responses, with different nuances, perspectives, and considerations. Similarly, for a question *Help me create a 2-page study notes to prepare for an exam*, there can be multiple good responses, say one with a summarized version of the textbook, one with the key insights and common misperceptions, one with a sample quiz and responses, and so on.

Further, in such settings, a single-shot answer generation framework that generates one final answer, like how most current QA systems are designed, rarely suffices, as users may not be satisfied with the first draft. Instead, a framework that enables users to iterate over the generated response by providing AI assistance to brainstorm ideas, refine the content, provide alternate perspectives, or additional data, can more smoothly lead the users to their preferred final version. Prior works to increase this kind of user engagement largely focused on suggesting follow-up questions or next steps (Lin et al., 2025; Huang et al., 2023; Wang et al., 2019). While these offer potential next steps, they also function with the underlying assumption that the generated answer is satisfactory to the user (Figure 1). In this work, we call out the need for open-ended question answering systems to be iterative in nature, and propose that AI assistance should be provided in brainstorming ideas or providing other perspectives to think along for the users, as opposed to being solely focused on improving the quality of the first version of the answers.

In this paper, we introduce the task of *document-grounded related insight recommendation* for open-ended document-grounded question answering, aimed at supporting iterative answer refinement. Given an open-ended question, an initial answer, and a collection of supporting documents, the model is required to generate additional insights that can help improve the existing answer. The generated insights must satisfy three conditions. First, they should be grounded in the provided documents. Second, they should be related to the question while avoiding repetition of information already present in the answer. Third, they should offer useful directions for refining or extending the answer. To get complementary context beyond the current answer, we reimagine K-Means to work on document chunks and build a graph representation of the collection. We then traverse this graph to find information that is related but not repetitive in the

current answer. Using a chain-of-thought reasoning prompt (Wei et al., 2022), this context is processed to generate novel and relevant insights. The graph structure guides the retrieval, enabling a shift from standard similarity to structural complementarity for open-ended insight generation.

This paper makes three main contributions. **(1)** We introduce a new task of document-grounded related insight recommendation for open-ended document QA, and release a new dataset, **SCOPE-QA** (Scientific Collections for Open-Ended QA), consisting of 3,000 open-ended questions spanning 20 research themes. **(2)** We propose a two-stage framework that first retrieves complementary context using a clustering-based neighborhood selection strategy, and then generates insights using a CoT-style reasoning module. **(3)** We compare our approach with naïve prompting and existing RAG-based context selection methods, and show that our method produces insights that are more novel, relevant, and informative.

## 2 Related Work and Background

**Related Question and Document Understanding:** Prior work most closely related to our task lies in related question generation and document understanding, which are typically applied after an initial answer to guide subsequent interaction. **Conversational Question Generation (CQG)** focuses on producing follow-up questions in multi-turn settings. Early approaches often ignored previous answers and yielded generic questions (Nakanishi et al., 2019; Do et al., 2023), while later methods improved coherence through answer awareness, reinforcement learning, and discourse modeling (Pan et al., 2019; Gu et al., 2021), as well as the use of external knowledge and synthetic data (Hwang and Lee, 2022; Faille et al., 2024). Similarly, **Suggested Question Generation (SQG)** produces related or clarifying questions, evolving from retrieval and rewriting approaches (Pan and Shaw, 2007) to retrieval-augmented generation with few-shot prompting and personalization (Tayal and Tyagi, 2024; Lin et al., 2025). Parallel work on document understanding extracts key information through summarization (Tan et al., 2017) or structured analysis of research papers (Song et al., 2025), although such methods are often constrained by predefined schemas and formats.

While these approaches operate in a similar post-answer setting, they differ fundamentally in objec-

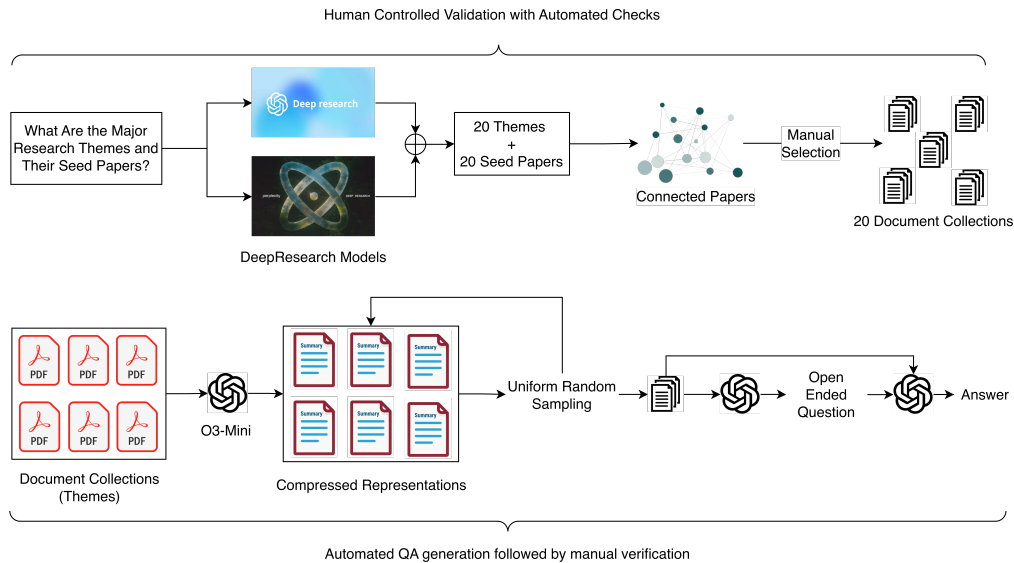


Figure 2: SCOpE-QA dataset curation pipeline overview, highlighting collection curation and QA generation.

tive. Related questions (CQG and SQG) aim to move the conversation forward by identifying what information should be sought next. For example, consider the question: “**How can a city redesign its public libraries to support digital-era creativity?**” A base answer might suggest maker spaces, digital workshops, startup partnerships, and collaborative work areas. A CQG or SQG system would generate follow-up questions such as “**What funding models could support these redesigns?**” or “**How have other cities done this?**”, which extend the discussion by requesting additional information (Figure 1).

In contrast, insight generation focuses on generating *insights* that analyze and improve the current answer rather than extend it. Insights are statements, not questions, and are designed to surface new perspectives, potential issues, or alternative directions. For the same example, an insight might highlight a concern (“**Some community members may not like turning libraries into commercial spaces**”), a trade-off (“**Spending more on digital tools may reduce support for traditional services**”), or a new idea (“**Libraries could host residency programs for digital artists**”). These statements do not request additional input, but instead help rethink, question, or strengthen the existing answer by identifying missing aspects, risks, or opportunities. In this sense, follow-up questions move the conversation forward, while insights deepen the analysis of what has already been produced.

Our formulation is also distinct from other related tasks such as multi-hop question answering, answer expansion, and multi-document summa-

rization. Multi-hop QA focuses on combining evidence to arrive at an answer, answer expansion adds further details to the same response, and multi-document summarization aggregates and condenses information from multiple sources. In contrast, insight generation operates on the answer itself, examining its assumptions, limitations, and implications. By highlighting trade-offs, surfacing potential issues, and suggesting new directions, insights improve answer quality by helping users reflect on and refine their responses, leading to better outcomes in open-ended tasks such as creative problem solving, critical review, and planning.

### 3 SCOpE-QA: Scientific Collections for Open-Ended QA

Document-grounded insight generation involves three inputs: a collection of related documents, an open-ended question, and its answer. The goal is to produce additional insights grounded in the documents, enabling users to better understand, interact with, or refine the answer. This naturally requires reasoning across multiple documents, reflecting real-world information synthesis. Existing multi-document QA datasets (Zhu et al., 2024; Bai et al., 2024; Minzheng Wang, 2024) focus on factual retrieval or structured questions and do not support open-ended cross-document reasoning. To address this, we introduce the **Scientific Collections for Open-Ended QA (SCOpE-QA)** dataset, a publicly released benchmark for multi-document, open-ended reasoning. It contains 20 research collections and 3000 QA pairs, each requiring integration of information across multiple documents.

The o3-mini model is used only for dataset construction (document compression and QA generation) and not for baselines or evaluation, making SCoPE-QA a reliable benchmark for studying cross-document insight generation.

**Collection creation and cleaning:** SCoPE-QA comprises 20 research paper collections across diverse scientific topics. Seed papers and major research themes are identified using ChatGPT<sup>4</sup> and Perplexity Deep Research<sup>5</sup>, which provide representative papers for each theme. Collections are expanded using Connected Papers<sup>6</sup>, generating similarity graphs based on co-citations and bibliographic coupling. Papers with high reference overlap are considered related, then filtered automatically via title relevance and abstract similarity, removing non-English, incomplete, or inaccessible documents. Manual review ensures topical relevance based on abstracts and introductions. This process yields 405 documents across 20 collections, each containing 10-35 papers, allowing analysis of context size effects while maintaining high-quality, theme-specific document sets.

**Document compression:** Each document is compressed with o3-mini to retain key information (main topics, conclusions, contributions) while removing non-essential content. This allows multi-document integration when generating QA pairs.

**Question Generation:** Questions are generated over combinations of compressed papers to encourage cross-document reasoning. For collections with up to 15 papers, 5-paper combinations are sampled, while larger collections use 10-paper combinations. Ten candidate questions are generated per combination, with 50 combinations sampled per collection, yielding about 500 candidates. These are filtered and refined using the o3-mini model to produce 200 high-quality questions per collection (250 for the Quantization).

**Answer Generation and Verification:** Answers are generated using the same sampled combinations of compressed documents as in question generation, with the model instructed to cite relevant sources. Redundant answers are removed using pairwise cosine similarity with a 0.9 threshold. The remaining QA pairs undergo human verification against the full documents to ensure factual accu-

racy. After de-duplication and verification, 150 QA pairs are retained per collection (200 for Quantization). For Quantization, these are further split into 150 test and 50 validation pairs.

The final dataset contains 20 collections and 3000 QA pairs. Appendix A describes the full dataset curation process, and Table 1 summarizes all relevant dataset statistics. Figure 2 summarizes the dataset construction and QA generation pipeline. Apart from the 20 research collections from SCoPE-QA, we also evaluate our pipeline on 15 internal, non-public document collections. These include financial reports, consulting notes, legal documents, and sales materials. Drawn from real organizations, these collections contain real-world questions such as drafting a sales strategy or understanding the business model of a company. We use the same pipeline configuration, prompts, and hyperparameters as in the scientific setting, without any domain-specific changes, so performance on these internal collections reflects the robustness and generalizability of our method.

## 4 Problem Formulation and Methodology

We aim to generate *related insights* as short, reflective prompts that enrich a given QA pair and a document collection. These related insights aim to provide additional perspectives, highlight missing points, and suggest alternative ideas that complement the current answer, helping users think critically and iteratively refine responses. We find two main challenges with using existing RAG-based retrieval and naïve prompting strategies to obtain this kind of *related context*. First, selecting the right context is difficult: the context should be related to the answer and provide complementary information, but most methods focus only on relevance to the question and ignore adjacency or neighborhood information. As we will show in Section 6, simple relevance-based retrieval strategies underperform in obtaining insightful reflections. Second, even with suitable context, the generated insights must be *insightful* (deepening the answer), *novel* (providing new information not already in the answer), and *relevant* (focused on the question).

We hypothesize that in order to generate related insights, we should select context that is related to the given QA pair and not repetitive. For this, we explicitly represent the document collections using a graph-like representation where the various document chunks are embedded and clustered in

<sup>4</sup><https://openai.com/index/introducing-deep-research/>

<sup>5</sup><https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research>

<sup>6</sup><https://www.connectedpapers.com/>

Research Theme	# Files	Avg. Tokens per File	Avg. Words per File	Avg. Question Length (words)	Avg. Answer Length (words)	Avg. Question Length (tokens)	Avg. Answer Length (tokens)
Inference Optimization	10	22,090.0	10,741.1	22.3	675.0	26.9	922.8
LLM Agents	10	23,611.9	11,680.3	22.1	698.1	28.4	1,005.5
Long-context RAG	10	23,489.4	10,617.1	21.3	645.7	26.5	891.6
Preference Optimization	10	27,719.5	16,148.1	21.8	770.7	26.2	1,043.3
Hate Speech Detection	15	20,461.7	10,346.1	21.6	674.9	26.2	949.3
Long Video Understanding	15	21,358.1	9,968.3	21.4	641.1	25.5	884.2
Representation Learning	15	12,797.0	6,741.3	22.4	643.1	28.1	896.6
Social Computing	15	10,538.7	6,769.5	21.9	670.0	24.9	889.8
Video Segmentation	15	14,916.3	7,370.3	21.9	591.1	27.3	810.8
Interpretability	20	14,119.2	7,102.6	22.8	786.0	27.1	1,066.1
Low-Resource NLP	20	21,014.8	8,199.6	23.7	832.0	31.4	1,211.4
Automatic Speech Recog.	25	8,319.0	4,663.8	24.0	799.7	31.3	1,191.1
Data Augmentation	25	20,189.6	9,917.1	23.6	814.9	28.7	1,118.8
Ethical Bias & Fairness	25	15,579.7	8,177.9	24.7	887.5	29.5	1,210.6
Legal NLP	25	19,057.4	9,225.8	21.9	822.3	26.9	1,163.8
LLM for Healthcare	25	28,137.2	12,311.9	22.2	804.9	27.5	1,156.3
Dialogue Systems	30	10,908.1	6,476.8	23.0	786.5	27.8	1,062.1
Quantization	30	26,699.4	10,034.4	23.0	865.8	30.7	1,311.3
Reinforcement Learning	30	14,316.1	7,969.5	23.5	747.4	28.3	1,022.4
Graph ML	35	14,621.5	7,810.6	24.1	777.7	29.6	1,080.4

Table 1: Summary statistics of the document collection across research themes. The table reports the number of files, average file length in tokens and words, and the average length of questions and answers in both words and tokens.

semantically similar themes, and in which the adjacency relations capture how similar or different any given themes are. Given a query, we then identify the most relevant themes to it, and consider them and their neighbors for obtaining appropriate “related” context for generating the reflections.

#### 4.1 Theme-based Document Representations

To capture the neighborhood relation between different themes or topics, we represent the given document collection using a theme-based graph structure by performing clustering over the document chunks. For this, we chunk the documents into semantically coherent chunks, preserving sentence boundaries (chunk size of approximately 2K tokens), and embed them using pre-trained Cohere embeddings<sup>7</sup>. We perform K-means clustering (Lloyd, 1982) on these embeddings, where each cluster approximately represents a semantic theme or sub-theme in the given collection (we set the number of clusters to `num_cluster`). Selecting an appropriate `num_cluster` is crucial: too small a value merges loosely related chunks, reducing the diversity of *relevant but non-repetitive* information, while too many clusters lead to overlapping themes and redundant content.

Clustering serves two main purposes: (1) providing a thematic representation of the document collection that clearly identifies key thematic regions, and (2) reducing redundancy by grouping semantically similar chunks. The cluster graph is constructed with edges between centroids representing distances in the embedding (semantic) space,

<sup>7</sup><https://cohere.com>

capturing relationships between clusters. Not all clusters are equally relevant to a query, but the *neighborhood surrounding the most relevant clusters* often contains supporting information that is contextually useful and complementary to what the answer itself would cover.

#### 4.2 Related Insight Generation

**Context Selection:** To identify context relevant to a given answer, we first determine which clusters the answer aligns with most closely. The answer is split into chunks of roughly 2K tokens, preserving semantic boundaries, and each chunk is embedded in the same vector space as the document chunks. For each chunk, we locate the nearest cluster centroids in the cluster graph, and the chunks in these clusters form the *answer-specific context*. To broaden the context, we select the top- $k$  clusters nearest to the answer-specific clusters based on centroid distances; the chunks in these clusters form the *related context*. These clusters contain information **connected to the main question but not already included in the answer**, providing additional perspectives or supporting details. The hyperparameters  $k$  and `max_hops` (the maximum distance from answer-specific clusters) control the extent of context expansion. Increasing the distance allows access to more **creative or tangential** information, while careful tuning ensures relevance.

**Related Insight Generation:** This module performs the selection of appropriate insight types and the generation of meaningful, actionable insights within a single, systematic process. We define a fixed set of insight types that support learning and

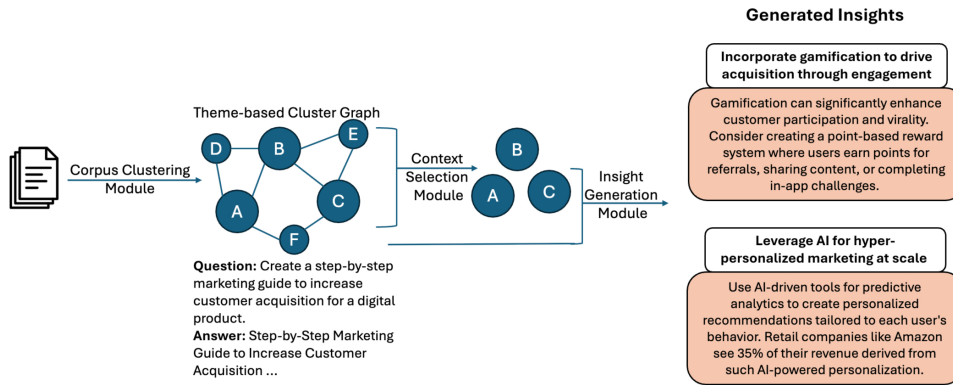


Figure 3: INSIGHTGEN pipeline showing theme-based clustering, context selection, and insight generation

creative exploration, including identifying *missing information*, *proposing new ideas*, suggesting *alternate answer framings*, creating *mind maps*, highlighting *potential issues or objections*, presenting *interesting facts*, designing *short quizzes*, providing *real-world applications or analogies*, and analyzing *tradeoffs*. This seed set is designed to aid common personas such as students, analysts, and researchers who routinely work with research paper collections. Given a question–answer pair, the model first infers the user’s intent and persona using a Chain-of-Thought prompt, and then generates insights that add new information beyond the original answer without repeating it, and associates each insight with an appropriate type and a brief justification. The process takes as input the original question, the answer, the selected supporting context, and the inferred user intents and goals, and for each generated insight the model produces self-assessment scores along four dimensions: *relevance*, *novelty*, *usefulness*, and *intent alignment*, each measured on a scale from 0 to 5.

## 5 Experimental Setup

### 5.1 Baselines

Alongside our proposed pipeline, we evaluate four baseline methods to analyze the importance of structured neighborhood selection. **Direct GPT** uniformly allocates a fixed token budget across all documents and generates insights using a single prompt that specifies requirements such as relevance, novelty, and diversity. **GPT+CoT** follows the same two-step Chain-of-Thought (CoT) procedure as our pipeline, first inferring user intent and then generating insights, but operates on a fully truncated global context. The CoT prompts are kept identical to avoid any evaluation bias. To isolate the role of clustering-based neighborhood retrieval, we also include two FAISS-based baselines

(Douze et al., 2024). **FAISS** retrieves the same number of chunks for each question as those passed to **INSIGHTGEN**, generating insights directly from these chunks, while **FAISS+CoT** applies the same CoT prompts used in **GPT+CoT** to the retrieved content. This ensures that any performance difference is due to the quality of the context rather than the amount of context. These baselines help us to verify the hypothesis that relying only on truncated context or similarity-ranked chunks, without explicitly defining local neighborhoods, leads to less diverse and less novel insights.

### 5.2 Hyperparameter Setup

Our pipeline uses three main hyperparameters:  $k$ , which determines the number of neighboring clusters selected;  $\text{max\_hops}$ , which controls how far the traversal proceeds from the answer chunks; and  $\text{num\_cluster}$ , which denotes the number of clusters used by the KMeans algorithm. Hyperparameter tuning on the validation set of the Quantization collection identifies a single optimal configuration with  $k=5$ ,  $\text{max\_hops}=2$ , and  $\text{num\_cluster}=\sqrt{n}$ , where  $n$  is the total number of chunks in the collection. The same configuration is applied consistently across all experiments, covering 35 domains and over 3000 questions to validate the robustness and generalizability of the pipeline. Detailed hyperparameter selection and ablation studies are provided in Appendix B.

### 5.3 Evaluation Metrics

Each method generates up to five related insights for every QA pair. We use **GPT-4o** and **Claude 3.5 Sonnet** as base models for insight generation, and **Gemini 2.5 Flash** and **Claude 4 Sonnet** as stronger Judge LLMs for evaluation.

We use two complementary evaluation settings that capture different aspects of insight quality. In

Model	Research Theme	Gemini-2.5-Flash					Claude-4-Sonnet				
		Direct GPT	GPT+CoT	FAISS	FAISS+CoT	INSIGHTGEN	Direct GPT	GPT+CoT	FAISS	FAISS+CoT	INSIGHTGEN
GPT-4o	Inference Optimization	3.79	3.02	3.87	3.03	3.87	2.79	1.92	3.06	2.06	3.05
	LLM as Agents	2.99	3.25	3.44	3.12	4.14	2.04	2.32	2.56	2.16	3.37
	Preference Optimization	3.12	2.22	3.68	3.37	4.10	2.00	1.28	2.74	2.35	3.32
	Long-context RAG	3.61	3.20	3.81	3.25	4.15	2.76	2.06	2.86	2.30	3.41
	Representation Learning	3.20	2.81	3.55	2.93	4.14	2.16	1.86	2.63	2.05	3.47
	Long Video Understanding	3.49	2.66	3.73	3.14	4.15	2.67	1.53	2.64	2.22	3.22
	Social Computing	2.82	3.27	3.21	3.06	4.52	2.16	2.26	2.60	2.24	3.84
	Video Segmentation	3.51	2.82	3.62	2.97	4.39	2.60	2.05	2.66	2.12	3.43
	Hate Speech Detection	2.91	2.76	3.84	3.30	4.48	1.97	2.24	2.66	2.42	3.52
	Interpretability	3.24	3.13	3.45	3.03	4.23	2.24	2.21	2.66	2.07	3.57
	Low-Resource NLP	3.35	2.80	3.76	2.85	4.34	2.20	1.82	2.86	2.13	3.60
	Data Augmentation	3.01	2.71	3.84	2.93	4.39	2.08	1.94	3.01	2.07	3.81
	Ethical Bias & Fairness	3.22	2.95	3.33	2.85	4.36	2.05	1.71	2.49	2.43	3.81
	Automatic Speech Recog.	2.99	2.65	3.47	2.97	4.45	0.96	0.96	2.57	2.13	3.36
	LLM for Healthcare	3.15	2.54	3.73	3.09	4.31	2.38	1.76	2.92	2.38	3.53
	Legal NLP	2.75	3.30	3.65	3.26	4.35	2.04	2.03	2.86	2.55	3.66
	Dialogue Systems	3.06	2.71	3.69	3.07	4.27	2.39	1.91	2.91	2.18	3.53
	Reinforcement Learning	2.88	1.90	3.75	3.29	4.27	2.08	1.14	2.68	2.29	3.56
	Quantization	3.59	2.20	3.92	3.15	3.98	2.96	1.24	2.78	2.09	2.89
	Graph ML	2.71	2.11	3.93	3.00	4.26	2.06	1.65	2.86	1.95	3.63
Claude-3.5-Sonnet	Inference Optimization	3.18	3.32	3.09	3.98	4.45	2.69	2.27	2.61	2.29	3.11
	LLM as Agents	2.23	2.31	2.53	4.08	4.44	1.77	1.20	1.89	2.71	3.46
	Preference Optimization	2.95	2.98	2.59	4.07	4.31	2.31	2.03	2.26	2.63	3.24
	Long-context RAG	2.29	3.54	2.81	4.17	4.41	1.67	2.56	2.21	2.85	3.33
	Representation Learning	2.37	2.50	3.13	3.94	4.33	1.68	1.27	2.45	2.24	3.27
	Long Video Understanding	2.32	3.26	3.12	4.33	4.06	1.97	2.15	2.50	2.92	2.86
	Social Computing	2.67	3.57	2.75	4.07	4.01	1.81	1.98	2.17	2.38	3.14
	Video Segmentation	2.85	2.92	3.12	3.68	4.39	2.64	1.95	2.35	2.02	3.07
	Hate Speech Detection	1.79	2.59	2.77	4.24	4.53	1.57	1.55	2.20	2.85	3.45
	Interpretability	2.24	2.86	3.05	4.19	4.36	2.27	1.67	2.45	2.44	3.15
	Low-Resource NLP	2.95	3.27	3.09	4.04	4.38	2.21	1.58	2.52	2.49	3.35
	Data Augmentation	3.56	3.17	3.01	3.79	3.97	2.62	2.07	2.47	2.57	3.19
	Ethical Bias & Fairness	3.17	2.37	3.19	3.95	4.34	2.33	1.77	2.47	2.45	3.41
	Automatic Speech Recog.	3.13	2.55	3.27	4.08	4.32	2.17	1.17	2.51	2.55	3.14
	LLM for Healthcare	2.15	2.31	2.73	4.19	4.51	1.77	1.41	2.11	2.80	3.53
	Legal NLP	2.86	3.21	2.85	3.90	4.28	2.11	2.35	2.25	2.45	3.46
	Dialogue Systems	2.11	2.05	2.97	3.79	4.54	1.83	1.22	2.48	2.45	3.52
	Reinforcement Learning	1.63	1.56	3.55	3.89	4.50	1.02	0.85	2.72	2.43	3.46
	Quantization	2.11	2.57	3.32	4.03	4.37	2.61	1.99	2.44	2.40	3.09
	Graph ML	2.28	2.23	3.46	3.69	4.07	1.99	1.23	2.69	2.20	3.21

Table 2: Set-level scores across models and themes on research paper collections (SCOpE-QA). Best results are shown in green and second-best in red. Insight-level scores are reported in Table 10.

**Set-Level evaluation**, the Judge evaluates the complete set of five insights generated by each method and assigns a single score in the range 0-5. The score reflects overall quality based on four criteria: **Novelty**, which measures how much new information or ideas the insights introduce; **Diversity**, which captures how distinct the insights are from one another; **Relevance**, which assesses how well the insights address the original question; and **Depth**, which indicates whether the insights are substantive rather than superficial. This setting is more robust in that it evaluates the insights as a coherent whole and reflects their overall usefulness. However, as the number of methods or the number of insights per method increases, the evaluation context becomes larger, which can cause the Judge model to under-attend to methods or insights that appear in the middle of the input, potentially affecting score reliability. To address this limitation, we additionally use an **Insight-Level evaluation**, where a single insight is randomly sampled from

each method for a given question and evaluated using the same criteria except **Diversity**. This process is repeated ten times per question and the scores are averaged, making this setting more scalable and less sensitive to long contexts. At the same time, this evaluation incurs higher computational cost as the number of samples increases and may be unreliable if a particularly poor-quality insight is repeatedly sampled. Despite these limitations, the two settings offer complementary perspectives, and using both allows us to evaluate the generated insights from multiple angles.

## 6 Results and Discussion

**RQ1: How do models perform across diverse domains?** We evaluate our insight generation approach across diverse domains, including 20 scientific paper collections from **SCOpE-QA**, (Tables 2, 10) and 15 non-academic internal collections (Tables 3, 11). Across all 35 collections, our method consistently matches or outperforms ex-

Model	Domain / Theme	Gemini-2.5-Flash					Claude-4-Sonnet				
		Direct GPT	GPT+CoT	FAISS	FAISS+CoT	INSIGHTGEN	Direct GPT	GPT+CoT	FAISS	FAISS+CoT	INSIGHTGEN
GPT-4o	Legal Business Analysis	2.94	1.12	3.59	3.77	4.65	2.00	0.77	2.77	3.00	3.47
	Instagram Marketing	3.62	3.24	3.33	3.10	4.76	2.29	2.52	2.24	2.19	3.67
	Climate Change Awareness	3.50	2.78	3.17	3.22	4.78	2.67	2.33	2.61	2.22	3.78
	Climate Change Policy	3.12	3.53	2.88	3.53	4.77	2.47	2.82	2.41	2.71	2.94
	Gut Health Insights	3.55	2.70	4.05	2.75	4.15	2.85	1.95	3.15	1.85	3.45
	Finance	3.11	3.33	3.22	3.67	4.44	2.11	2.67	2.11	3.00	3.11
	Finance - Investment 2	2.64	3.41	3.14	3.46	4.73	2.05	2.41	2.14	2.23	3.50
	Legal & Regulatory Compliance	3.40	2.80	3.80	3.15	4.25	2.50	2.10	2.90	2.05	3.00
	Finance - Investment 3	2.73	3.73	3.07	3.80	3.73	2.07	2.67	2.20	2.53	2.67
	Hotel Sales Strategies	3.00	3.11	3.63	3.90	4.53	2.26	2.53	2.74	2.68	3.32
	Responsible AI Consulting	3.81	3.00	3.94	2.69	4.50	2.56	2.50	3.13	2.19	3.31
	Revenue & Finance Reports	2.70	3.25	3.35	3.25	4.45	2.15	2.45	2.55	2.75	3.25
	Responsible AI Consulting 2	2.70	3.35	3.35	3.05	4.50	1.65	2.55	2.35	2.25	3.45
	Summarization of Articles	2.78	3.09	3.78	3.26	4.61	2.09	2.26	2.70	2.35	3.61
	Twitter & Mental Health	3.13	3.25	3.88	2.38	4.31	2.31	2.25	3.13	2.13	3.63
	Claude-3.5-Sonnet	Legal Business Analysis	2.00	3.41	2.47	4.65	4.41	1.71	2.59	2.29	2.82
Instagram Marketing		3.00	3.71	1.43	4.24	4.38	1.71	2.48	1.52	2.62	3.43
Climate Change Awareness		2.61	4.11	1.28	3.94	4.61	2.06	3.00	1.28	2.67	3.78
Climate Change Policy		2.71	3.82	1.88	4.12	4.53	1.88	2.82	1.65	2.88	2.88
Gut Health Insights		3.60	3.95	1.75	3.70	3.65	3.10	2.10	2.35	1.80	2.45
Finance		2.44	4.33	1.67	3.44	4.56	1.78	3.11	1.67	2.89	3.33
Finance - Investment 2		2.46	3.82	1.82	4.09	4.41	1.68	2.73	1.50	2.68	3.41
Legal & Regulatory Compliance		1.95	4.05	2.10	3.60	4.35	1.25	2.50	1.95	2.50	2.95
Finance - Investment 3		2.40	4.20	2.13	3.87	4.73	1.65	2.82	1.82	2.59	3.65
Hotel Sales Strategies		3.21	3.63	2.37	4.32	4.53	2.00	2.42	2.05	2.79	3.21
Responsible AI Consulting		2.63	3.25	2.31	3.88	4.31	1.44	2.13	1.94	2.50	3.50
Revenue & Finance Reports		1.80	3.00	2.65	4.25	4.65	1.45	2.05	2.15	2.95	3.35
Responsible AI Consulting 2		2.10	1.20	2.15	4.25	4.70	1.25	0.60	1.75	2.80	3.35
Summarization of Articles		2.96	3.48	1.87	4.13	4.39	1.91	1.87	2.04	2.22	3.39
Twitter & Mental Health		2.63	2.88	2.69	4.50	4.38	2.19	1.88	2.31	2.31	3.25

Table 3: Set-level scores across models and domains on internal non-scientific document collections. Best results are shown in green and second-best in red. Insight-level scores are reported in Table 11.

isting baselines, demonstrating the effectiveness of clustering-driven representations and the importance of neighborhood structures for comprehensive insight traversal. In contrast, naively selecting context, whether by taking the first  $k$  tokens, the top relevant chunks, or using a chain-of-thought prompt, can result in reduced performance. This highlights that careful context selection is critical for effectively leveraging chain-of-thought reasoning and generating high-quality insights.

**RQ2: How does collection size affect insight generation?** In our study of how document collection size affects performance, we observe that for small collections, such as non-academic datasets, Sales, Climate Change, Finance-Investment, Social Computing, or Inference Optimization, full-context methods are often sufficient to generate competitive insights. In these settings, our pipeline maintains strong performance, achieving scores above 4.5 as judged by Gemini. However, as the collection grows larger, including collections like Graph ML, Dialogue Systems, or Quantization, the advantage of clustering-based representations becomes clearly visible. Direct prompting on large collections typically scores between 2-2.5 according to Gemini and even falls below 1 when evaluated by Claude, while our approach consistently achieves scores above 4

with Gemini and above 3 with Claude. This widening gap highlights that clustering-based methods capture broader context and produce higher-quality insights as collection size increases.

**RQ3: How stable is INSIGHTGEN with respect to hyperparameter choices?** We use the hyperparameter configuration described in subsection 5.2. The same setup generalizes well across all 35 collections, covering more than 3,000 QA pairs, two evaluation models, and two evaluation strategies. To further assess robustness, we conduct a detailed ablation analysis over key hyperparameters, including the clustering method, the number of clusters used in KMeans, the number of neighboring clusters ( $k$ ), and the maximum number of hops. Tables 6, 7, 8, and 9 demonstrate that performance remains stable across ablations, confirming the robustness of our selected configuration. Additional details are provided in Appendix B.

**RQ4: How robust are our conclusions to the choice of LLM judge?** To examine the stability of our conclusions, we compare evaluations from two judges, Claude and Gemini, both of which assign scores on the same 0-5 scale. Although the scoring range is identical, the two evaluators exhibit clearly different calibration behaviors. Claude is substantially stricter, concentrating most of its scores in

the lower half of the scale (scores  $\leq 3$ ), resulting in a lower overall mean (approximately 2.4) and very few maximum ratings (around 1% of scores equal to 5). In contrast, Gemini tends to favor the upper half of the scale (scores  $\geq 3$ ), yielding a higher mean score (approximately 3.3) and assigning the top score much more frequently (about 25%). Despite these differences in score distributions, both evaluators consistently rank INSIGHTGEN as the best-performing method across all settings, indicating that our main conclusions are robust to the choice of LLM judge.

**RQ5: How well does Neighborhood-Aware Retrieval perform compared to advanced Agentic RAG pipelines for insight generation?** We investigate whether insight generation requires fundamentally different retrieval strategies than standard and agentic RAG systems. Prior approaches improve relevance through similarity optimization, query expansion, and iterative refinement, but do not explicitly model structural diversity or inter-document relationships. To evaluate this, we compare INSIGHTGEN with three progressively stronger baselines: Iterative RAG, which uses query decomposition and multi-pass self-refinement (Asai et al., 2023; Chen et al., 2025); Multi-Query RAG, which applies query reformulations with LLM-based re-ranking (Ma et al., 2023); and Agentic RAG, which integrates techniques such as HyDE-style embeddings (Gao et al., 2023), multi-query expansion (Jagerman et al., 2023), decomposition (Ammann et al., 2025), iterative refinement, and re-ranking within a unified framework. For a controlled comparison, all methods use the same hyperparameter and configurations as INSIGHTGEN. Given the rapid evolution of RAG methods, instead of adopting any single end-to-end pipeline as-is, we utilize and combine the core components from these approaches to construct these RAG baselines for our analysis.

INSIGHTGEN consistently outperforms these baselines (Table: 4), showing that even highly sophisticated, self-refining retrieval pipelines focused on similarity and recall are insufficient for insight generation. In contrast, its neighborhood-aware evidence selection captures complementary structural signals beyond relevance, enabling more effective cross-document synthesis. These results highlight that modeling inter-document structure, rather than solely optimizing retrieval quality, is critical for generating deeper insights.

**RQ6: How robust is our evaluation across**

Theme	InsightGen	Agentic RAG	Iterative	Multi-Query
<i>Gemini 2.5 Flash</i>				
Inference Optimization	4.21	3.54	3.74	3.44
Dialogue System	4.14	3.50	3.55	3.49
Long Video Understanding	4.34	3.73	3.44	3.60
Social Computing	4.51	3.52	3.27	3.13
Video Segmentation	4.49	3.70	3.42	3.44
Quantization	4.09	3.57	3.61	3.56
<i>Claude 4 Sonnet</i>				
Inference Optimization	2.99	2.58	2.48	2.56
Dialogue System	2.75	2.64	2.56	2.49
Long Video Understanding	2.92	2.67	2.62	2.52
Social Computing	3.38	2.62	2.51	2.37
Video Segmentation	3.18	2.58	2.36	2.36
Quantization	2.67	2.51	2.54	2.43

Table 4: Comparison with advanced RAG baselines across themes and judge LLMs.

**different LLM judges?** Across 3,000 questions, cross-judge agreement between Gemini and Claude remains high, with 87% pairwise ordering agreement, 73% top-method overlap (top-2 Jaccard: 72%), and strong rank correlation (median Spearman  $\rho = 0.85$ ;  $p_{75} = 0.95$ ;  $p_{90} = 1.00$ ; mean within-domain  $\rho = 0.83$ ). All 32 Wilcoxon signed-rank tests comparing INSIGHTGEN against baselines remain significant after Bonferroni correction ( $\alpha' = 0.00625$ ), with large effect sizes (median  $r = 0.65$ ), indicating agreement across evaluators.

## 7 Conclusions

Open-ended question answering over multiple documents requires identifying relevant information across sources while avoiding redundancy, a challenge that standard QA systems often fail to address. To tackle this, we introduce the task of *document-grounded related insight recommendation*, which aims to generate additional insights that help users iteratively refine and explore answers. We curate and release **SCOPE-QA (Scientific Collections for Open-Ended QA)**, a high-quality dataset covering 20 research paper collections with over 3,000 human-validated question-answer pairs. Building on this dataset, we propose INSIGHTGEN, which uses clustering-based context retrieval to generate related insights that promote ideation, brainstorming, highlight overlooked facts, and surface potential limitations in existing answers. Evaluation across 20 research and 15 internal collections, over 3000 QA pairs and 500 documents, using GPT-4o and Claude-3.5 as generation models and Gemini-2.5 and Claude-4 as judges, demonstrates that INSIGHTGEN produces more novel, diverse, and higher-quality insights than baseline methods, with benefits increasing as document collections grow larger and more complex.

## 8 Limitations and Future Work

While our results are promising, this work has some limitations that suggest directions for future research. First, the number of generated insights is currently restricted to five due to API constraints. Future work could allow different questions to have a dynamic number of insights based on their complexity and explore strategies to optimize token usage and determine the optimal number of insights per question, ensuring both relevance and coverage while improving scalability and overall quality. Second, our study focuses only on English-language text; extending the approach to multilingual datasets would enable evaluation across different languages and scripts and enhance the generality of the method. Third, the current framework is limited to text data; incorporating multimodal datasets, such as images or structured data, would allow richer insights and better reflect real-world document usage. Addressing these limitations can further improve the flexibility, applicability, and impact of the framework.

## 9 Ethical Concerns

All 20 collections use publicly available, free-to-use papers to avoid copyright issues. Any non-public document details have been appropriately masked to protect copyright and maintain anonymity. Therefore, our work does not raise ethical concerns.

## References

- Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdiah Soleymani Baghshah, and Ehsaneddin Asgari. 2025. [Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation](#). *Preprint*, arXiv:2502.08826.
- Paul J. L. Ammann, Jonas Golde, and Alan Akbik. 2025. [Question decomposition for retrieval-augmented generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 497–507, Vienna, Austria. Association for Computational Linguistics.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-rag: Learning to retrieve, generate, and critique through self-reflection](#). In *The Twelfth International Conference on Learning Representations*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, and 1 others. 2022. [Constitutional ai: Harmlessness from ai feedback](#). <https://arxiv.org/abs/2212.08073>. ArXiv:2212.08073.
- Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. [Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks](#). *arXiv preprint arXiv:2412.15204*.
- Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. [Density-based clustering based on hierarchical density estimates](#). In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.
- Guanhua Chen, Yutong Yao, Lidia S. Chao, Xuebo Liu, and Derek F. Wong. 2025. [SGIC: A self-guided iterative calibration framework for RAG](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28357–28370, Vienna, Austria. Association for Computational Linguistics.
- Google DeepMind. 2023. [Gemini 1 technical report](#). <https://arxiv.org/abs/2312.11805>. ArXiv:2312.11805.
- Xuan Long Do, Bowei Zou, Shafiq Joty, Tran Tai, Liangming Pan, Nancy Chen, and Ai Ti Aw. 2023. [Modeling what-to-ask and how-to-ask for answer-unaware conversational question generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10785–10803, Toronto, Canada. Association for Computational Linguistics.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. [The faiss library](#).
- Juliette Faille, Quentin Brabant, Gwenole Lecorve, Lina M. Rojas-Barahona, and Claire Gardent. 2024. [Question generation in knowledge-driven dialog: Explainability and evaluation](#). *Preprint*, arXiv:2404.07836.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. [Precise zero-shot dense retrieval without relevance labels](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.
- Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. [ChainCQG: Flow-aware conversational question generation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2061–2070, Online. Association for Computational Linguistics.
- Baorong Huang, Juhua Dou, and Hai Zhao. 2023. [Reading bots: The implication of deep learning on guided reading](#). *Frontiers in Psychology*, 14:980523.

- Seonjeong Hwang and Gary Geunbae Lee. 2022. **Conversational QA dataset generation with answer revision**. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1636–1644, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653*.
- Zhonghao Li, Kunpeng Zhang, Jinghui Ou, Shuliang Liu, and Xuming Hu. 2025. **Treehop: Generate and filter next query embeddings efficiently for multi-hop question answering**. *Preprint*, arXiv:2504.20114.
- Zihao Lin, Zichao Wang, Yuanting Pan, Varun Manjunatha, Ryan A. Rossi, Angela Lau, Lifu Huang, and Tong Sun. 2025. **Persona-SQ: A personalized suggested question generation framework for real-world documents**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 210–247, Albuquerque, New Mexico. Association for Computational Linguistics.
- Stuart P. Lloyd. 1982. **Least squares quantization in pcm**. *IEEE Transactions on Information Theory*, 28(2):129–137.
- Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.
- Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2025. **A survey of multimodal retrieval-augmented generation**. *Preprint*, arXiv:2504.08748.
- Cheng Fu Shengyi Liao Xinghua Zhang Bingli Wu Haiyang Yu Nan Xu Lei Zhang Run Luo Yunshui Li Min Yang Fei Huang Yongbin Li Minzheng Wang, Longze Chen. 2024. **Leave no document behind: Benchmarking long-context llms with extended multi-doc qa**. In *Proceedings of EMNLP*, pages 5627–5646.
- Mao Nakanishi, Tetsunori Kobayashi, and Yoshihiko Hayashi. 2019. **Towards answer-unaware conversational question generation**. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 63–71, Hong Kong, China. Association for Computational Linguistics.
- OpenAI. 2023. GPT-4 technical report. <https://arxiv.org/abs/2303.08774>. ArXiv:2303.08774.
- Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. 2019. **Reinforced dynamic reasoning for conversational question generation**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2124, Florence, Italy. Association for Computational Linguistics.
- Shimei Pan and James Shaw. 2007. Natural language query recommendation in conversation systems. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI’07*, page 1701–1706, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Dan Pelleg, Andrew W Moore, and 1 others. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml*, volume 1, pages 727–734. Stanford, CA.
- Chaitanya Sharma. 2025. **Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers**. *Preprint*, arXiv:2506.00054.
- Zihan Song, Gyo-Yeob Hwang, Xin Zhang, Shan Huang, and Byung-Kwon Park. 2025. A scientific-article key-insight extraction system based on multi-actor of fine-tuned open-source large language models. *Scientific Reports*, 15(1):1608.
- Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. **From neural sentence summarization to headline generation: A coarse-to-fine approach**. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4109–4115.
- Anuja Tayal and Aman Tyagi. 2024. **Dynamic contexts for generating suggestion questions in rag based conversational systems**. In *Companion Proceedings of the ACM Web Conference 2024, WWW ’24*, page 1338–1341, New York, NY, USA. Association for Computing Machinery.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. <https://arxiv.org/abs/2307.09288>. ArXiv:2307.09288.
- Siyuan Wang, Zhongyu Wei, Zhihao Fan, Yang Liu, and Xuanjing Huang. 2019. **A multi-agent communication framework for question-worthy phrase extraction and question generation**. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7168–7175.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Ruofan Wu, Youngwon Lee, Fan Shu, Danmei Xu, Seung won Hwang, Zhewei Yao, Yuxiong He, and Feng Yan. 2025. **Composerag: A modular and composable rag for corpus-grounded multi-hop question answering**. *Preprint*, arXiv:2506.00232.
- Xiaoming Zhang, Ming Wang, Xiaocui Yang, Daling Wang, Shi Feng, and Yifei Zhang. 2024. **Hierarchical retrieval-augmented generation model with rethink for multi-hop question answering**. *Preprint*, arXiv:2408.11875.

Zhuocheng Zhang, Yang Feng, and Min Zhang. 2025. [Levelrag: Enhancing retrieval-augmented generation with multi-hop logic planning over rewriting augmented searchers](#). *Preprint*, arXiv:2502.18139.

Zhonghua Zhao, Shanqing Guo, Qiuliang Xu, and Tao Ban. 2008. G-means: a clustering algorithm for intrusion detection. In *International Conference on Neural Information Processing*, pages 563–570. Springer.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37.

## A Scientific Collections for Open-Ended QA (SCOpE-QA) dataset

### A.1 Overview of Collection and Dataset Construction

Document-grounded related insight generation works with three inputs: a collection of related documents, an open-ended question, and its answer. The goal is to generate additional insights that are grounded in the documents, helping users better understand the question and interact with or refine the answer. This setting naturally requires reasoning across multiple documents, which reflects how people read and combine information in real-world scenarios. However, existing multi-document QA datasets such as FanOutQA (Zhu et al., 2024), LongBench v2 (Bai et al., 2024), and Loong (Minzheng Wang, 2024) mainly focus on factual retrieval or structured question formats. They do not capture the open-ended reasoning and cross-document synthesis needed for insight generation. To fill this gap, we introduce the **Scientific Collections for Open-Ended QA (SCOpE-QA)** dataset. SCOpE-QA is a publicly released benchmark designed for open-ended, multi-document reasoning. It includes 20 research themes with collections of varying sizes, and a total of 3000 open-ended question-answer pairs. Each question requires integrating and condensing information spread across multiple documents rather than retrieving isolated facts. The o3-mini model is used only during dataset construction for document compression and question-answer generation, and is not used in insight generation, baselines, or evaluation. These design choices make SCOpE-QA a reliable benchmark for studying document-grounded insight generation, supporting cross-document reasoning and open-ended synthesis of information.

**1) Collection creation and cleaning:** The first component of the SCOpE-QA dataset is 20 research paper collections covering a diverse set of scientific topics. To construct these collections, we first identify seed papers and popular research themes using the chat interfaces of ChatGPT<sup>8</sup> and Perplexity’s Deep Research mode<sup>9</sup>. Both tools are prompted to list major research themes along with representative papers for each theme. Once seed papers are identified, each collection is expanded using Connected Papers<sup>10</sup>, which generates a similarity graph based on co-citation patterns and bibliographic coupling. Papers that share a substantial number of references or citations with the seed paper are considered topically related. From the resulting graph, papers are first filtered using automated checks: title relevance and abstract similarity to the seed paper are measured using cosine similarity, and non-English, incomplete, or inaccessible documents are removed. After these automated checks, a careful manual review is performed to validate the topical relevance of each paper based on its abstract and introduction. This combination of automated filtering and manual verification ensures that each collection is focused, coherent, and reliable for downstream analysis. The final dataset contains 405 documents organized into 20 research collections, with each collection containing between 10-35 papers. The variation in collection size allows us to study the effect of context size on model performance, while maintaining high-quality, theme-specific document sets.

**2) Document Compression:** Once the collections are finalized, each document is processed using the o3-mini model to create a compressed, highly informative representation. The compression is performed so that multiple papers can fit within the context when generating question-answer pairs. Since the task focuses on open-ended questions, these compressed representations retain all key information from each paper, including main topics, conclusions, contributions, and essential findings, while removing non-essential material such as formatting markers, reference lists, and boilerplate text. This enables the generation of answers that integrate ideas and contributions spanning multiple documents.

<sup>8</sup><https://openai.com/index/introducing-deep-research/>

<sup>9</sup><https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research>

<sup>10</sup><https://www.connectedpapers.com/>

**3) Question Generation:** Using the compressed representations, question generation is performed over combinations of multiple papers to ensure that questions can span ideas across documents. For collections with up to 15 documents, 5-paper combinations are sampled, while for larger collections, 10-paper combinations are used. For each combination, 10 candidate questions are generated. To ensure balanced coverage of the collection, 50 combinations are randomly sampled per collection, resulting in approximately 500 candidate questions. These candidates are then filtered and refined using the o3-mini model to produce 200 high-quality questions per collection (250 for the Quantization theme).

**4) Answer Generation and Verification:** Answers are generated by prompting the o3-mini model with the same sampled combinations of compressed documents used for question generation. The model is instructed to cite relevant documents whenever a specific source is referenced. To remove highly redundant answers, pairwise cosine similarity is computed between candidate answers, and any answers exceeding a similarity threshold of 0.9 are removed. The remaining question-answer pairs are then subjected to careful human verification for factual correctness. Human annotators consult the full original documents to ensure that answers accurately reflect the content of the papers. After de-duplication and verification, 150 question-answer pairs are retained per collection (200 for the Quantization theme). For Quantization, the QA pairs are further split into 150 for test and 50 for validation set to support hyperparameter tuning.

The final dataset consists of 20 document collections and 3000 QA pairs generated through the multi-step pipeline that combines automated processing with human validation, ensuring high quality for cross-document open-ended question answering. Appendix A presents an overview of the dataset curation pipeline while Tables 1 and 5 report key statistics, including the number of files, tokens, and words in each collection.

Apart from the 20 research themes from SCOpE-QA, we also evaluate our pipeline on 15 internal, non-public document collections. These include financial reports, consulting notes, legal documents, and sales materials. Drawn from real organizations, these collections contain real-world questions such as, such as drafting a sales strategy or understanding the business model of a company. We use the same pipeline configuration, prompts, and hyper-

parameters as in the scientific setting, without any domain-specific changes, so performance on these internal collections reflects the robustness and generalizability of our method.

## A.2 Collection Statistics

We provide an overview of the document collections used in our study, summarizing key characteristics across various research themes. Tables 1 and 5 present the number of files, average file length in tokens and words, as well as average lengths of questions and answers in both words and tokens. These statistics offer insight into the scale and complexity of each collection, which is relevant for understanding model performance and resource requirements.

## B Hyperparameter Setup

Our pipeline uses three core hyperparameters. The parameter  $k$  controls the number of neighboring clusters selected during retrieval. The parameter  $\text{max\_hops}$  determines the maximum traversal depth starting from the answer chunks. The parameter  $\text{num\_cluster}$  specifies the number of clusters used by the KMeans algorithm. Hyperparameter tuning on the validation split of the Quantization collection yields a single optimal configuration with  $k=5$ ,  $\text{max\_hops}=2$ , and  $\text{num\_cluster}=\sqrt{n}$ , where  $n$  denotes the total number of chunks in a collection. This configuration is fixed for all experiments and applied uniformly across 35 domains covering more than 3000 questions.

For insight generation, we use GPT-4o and Claude-3.5-Sonnet as the base models. All documents are segmented into chunks of up to 2k tokens using a sentence-boundary-aware chunking strategy. Embeddings for both our pipeline and the FAISS baselines are generated using `cohere-embed-v3-english`. Clustering is performed using `scikit-learn`'s KMeans implementation with  $\sqrt{n}$  clusters and a fixed random seed of 42 to ensure reproducibility. Evaluation is carried out using two judge LLMs and two evaluation strategies. The consistent performance observed across domains, evaluators, and evaluation protocols indicates that the chosen hyperparameters are stable and not specific to a particular evaluation setup.

**Ablation Setup** To further validate the robustness of our hyperparameter choices, we conduct ablation studies on six themes, consisting of the top three and bottom three performing themes. All

Domain	# Files	Avg. Tokens per File	Avg. Words per File	Avg. Question Length (words)	Avg. Answer Length (words)	Avg. Question Length (tokens)	Avg. Answer Length (tokens)
Legal Business Analysis	11	15,186.1	10,534.4	19.9	352.1	23.5	434.1
Instagram Marketing	7	2,853.0	2,169.4	18.3	363.8	20.8	457.1
Climate Change Awareness	7	797.7	586.6	21.0	317.6	23.3	375.1
Climate Change Policy	9	12,272.4	8,278.3	21.7	354.1	25.8	440.9
Gut Health Insights	8	7,619.1	4,113.9	24.0	377.0	29.2	519.5
Finance	7	6,825.3	5,398.6	27.0	340.1	33.3	439.0
Finance - Investment 2	7	6,220.1	4,894.0	19.6	327.3	24.0	410.0
Legal & Regulatory Compliance	8	22,948.3	16,233.5	26.1	388.4	30.0	480.6
Finance - Investment 3	9	15,812.1	5,957.8	28.8	336.1	35.7	428.4
Hotel Sales Strategies	13	2,987.6	2,260.1	28.8	323.0	33.8	395.5
Responsible AI Consulting	13	11,631.0	8,219.2	20.9	334.9	23.3	431.2
Revenue & Finance Reports	13	26,116.3	15,413.4	21.3	350.3	25.5	431.8
Responsible AI Consulting 2	13	12,180.8	7,971.5	21.6	214.5	26.9	262.7
Summarization of Articles	11	10,699.7	5,010.6	22.4	466.9	30.0	647.4
Twitter & Mental Health	11	6,614.7	3,788.9	25.8	382.8	29.2	495.8

Table 5: Summary statistics of the document collection across domains. The table reports the number of files, average file length in tokens and words, and the average length of questions and answers in both words and tokens.

Research Theme	Gemini 2.5 Flash				Claude 4 Sonnet			
	$\sqrt{n}$	$n/3$	$n/5$	$\sqrt[3]{n}$	$\sqrt{n}$	$n/3$	$n/5$	$\sqrt[3]{n}$
Inference Optimization	<b>4.28</b>	4.11	3.52	3.28	<b>2.98</b>	2.94	2.55	2.14
Dialogue Systems	<b>4.04</b>	3.58	3.89	3.48	<b>3.05</b>	2.69	2.67	2.50
Long Video Understanding	<b>4.37</b>	4.02	3.72	3.13	<b>3.03</b>	2.77	2.52	2.17
Social Computing	<b>4.37</b>	3.68	3.63	3.48	<b>3.29</b>	2.88	2.70	2.51
Video Segmentation	<b>4.37</b>	3.66	3.64	3.12	<b>3.02</b>	2.54	2.62	2.24
Quantization	<b>4.33</b>	3.72	3.63	2.99	<b>2.85</b>	2.57	2.54	2.07

Table 6: Ablation on the number of clusters used by KMeans. The hyperparameter num\_cluster controls how many clusters are formed over the chunk set. Results are reported on six research themes selected from INSIGHTGEN. Setting the number of clusters to  $\sqrt{n}$  gives the best and most consistent performance and is used in the final system.

ablations are evaluated using both evaluation models. Unless stated otherwise, the best-performing configuration from validation is used as the default.

**Number of Clusters** We first study the effect of the number of clusters used for KMeans. We evaluate four settings:  $\sqrt{n}$ ,  $n/3$ ,  $n/5$ , and  $\sqrt[3]{n}$ , where  $n$  denotes the number of chunks in a collection. Table 6 shows that the  $\sqrt{n}$  configuration consistently performs best across themes and evaluation models. Smaller values tend to merge loosely related chunks, which reduces the diversity of *relevant but non-repetitive* information. In contrast, larger values produce overlapping clusters, leading to redundant context. Overall,  $\sqrt{n}$  provides the best balance between coverage and diversity, and is used in all experiments.

**Maximum Hops** Next, we analyze the effect of the maximum traversal distance from answer-specific clusters. We test three values: 1-hop, 2-hop, and 3-hop neighborhood traversal. As shown in Table 7, setting max\_hops=2 achieves the best performance. A 1-hop traversal restricts context to very local information, often leading to repet-

itive insights. Increasing the distance to 3 hops introduces more tangential content, which reduces relevance. The 2-hop setting provides a controlled expansion of context while maintaining relevance, and is used as the default.

**Clustering Method** We also compare different clustering methods, including KMeans, X-Means(Pelleg et al., 2000), G-Means(Zhao et al., 2008), and HDBSCAN(Campello et al., 2013). X-Means and G-Means extend KMeans by automatically adapting the number of clusters based on the data distribution. For X-Means, we initialize the algorithm with KMeans and allow cluster splits based on a model selection criterion, with a fixed random seed of 42 for reproducibility. For G-Means, cluster splitting is guided by a Gaussianity test, using a significance level of 10% to allow moderate splitting in high-dimensional embeddings. HDBSCAN is a density-based clustering method that does not require specifying the number of clusters. We set the minimum cluster size to 5 and use cosine distance for clustering. Points labeled as noise are reassigned to the nearest cluster centroid to ensure full coverage. Table 8 shows that KMeans provides

Research Theme	Gemini 2.5 Flash			Claude 4 Sonnet		
	1 Hop	2 Hop	3 Hop	1 Hop	2 Hop	3 Hop
Inference Optimization	3.57	<b>4.47</b>	3.78	2.42	<b>3.08</b>	2.68
Dialogue Systems	3.60	<b>4.47</b>	3.68	2.38	<b>3.23</b>	2.53
Long Video Understanding	3.72	<b>4.47</b>	3.92	2.39	<b>3.20</b>	2.68
Social Computing	3.73	<b>4.61</b>	3.73	2.53	<b>3.42</b>	2.69
Video Segmentation	3.73	<b>4.53</b>	3.82	2.40	<b>3.23</b>	2.60
Quantization	3.28	<b>4.53</b>	3.43	2.18	<b>2.95</b>	2.28

Table 7: Ablation on traversal depth from answer chunks. The hyperparameter `max_hops` controls how far retrieval expands from the initial answer chunks. Setting `max_hops=2` gives the best and most consistent performance.

Research Theme	Gemini 2.5 Flash				Claude 4 Sonnet			
	KMeans	G-Means	HDBSCAN	X-Means	KMeans	G-Means	HDBSCAN	X-Means
Inference Optimization	<b>4.12</b>	3.62	3.82	3.88	<b>2.84</b>	2.69	2.48	2.69
Dialogue Systems	<b>4.22</b>	3.47	3.72	3.52	<b>3.08</b>	2.68	2.69	2.44
Long Video Understanding	<b>4.38</b>	3.52	3.42	3.31	<b>3.08</b>	2.58	2.54	2.23
Social Computing	<b>4.34</b>	3.59	3.54	3.61	<b>3.27</b>	2.92	2.53	2.69
Video Segmentation	<b>4.24</b>	3.68	3.70	3.59	<b>2.80</b>	2.60	2.79	2.57
Quantization	<b>4.32</b>	3.37	3.08	3.40	<b>2.83</b>	2.37	2.23	2.33

Table 8: Ablation on the clustering method used in the pipeline. KMeans gives the most stable and best overall performance and is used in the final system.

the most stable performance across themes. Other methods often produce either fewer or more clusters than  $\sqrt{n}$ , which affects neighborhood structure and downstream retrieval. Using KMeans with  $\sqrt{n}$  clusters results in better coverage and a more reliable neighborhood graph.

**Number of Neighboring Clusters** Finally, we study the effect of the number of neighboring clusters selected during retrieval. We evaluate three values for `top-k`. Results in Table 9 indicate that selecting 5 neighboring clusters provides the best tradeoff. Smaller values limit context diversity, while larger values introduce redundant or weakly relevant information. Based on this ablation, we set `top-k=5` for all experiments.

**Summary** These ablation studies demonstrate that our final configuration is robust, despite being tuned on a single validation collection. The same hyperparameter values are applied across all 35 themes without further adjustment, and consistently yield strong performance.

## C Design and Evaluation Setup

This section describes the workflow for generating and evaluating domain-specific questions, answers, and insights, along with the key criteria used to assess their quality.

**Question Generation** We use the `o3-mini` model with `temperature=0.7`, `max_tokens=2000` to generate questions that are exploratory, open-ended, and conceptually rich. Questions are designed to

encourage critical thinking and discussion across the domain. For example:

“How do different strategies for X reflect underlying assumptions about Y?”

During generation, questions are constrained to 15–35 words, avoid purely factual or dataset-specific queries, and emphasize conceptual understanding, implications, and trade-offs.

**Answer Generation** Answers are also produced using `o3-mini` with `temperature=0.7`. Each answer is structured to be direct, comprehensive, and grounded in the provided collection. Key goals include synthesizing insights from multiple sources, citing supporting evidence with exact paper titles, systematically addressing all aspects of the question, and avoiding information not present in the sources.

**Insight Generation** Insights provide supplementary perspectives that go beyond the base answer. Each insight consists of a short descriptive hook, a detailed explanation grounded in evidence or examples, and a clear statement of the expected understanding or realization it enables.

We explore two generation settings. In the full context setting, the `GPT-4o` model (`temperature=0.7`, `max_tokens=4000`) receives the entire text collection as input. Generated insights can involve alternative viewpoints, creative brainstorming, knowledge testing, or concept mapping. In the retrieval-based setting, only the `top-k` most relevant

Research Theme	Gemini 2.5 Flash			Claude 4 Sonnet		
	Top-k = 3	Top-k = 5	Top-k = 7	Top-k = 3	Top-k = 5	Top-k = 7
Inference Optimization	3.28	<b>4.39</b>	3.00	2.15	<b>3.16</b>	2.11
Dialogue Systems	3.31	<b>4.34</b>	3.00	2.28	<b>3.04</b>	2.09
Long Video Understanding	2.65	<b>4.63</b>	3.02	1.86	<b>3.16</b>	2.27
Social Computing	3.11	<b>4.53</b>	3.18	2.16	<b>3.37</b>	2.32
Video Segmentation	2.96	<b>4.40</b>	3.28	2.11	<b>3.07</b>	2.38
Quantization	3.06	<b>4.50</b>	3.21	2.12	<b>2.94</b>	2.35

Table 9: Ablation on the number of neighboring clusters retrieved. Setting top-k=5 gives the best and most consistent performance and is used in the final system.

text chunks are used as input via dense retrieval with FAISS and GPT-4o, allowing evaluation of how focused, retrieval-conditioned context affects insight quality and diversity.

**Evaluation Setup** Insights are evaluated using Gemini 2.5 Flash with scores from 0–5. Evaluation occurs at both the set and individual level. At the set level, four metrics are used: novelty (new ideas beyond the base answer), relevance (connection to the question and user intent), meaningfulness (actionable or informative content), and diversity (coverage of distinct perspectives). At the individual level, diversity is not applied. Each method receives a score along with a brief rationale describing its relative strengths and distinguishing features. This structured evaluation supports both qualitative and quantitative analysis of insight generation performance.

## D Insight-level Evaluation

Different methods tend to generate different types of insights for the same question. The underlying type of insight can also vary depending on how much context is visible during insight generation. Because of this design choice, we adopt a **set-level evaluation** strategy. In this setting, the LLM judge evaluates the generations from all five methods as a single set. This helps avoid over-penalizing one poor insight or over-rewarding a single strong insight within the set.

However, when the number of insights varies across methods, or when the number of methods or insights increases, context length can become a limitation. To test the robustness of set-level evaluation, we also propose an **insight-level evaluation**. For each question, one insight is randomly sampled from each method and evaluated using the same criteria, except for Diversity. This process is repeated ten times per question, and the final score is computed by averaging across runs.

Since insights are highly subjective, both evaluation settings use comparative prompts. The order of methods is randomly shuffled for every question to reduce positional bias. Tables 10 and 11 summarize the key findings. INSIGHTGEN consistently outperforms the baselines on the 20 research paper collections and remains competitive on the 15 internal non-scientific collections. This demonstrates the robustness of our method across evaluation strategies. The insight-level evaluation serves as a secondary confirmation of the set-level results, showing that our method remains stronger even when evaluation is performed at the level of individual insights.

## E Thematic Summaries of Collections

To highlight the domains, sub-domains, and topics shared across the 35 collections (20 public research collections and 15 internal collections), we provide thematic summaries to better understand the focus of the papers.

### E.1 Research Paper Collections: Fully Public

- **Inference Optimization:** This collection focuses on optimizing inference for large neural networks, mainly LLMs and Transformer-based models. It covers techniques like early exiting, adaptive layer skipping, block skipping, and parallel decoding. The main goal is to reduce inference latency and computational cost while maintaining model performance.
- **LLM Agents:** This collection explores LLM Agents, emphasizing reasoning, acting, and planning. It includes reinforcement learning, self-reflection, and tool use, aiming to build robust frameworks and architectures to improve the autonomy and capabilities of language model agents.
- **Preference Optimization:** This collection focuses on LLM alignment, exploring pref-

Model	Research Theme	Gemini-2.5-Flash					Claude-4-Sonnet				
		Direct GPT	GPT+COT	FAISS	FAISS+COT	INSIGHTGEN	Direct GPT	GPT+COT	FAISS	FAISS+COT	INSIGHTGEN
GPT-4o	Inference Optimization	4.02	3.48	3.98	3.42	3.98	2.78	2.13	2.89	2.30	3.03
	LLM as Agents	3.67	3.08	3.73	3.56	4.04	2.17	2.54	2.53	2.34	3.18
	Preference Optimization	3.37	3.73	3.70	3.45	4.07	2.33	1.81	2.56	2.38	3.13
	Long-context RAG	3.76	3.72	3.86	3.61	4.20	2.67	2.62	2.69	2.41	3.31
	Representation Learning	3.22	3.60	3.44	3.47	4.33	2.25	2.18	2.50	2.28	3.20
	Long Video Understanding	3.37	3.39	3.57	3.38	4.08	2.65	2.10	2.68	2.27	3.24
	Social Computing	3.56	3.35	3.71	3.32	4.24	2.30	2.47	2.50	2.35	3.52
	Video Segmentation	3.20	3.52	3.79	3.62	4.31	2.56	2.29	2.62	2.28	3.27
	Hate Speech Detection	3.70	3.25	3.83	3.57	4.20	2.18	2.41	2.61	2.55	3.39
	Interpretability	3.51	3.53	3.61	3.26	4.09	2.41	2.34	2.57	2.30	3.27
	Low-Resource NLP	3.49	3.54	3.86	3.40	4.20	2.37	2.35	2.74	2.31	3.37
	Data Augmentation	3.51	3.93	3.72	3.44	4.24	2.32	2.20	2.70	2.28	3.39
	Ethical Bias & Fairness	3.42	3.36	3.76	3.56	4.30	2.37	2.37	2.47	2.32	3.49
	Automatic Speech Recog.	3.49	3.32	3.63	3.28	4.19	1.39	1.31	2.38	2.13	3.21
	LLM for Healthcare	3.57	3.64	3.52	3.30	4.15	2.43	2.13	2.65	2.51	3.40
	Legal NLP	3.30	3.27	3.83	3.37	4.21	2.34	2.70	2.73	2.43	3.43
	Dialogue Systems	3.23	2.66	3.70	3.27	4.09	2.28	2.41	2.63	2.30	3.25
	Reinforcement Learning	3.31	3.35	3.72	3.34	4.18	2.11	1.70	2.51	2.27	3.20
	Quantization	3.80	2.99	3.86	3.39	3.89	2.83	1.90	2.78	2.26	2.99
	Graph ML	3.10	3.02	3.74	3.33	4.16	2.16	1.88	2.62	2.31	3.25
Claude-3.5-Sonnet	Inference Optimization	3.70	4.00	3.67	4.09	4.49	2.57	2.51	2.67	2.64	3.13
	LLM as Agents	3.29	3.51	3.25	4.05	4.35	2.07	1.52	2.05	2.71	3.29
	Preference Optimization	2.92	3.01	2.94	4.13	4.47	2.32	2.25	2.32	2.72	3.15
	Long-context RAG	2.91	3.86	3.22	4.24	4.48	2.02	2.62	2.31	2.95	3.35
	Representation Learning	3.02	3.75	3.13	4.12	4.07	1.93	1.87	2.48	2.50	3.16
	Long Video Understanding	3.15	3.21	3.64	3.88	4.37	2.09	2.39	2.42	2.93	3.10
	Social Computing	3.23	3.28	3.54	3.85	4.52	2.09	2.39	2.24	2.76	2.97
	Video Segmentation	2.43	3.03	3.15	4.17	4.47	2.38	2.13	2.44	2.58	3.24
	Hate Speech Detection	2.94	3.57	3.51	4.29	4.33	1.74	1.99	2.18	2.94	3.33
	Interpretability	3.52	3.59	3.60	4.03	4.19	2.60	2.24	2.48	2.51	2.88
	Low-Resource NLP	3.41	3.71	3.67	4.13	4.43	2.40	2.31	2.56	2.73	3.22
	Data Augmentation	3.45	3.87	3.35	4.07	4.39	2.50	2.46	2.62	2.83	3.11
	Ethical Bias & Fairness	2.90	3.22	3.12	4.21	4.50	2.54	2.43	2.51	2.65	3.20
	Automatic Speech Recog.	3.45	3.40	3.62	4.13	4.42	2.21	1.89	2.44	2.71	3.10
	LLM for Healthcare	3.60	3.67	3.51	4.02	4.32	2.00	1.84	2.17	2.96	3.44
	Legal NLP	3.61	3.75	3.64	4.15	4.27	2.40	2.67	2.32	2.83	3.33
	Dialogue Systems	2.71	2.63	3.63	3.80	4.43	2.14	1.86	2.36	2.60	3.34
	Reinforcement Learning	2.92	2.85	3.41	3.82	4.52	1.32	1.32	2.54	2.49	3.20
	Quantization	3.08	3.33	3.72	4.01	4.42	2.48	2.30	2.56	2.66	3.18
	Graph ML	2.75	2.78	3.70	3.68	4.18	1.92	1.74	2.55	2.34	2.95

Table 10: Insight-level scores are reported across models and themes on research paper collections. Best results are highlighted in green, with second-best results in red.

ference optimization methods like RLHF and Direct Preference Optimization. It covers algorithms and theoretical frameworks aimed at improving LLM behavior to match human preferences for helpfulness and harmlessness.

- **Long-context RAG:** This collection studies Long-context Retrieval-Augmented Generation (RAG) and LLMs, focusing on long-context understanding, retrieval vs extended context comparison, and novel architectures for handling long inputs efficiently.
- **Representation Learning:** This collection covers representation learning in NLP, including vector representations for words, phrases, and sentences. It explores neural architectures and training paradigms to develop effective and transferable representations.

- **Long Video Understanding:** This collection focuses on understanding long videos with multimodal LLMs, addressing context length, computation, and redundancy. Techniques include token compression, frame selection, and temporal reasoning for hour-long videos.
- **Social Computing:** This collection studies user characteristics and dynamics on social platforms like Twitter. Topics include measuring influence, identifying experts, and understanding interests, leveraging content and social network structure.
- **Video Segmentation:** This collection investigates Video Object Segmentation, focusing on unsupervised and zero-shot approaches. It explores architectures and methods that combine appearance, motion, and other modalities

Model	Domain / Theme	Gemini-2.5-Flash					Claude-4-Sonnet				
		Direct GPT	GPT+COT	FAISS	FAISS+COT	INSIGHTGEN	Direct GPT	GPT+COT	FAISS	FAISS+COT	INSIGHTGEN
GPT-4o	Legal Business Analysis	3.47	1.56	3.77	4.12	4.56	2.35	0.96	2.65	2.82	3.52
	Instagram Marketing	3.90	3.75	3.67	3.54	4.59	2.49	2.49	2.40	2.39	3.48
	Climate Change Awareness	3.64	3.41	3.59	3.42	4.64	2.66	2.39	2.61	2.53	3.53
	Climate Change Policy	3.82	3.75	3.35	3.66	4.34	2.66	2.58	2.38	2.59	3.35
	Gut Health Insights	3.62	3.17	4.01	2.97	4.50	2.70	2.21	2.97	2.03	3.57
	Finance	3.59	3.63	3.63	3.84	4.67	2.20	2.64	2.47	2.61	3.37
	Finance - Investment 2	3.38	3.75	3.61	3.46	4.64	2.44	2.67	2.47	2.43	3.56
	Legal & Regulatory Compliance	3.35	3.22	3.63	3.60	4.26	2.22	2.14	2.63	2.53	3.16
	Finance - Investment 3	3.33	3.96	3.47	3.83	4.31	2.25	2.73	2.46	2.77	3.08
	Hotel Sales Strategies	3.44	3.52	3.77	3.73	4.52	2.33	2.50	2.54	2.62	3.42
	Responsible AI Consulting	3.76	3.56	4.09	3.36	4.64	2.56	2.50	2.79	2.15	3.48
	Revenue & Finance Reports	3.35	3.81	3.92	3.84	4.38	2.28	2.54	2.66	2.65	3.03
	Responsible AI Consulting 2	3.19	3.92	3.42	3.69	4.36	2.07	2.53	2.38	2.19	3.35
	Summarization of Articles	3.38	3.53	3.67	3.58	4.38	2.40	2.34	2.68	2.37	3.27
	Twitter & Mental Health	3.44	3.57	3.63	3.28	4.57	2.58	2.52	2.69	2.23	3.49
	Claude-3.5-Sonnet	Legal Business Analysis	3.05	4.01	2.82	4.60	4.71	1.94	2.61	1.91	3.18
Instagram Marketing		3.19	4.06	2.02	4.40	4.76	2.25	2.68	1.50	2.95	3.37
Climate Change Awareness		3.42	4.16	1.47	3.92	4.72	2.33	2.81	1.16	2.69	3.77
Climate Change Policy		3.18	4.15	2.63	4.28	4.19	2.11	2.97	1.92	3.07	2.60
Gut Health Insights		3.62	4.04	2.54	3.88	4.04	2.72	2.48	1.81	2.53	2.80
Finance		2.96	4.39	1.63	3.57	4.49	1.97	3.06	1.07	2.67	3.24
Finance - Investment 2		3.31	4.17	2.01	3.97	4.67	2.10	2.81	1.33	2.77	3.38
Legal & Regulatory Compliance		2.25	3.87	2.38	3.88	4.45	1.71	2.41	1.66	2.59	3.18
Finance - Investment 3		2.97	4.35	2.84	4.07	4.73	1.82	2.88	1.72	2.52	3.44
Hotel Sales Strategies		3.56	3.94	2.29	4.27	4.36	2.35	2.62	1.63	3.02	3.32
Responsible AI Consulting		2.70	3.88	2.69	4.19	4.56	1.75	2.58	1.88	2.74	3.36
Revenue & Finance Reports		2.66	3.54	3.10	4.40	4.62	1.52	2.16	2.04	3.02	3.38
Responsible AI Consulting 2		2.97	2.46	2.40	4.08	4.57	1.46	0.81	1.47	2.65	3.26
Summarization of Articles		3.34	3.67	2.39	3.68	4.46	2.38	2.37	1.70	2.45	3.11
Twitter & Mental Health		2.99	3.34	3.16	3.98	4.46	2.24	2.18	2.46	2.56	3.36

Table 11: Insight-level scores across models and domains on internal document collections (non-scientific domain). Best results are highlighted in green, with second-best results in red.

for robust automatic segmentation.

- **Hate Speech Detection:** This collection studies detecting and explaining hateful content in multimodal memes. Approaches include classification, target identification, and generating rationales, often using language and vision-language models.
- **Interpretability:** This collection explores the interpretability of deep neural networks, mainly LLMs like Transformers. It focuses on understanding linguistic knowledge encoded in attention mechanisms and contextual representations.
- **Low-Resource NLP:** This collection studies low-resource and multilingual NLP, including cross-lingual transfer, embedding alignment, vocabulary generation, and architecture design to improve performance across diverse languages.
- **Data Augmentation:** This collection covers data augmentation in NLP, including rule-based, contextual, and generative methods applied to tasks like text classification and

sequence labeling, aiming to address data scarcity and improve model performance.

- **Ethical Bias & Fairness:** This collection investigates bias and fairness in AI, focusing on NLP and computer vision. It studies detection, quantification, and mitigation of biases like gender, racial, and intersectional stereotypes in embeddings, models, and classifiers.
- **Automatic Speech Recognition:** This collection focuses on improving end-to-end ASR models, including contextual biasing, personalization, and handling rare or domain-specific words. It explores integrating external knowledge, language models, and phonetic information.
- **LLM for Healthcare:** This collection advances large multimodal models for healthcare, mainly in medical vision-language tasks. Topics include foundation models, domain-specific datasets, visual question answering, report generation, and diagnosis.
- **Legal NLP:** This collection studies NLP for legal documents, covering automatic summa-

rization of cases, judgments, and verdicts, as well as prediction and explanation of legal outcomes, across languages and jurisdictions.

- **Dialogue Systems:** This collection explores advanced ML approaches for dialogue systems, mainly task-oriented dialogue. It emphasizes reinforcement learning for dialogue policy optimization, exploration, sample efficiency, and robustness.
- **Reinforcement Learning:** This collection focuses on reinforcement learning, exploring policy optimization techniques such as trust region methods, maximum entropy, and deterministic policy gradients for continuous control, with emphasis on stability and sample efficiency.
- **Quantization:** This collection studies quantization for neural networks, mainly LLMs, covering post-training and quantization-aware methods for weights, activations, and KV caches to improve efficiency on resource-limited devices.
- **Graph ML:** This collection focuses on graph machine learning and Graph Neural Networks, covering representation learning, semi-supervised learning, and architectural innovations like graph convolutions, attention, and hierarchical pooling for complex graph-structured data.

## F Qualitative Examples

### F.1 Example insights generated for different questions

Figures 4 and 5 show what we mean by insights. For an open-ended question like “What are the key findings of these documents?” on a collection discussing mental health analysis on Twitter, these insights go beyond simply summarizing the text. They draw attention to interesting facts, suggest new ideas, or highlight actionable points that can be included to improve answers. In this way, they become useful tools to enhance user interaction, make existing answers better, and improve overall user experience and engagement.

Table 12 presents selected insights generated by different baselines. To keep the table clear, we show only the best-performing variant for each method, with or without CoT. The examples clearly

illustrate the superior quality of INSIGHTGEN and show that Gemini’s judgments closely match manual inspection.

### F.2 Collection Visualization and Thematic Clustering

To better understand the structure and main topics of our document collections, we perform visualization and clustering of document segments. Each collection is first divided into smaller chunks to capture coherent information. Then, we use semantic embeddings to group segments that are similar in meaning. This process allows us to see clusters of related content and the overall thematic organization in each domain. Figures 6–8 illustrate these cluster graphs.

**Graph ML Collection** For the Graph Machine Learning (Graph ML) collection, clustering reveals several key areas of research. Each node in the graph represents a document segment, and edges indicate semantic relationships between them. The main themes identified include:

- Performance analysis of different graph network architectures
- Techniques for learning node embeddings in graphs
- Recent advancements in graph neural networks
- Considerations for ethical oversight in graph-based models
- Recommendations and applications at web scale

**Reinforcement Learning Collection** In the Reinforcement Learning (RL) collection, clustering allows us to identify recurring research directions and methods. The key themes observed are:

- Advances in distributional reinforcement learning
- Maximum entropy reinforcement learning approaches
- Regularized policy optimization techniques
- Random search methods in reinforcement learning
- Applications of RL in robotics

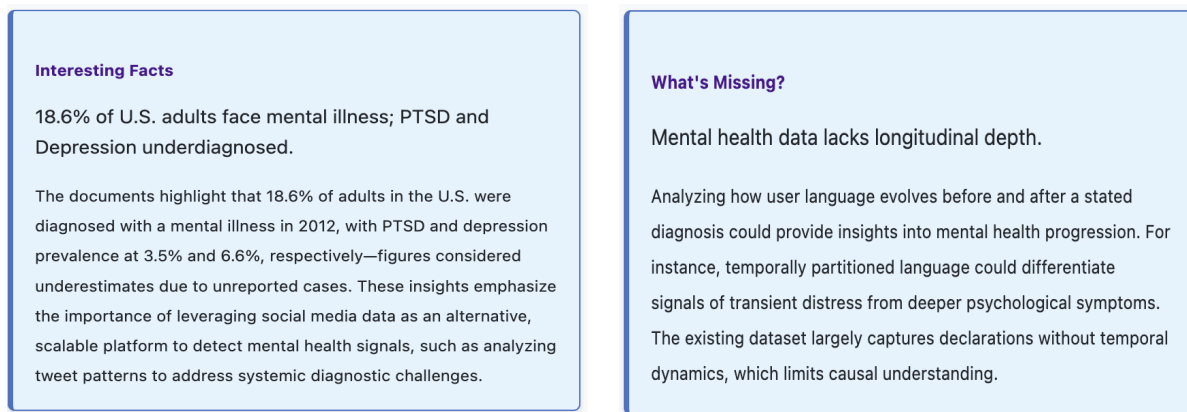


Figure 4: Example insights showing differences in type and content. These insights provide additional observations and reflections from a set of documents on analyzing mental health on Twitter.

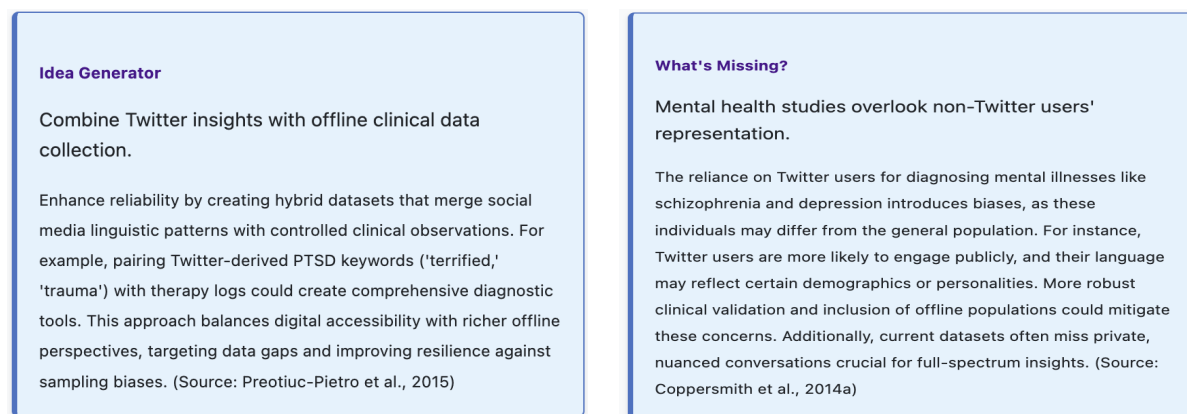


Figure 5: Example insights showing differences in type and content. These insights provide additional observations and reflections from a set of documents on analyzing mental health on Twitter.

- Analysis of continuous control algorithms
- Dynamics of rigid body motion in RL environments

**Quantization Collection** The Quantization collection focuses on research related to reducing model size and improving computational efficiency. Clustering shows several distinct themes:

- General quantization techniques for neural networks
- Analysis of outlier quantization methods
- Compression strategies for large language models (LLMs)
- Low-precision quantization and its effects
- Performance of LLaMA models under quantization

Overall, these visualizations provide a clear view of how different research themes are distributed

across the collections. By examining the clusters, we can identify major areas of focus, connections between topics, and the diversity of content within each domain.

### F.3 Example insights from our method for the interpretability question

- **Insight 1: Modular Syntax vs Task Layers**

**Hook:** Separate layers that learn syntax from layers that learn task details.

**Body:** Alternate layer roles across depth. Use lower layers to capture broad syntactic structure with low-rank projections, while higher layers handle task-specific reasoning using a small set of high impact attention heads. This makes the model easier to interpret and removes redundant computation.

- **Insight 2: Dynamic Rank and Head Adaptation**

**Hook:** Let the model adapt its rank and head usage based on the task.

Question	Domain	Method	Response
Give me campaign ideas for a product launch of a beauty brand based on these guidelines	Marketing	INSIGHTGEN	Inspire users with a 7-day transformation challenge. Engage audiences via a #7DaysToGlow campaign using daily Stories and branded hashtags, encouraging UGC and giveaways for the best transformations.
		GPT	Instagram is a critical platform for beauty brands due to its visual engagement. 81% use it to inform purchases, making it vital for visually-driven products.
		FAISS	Leverage Instagram’s power where 81% of users research products. Visual content is key to driving conversions and business site visits.
		Gemini Judgement	Method 1 offers novel, actionable, and diverse insights, while Method 2 repeats key points and lacks consistent novelty, and Method 3 mostly rephrases existing ideas with limited depth or originality.
Can you revise "The <HOTEL-CHAIN>Effect" paper to make it more engaging and convincing?	Sales	INSIGHTGEN	The paper overlooks recent travel trends shaping consumer behaviors. Consider integrating traveler archetypes (e.g., culture seekers, sun-and-beach lovers) and data showing 18% prioritize authentic experiences, using <HOTEL-CHAIN>’s regional cuisine programs and local artisan tie-ups as examples.
		GPT	Highlight <HOTEL-CHAIN>’s milestones to underscore its transformative role in the travel industry. Mention TV installations (1947), room service (1930), and global reservation systems (1947) to reinforce its legacy.
		FAISS	Incorporate facts about <HOTEL-CHAIN>’s innovations, TVs in rooms, reservation hotlines, and Connected Room concepts, to capture attention, though these are mostly restatements from source chunks.
		Gemini Judgement	Method 1 consistently adds novel, strategic insights that go beyond the original answer, unlike Method 2’s moderate expansions and Method 3’s occasional repetition.
How has the integration of contrastive reasoning, retrospective optimization, and episodic memory evolved in frameworks aimed at enhancing LLM as Agents’ decision-making capabilities?	Scientific	INSIGHTGEN	Explore how contrastive reasoning, optimization, and memory can enhance real-world autonomous systems, e.g., dynamic routing in AVs, action refinement in robotics, and personalized learning based on historical interactions, as shown in Reflexion’s cross-domain gains in robustness and success rates.
		GPT	Episodic memory enriches retrospective optimization by providing contextual feedback, enabling dynamic learning from failure. Reflexion leverages this for self-reflective learning, which could be enhanced with policy gradients for iterative policy refinement.
		FAISS	Episodic memory feeds retrospective optimization with contextual feedback, enabling adaptive failure-driven learning. Reflexion demonstrates this, with potential gains from adding policy gradients for iterative policy refinement.
		Gemini Judgement	Method 1 consistently delivers novel, deep, and diverse insights with clear examples, while Method 2 lacks novelty, and Method 3 suffers from rule violations.

Table 12: Comparison of insights across domains and methods. Method 1 is our pipeline, Method 2 uses direct GPT-4o prompting, and Method 3 combines FAISS-based retrieval with GPT-4o. Our pipeline consistently generates novel and actionable insights, while other methods tend to rephrase content. For Methods 2 and 3, we show the variant (with or without chain-of-thought prompts) that achieves the higher score for a clearer comparison.

**Body:** Build modules that adjust attention head usage and representation rank to match task complexity. For syntax-light tasks, prune most heads and use compact subspaces. For linguistically rich tasks, keep the essential heads and allow larger rank. Active components will then directly reflect task needs and are easier to inspect.

Pair these with low-rank encodings to make models faster and simpler to analyze. Performance can then be traced to a small number of clear components.

• **Insight 3: Interpretable Models for Real Time Use**

**Hook:** Apply simple, interpretable pruning for on device NLP systems.

**Body:** In real time or low resource settings, retain only a few interpretable attention heads that encode syntax, alignment, or position.

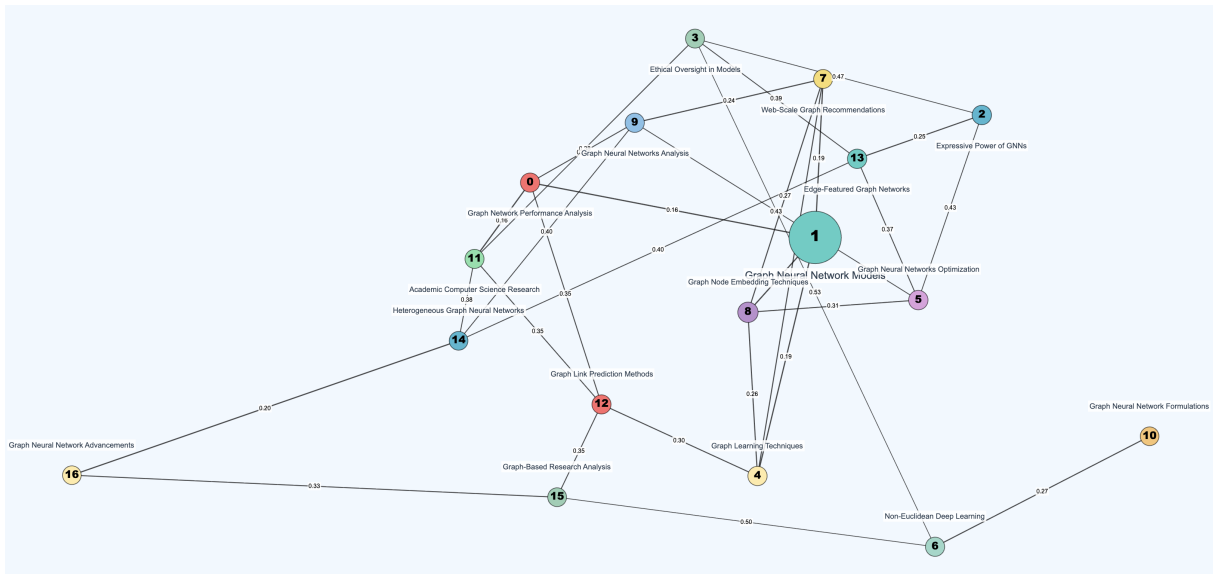


Figure 6: Thematic clusters from the Graph ML collection. Each node corresponds to a document segment, and edges represent semantic relationships. Clustering highlights distinct research themes within the Graph ML domain.

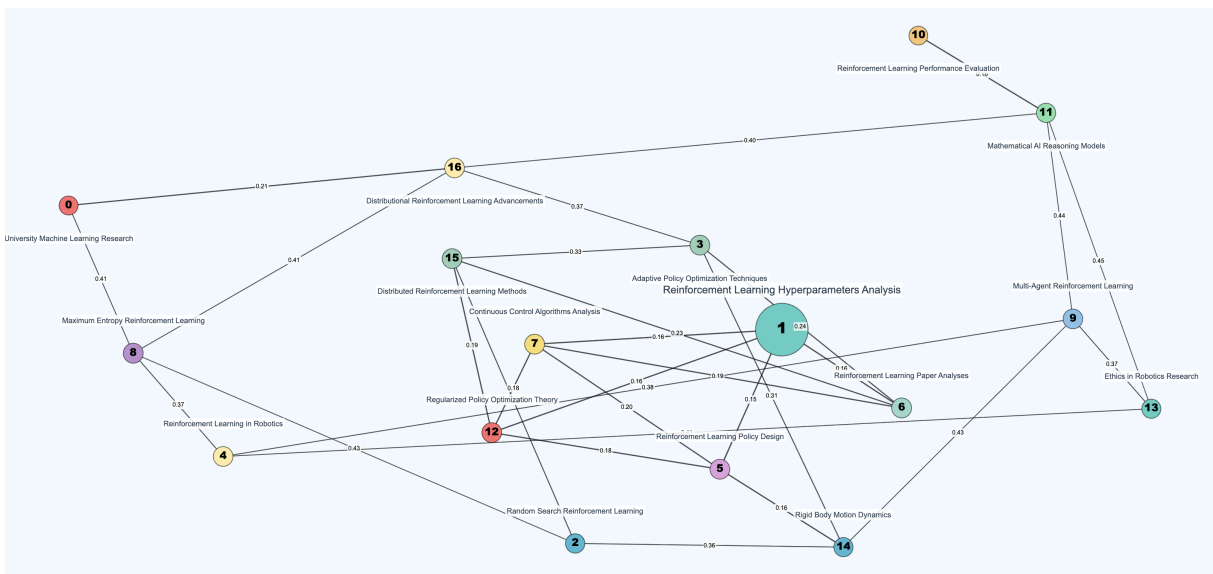


Figure 7: Clusters within the RL collection. Nodes represent semantically grouped document segments, showing recurring research topics and methodological patterns.

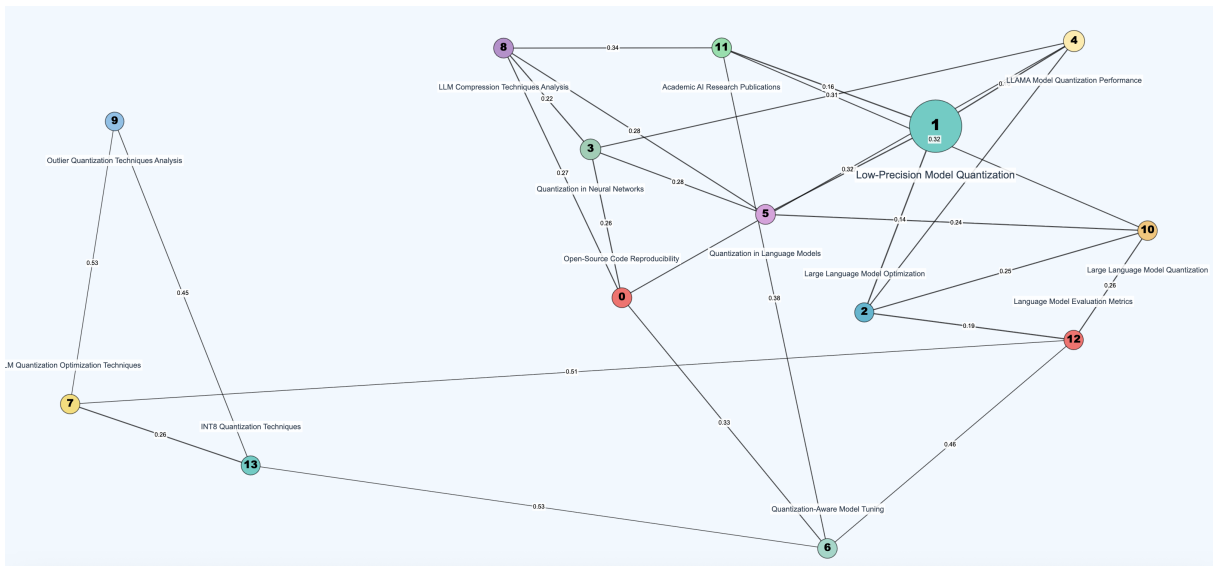


Figure 8: Clusters in the Quantization collection. Nodes correspond to document segments, and edges indicate semantic relationships. Clustering highlights both technical and application-specific topics within quantization research.