Knowledge Navigator: LLM-guided Browsing Framework for Exploratory Search in Scientific Literature

Uri Katz[1] **Mosh Levy**[1] **Yoav Goldberg**[1,2]
[1]Bar-Ilan University [2]Allen Institute for AI
{urikacid,moshe0110,yoav.goldberg}@gmail.com

Abstract

The exponential growth of scientific literature necessitates advanced tools for effective knowledge exploration. We present Knowledge Navigator, a system designed to enhance exploratory search abilities by organizing and structuring the retrieved documents from broad topical queries into a navigable, two-level hierarchy of named and descriptive scientific topics and subtopics. This structured organization provides an overall view of the research themes in a domain, while also enabling iterative search and deeper knowledge discovery within specific subtopics by allowing users to refine their focus and retrieve additional relevant documents. Knowledge Navigator combines LLM capabilities with cluster-based methods to enable an effective browsing method. We demonstrate our approach's effectiveness through automatic and manual evaluations on two novel benchmarks, CLUSTREC-COVID and SCITOC. Our code, prompts, and benchmarks are made publicly available.

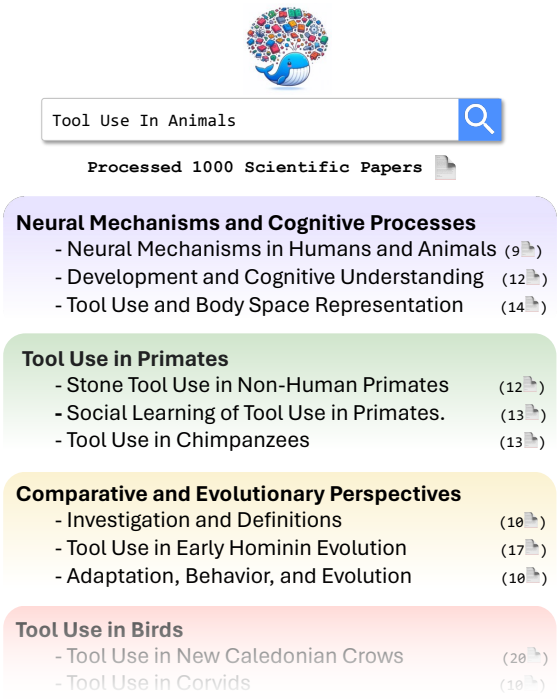


Figure 1: Hierarchical knowledge map generated by Knowledge Navigator, illustrating the primary themes and subtopics identified within a corpus of scientific literature retrieved for the query "Tool Use in Animals." This map demonstrates the system's ability to organize and structure knowledge on a broad topic.

1 Introduction

Traditional search engines, while adept at retrieving relevant documents for specific queries, are suboptimal when dealing with broad, topical queries. Such queries typically return lengthy ranked lists of potentially relevant papers, which, while comprehensive, overwhelms researchers with an information overload, obscuring the underlying structure of the topic and hindering the discovery of relevant subtopics and novel connections. Simply put, researchers are presented with an extensive inventory of documents without a clear map to guide their exploration, while they are interested in understanding broader topical trends.

The limitations of ranked list search results have driven a longstanding interest in methods for grouping and categorizing retrieved documents (Käki, 2005; Hearst, 2006). Over the years, a vast amount of research was devoted to designing different methods to support the paradigm of grouping and categorizing retrieved documents to organize them into a meaningful structure of knowledge, in many cases taking the form of hierarchical, cluster-based navigation based on automatically induced topical clusters (See §2). However, these browsing methods ultimately did not achieve widespread use and were not adopted by modern search engines, largely due to the insufficient quality of the automatically derived structures for practical application.

In this work, we demonstrate that importing the cluster-based navigation paradigm to the era of large language models (LLMs), combined with

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 8838–8855

modern NLP and IR methods, can overcome many of the obstacles faced by previous approaches. Our research shows that the components required to support this paradigm perform well as stand-alone tasks, and the entire framework functions effectively end-to-end.

We propose Knowledge Navigator[1], an LLM-based framework that transforms a large corpus of retrieved scientific literature into multi-level, organized themes of subtopics. Given a broad query, Knowledge Navigator generates a list of high-quality subtopics, each accompanied by a readable and interpretable summary grounded in the documents within the corpus. Each subtopic of interest can be expanded through ad-hoc secondary retrieval of fine-grained documents within that specific area. See Figure 1 for a graphical illustration of the system's real-world outputs. These subtopics represent meaningful research clusters, enabling searchers to identify areas of interest, uncover novel connections, and explore specific domains within the broader topic.

By shifting the focus from individual documents to organized subtopic clusters, Knowledge Navigator offers a potential solution to the challenges inherent in traditional ranked list presentations for scientific literature search.

We evaluate the Knowledge Navigator framework components on various LLMs and representation models, demonstrating its viability with both proprietary models (*GPT-4o*) and open-source models (*Mixtral-8x7B*). To evaluate the system's components and overall performance, we use (a) CLUSTREC-COVID, a modified novel form of the TREC-COVID benchmark (Voorhees et al., 2021), which we adapted for subtopic clustering, cluster-based aspect generation, and query generation tasks; and (b) SCITOC, a new dataset of scientific review table of contents, constructed from "Annual Reviews" open-access journals in a variety of scientific fields. We publish those datasets for future work in NLP research.

Using these benchmarks in both automatic and domain expert evaluation, we demonstrate that Knowledge Navigator performs efficiently in each of its component tasks, as well as in its overall function of organizing and outlining scientific knowledge. To our knowledge, this work is the first to showcase the feasibility of modern LLMs for supporting cluster-based navigation paradigms and aims to contribute to both the research and development of modern browsing systems.

[1]Knowledge Navigator code, evaluation datasets, and the Streamlit app can be found in https://knowledge-navigators.github.io

2 Exploratory Information Seeking through Knowledge Navigation

When scientists explore a new topic and review the literature, their information-seeking behavior is an instance of "exploratory search" (Meho and Tibbo, 2003; Soufan et al., 2022). That is when a searcher does not have a particular document or topic in mind but rather is interested in finding out the different aspects reflected in the *document collection*, to both gain an understanding of the overall structure of the domain, as well as to look for subtopics that may interest them and/or fit their expertise and interests (Marchionini, 2006; White and Roth, 2009).

Currently, this process is not well supported by modern search systems, although it has been found to be a common search behavior among scientists (Nedumov and Kuznetsov, 2019; Tahri et al., 2023). A major obstacle is the inability of searchers to effectively consume hundreds to thousands of documents and distill topics from them. Rather, they browse tens of document titles at a time, revise their mental model of the domain based on this subset, maybe take notes, adapt their query to reflect their new mental image and interests, and navigate into a specific aspect or subtopic of interest.

Knowledge Navigator enhances this exploration process by consuming hundreds of results, finding common themes, and organizing them into a two-level hierarchy of subtopics, allowing users to systematically explore different facets of a broad topic, thereby addressing the complex and multifaceted nature of their information needs. By transforming search results into organized subtopic clusters, Knowledge Navigator supports a holistic way of absorbing new information, helping them to identify areas of interest and discover novel connections even when their queries are initially vague or evolving.

Cluster-based browsing has been extensively studied in the past (Cutting et al., 1992; Hearst, 1999; Zamir and Etzioni, 1999; Osinski and Weiss, 2005), but despite this, these methods were not widely adopted. Approaches like Scatter/Gather (Cutting et al., 1992), which aimed to organize documents into coherent clusters for easier navigation, failed to produce good representations of docu-

ments in practice, leading to clusters that did not accurately differentiate between subtopics (Hearst, 1999). Additionally, these methods struggled to generate clusters that were easily interpretable by users due to reliance on keyword extraction techniques that were difficult for searchers to understand (Zhang et al., 2014). This was largely due to the immature state of Information Retrieval (IR) and Natural Language Processing (NLP) methodologies at the time.

The introduction of instruction-tuned LLMs, with their advanced world knowledge and text understanding abilities (Ouyang et al., 2022; Achiam et al., 2023), has recently led to renewed research utilizing LLMs for information organization across various tasks (Pham et al., 2023; Viswanathan et al., 2023; Zhang et al., 2023). We demonstrate that LLMs can be effectively used in assisting scientific literature consumption, by organizing document collections that result from a query into a digestible "table-of-content" of the topic, where each subtopic is grounded in concrete research works. This builds on the parametric knowledge of the LLM about scientific concepts, their categorization, the relations between them and many other facets of scientific knowledge gained in training, but relying on this knowledge in a fully grounded way, using it solely for the purpose of cataloging and organizing a given set of human authored documents, which result from a provided query.

3 Knowledge Navigator

Knowledge Navigator system takes a corpus of retrieved scientific documents for a given query and outputs an organized two-level thematic structure of subtopics spanning that topical query. The system's functionality is supported by the following conceptual steps: corpus construction, embedding and clustering of documents, describing and naming clusters, filtering irrelevant clusters, grouping the clusters into a thematic hierarchy, and subtopic query generation. This is implemented in a five-component architecture that largely follows the conceptual steps.

This architecture enables LLMs to generate grounded outputs based on a large number of source documents, a crucial requirement for organizing and structuring large corpora.

System Design LLMs are incredibly effective at consuming information but are expensive to run and bottlenecked by the amount of information they can effectively process in their prompts. Thus, we design the system around these constraints, attempting to minimize the number of LLM calls and aiming to make the most effective use of each one. Each step is designed to reduce the information size and transform it into a suitable form to be fed as input to the next LLM step. The system is thus designed to work bottom-up, progressively abstracting information at each stage.

Starting from hundreds of search-query results, represented as titles, abstracts and snippets, we employ a relatively cheap operation of contextual embeddings followed by clustering, to organize them into smaller—but cohesive—groups. Each group is then fed, as a whole but separately from other groups, into the Cluster Reader, an LLM-based component that is in charge of analyzing it, describing its common themes, naming it, and scoring its relevance to the query. Then, the names and descriptions of all the relevant clusters are fed (together with the initial query) into a subsequent LLM-based component, which organizes them into thematic groups, and names these groups. The result is a corpus-level hierarchical organization of the search results, which can serve as a map to the scientific topic, and whose construction relied on the entire corpus: the clustering is a global operation; the naming is separate per cluster but takes the query into account and considers many documents in assigning the description and name; and the second-stage thematic organization again provides a global view based on all the clusters who passed the relevance filter.

The searcher can then browse the generated topical outline, identify a sub-topic of interest, and initiate the Subtopic Expander to automatically generate a query to retrieve additional documents on the fine-grained sub-topic.

We now describe each component.

3.1 Topical Corpus construction

The initial step starts with a search query reflecting a relatively broad scientific topic T (e.g. "Tool use in animals"), and results in a topical corpus C comprising the top K documents ranked by a search engine for this query.[2] We select a large K (up to 1000) to ensure a diverse set of research papers that represent the full spectrum of the topic.

[2]In our research, we utilize Google Scholar via SerpAPI service, though the method is not confined to any specific search engine or retrieval technique as long as they are capable of retrieving a pool of relevant documents.

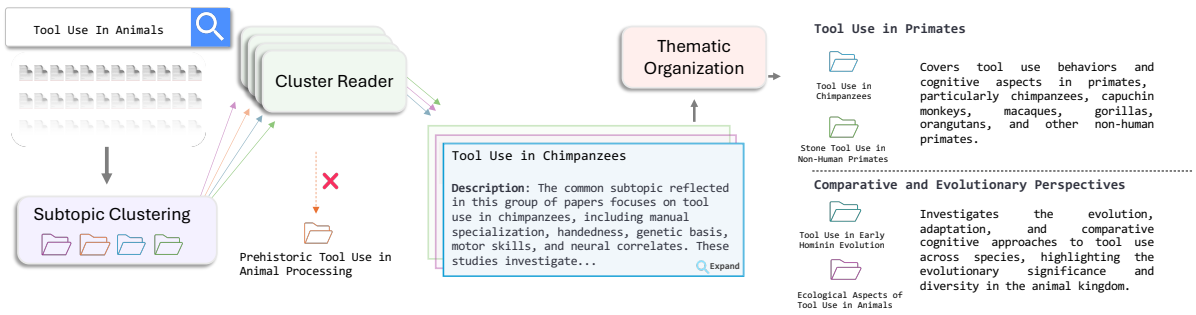


Figure 2: Knowledge Navigator Workflow: Starting with a query to a scientific literature retriever (e.g., Google Scholar), retrieved documents are embedded and clustered. The Cluster Reader then generates descriptive titles and descriptions for each cluster and filters for relevance. Finally, the Thematic Organization module groups the subtopics into a structured outline

3.2 Subtopic Clustering

Next, we aim to divide the corpus C into subtopic groups, each reflecting a sub-topic t_i of the broad topic T. This is done via clustering of contextual embedding vectors. We represent each retrieved document as a single embedding vector derived from the paper's title and its snippet and abstract (section 5.1 compares various embedders).

For clustering, we use Gaussian Mixture Model (GMM), a probabilistic soft-clustering algorithm that can assign each sample to one or more clusters.

The optimal number of clusters is determined using the Silhouette score (Rousseeuw, 1987), balancing cohesiveness and separation without penalizing the complexity of the GMM. Given the high-dimensional nature of modern text embeddings, and relatively low sample count, clustering methods face challenges in accurately assessing sample proximity, often resulting in poor quality clusters (Aggarwal et al., 2001). To address this, prior to clustering, we reduce the dimensionality of the vectors using UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018).

3.3 Cluster Reader

The Cluster Reader is an LLM-based component that operates independently on each subtopic cluster, receiving as input the titles and abstracts of all documents within that specific cluster, along with the initial topical query. This component serves two functions, –naming and filtering–, which are achieved in a single prompt (See C.1 for the exact prompt). Its output is a subset of the clusters, each with an associated name and description.

Describing and Naming Subtopics The Cluster Reader first reads the initial query, paper titles, and abstracts within each cluster to identify and articulate the specific subtopic they address. It generates a detailed description that encapsulates the thematic essence of the cluster in relation to the broader topic (T). Based on this description, it then generates a meaningful title for the subtopic. The output of this process is a detailed description of the subtopic, along with a subtopic title. The performance of the subtopic naming function is evaluated by a domain expert in §5.1.

Subtopic Filtering In the same LLM call, after the Cluster Reader generates a description and title for each subtopic cluster, it then scores the subtopic's relevance to the original topic T on a scale of 1 to 5. Based on this score, it determines whether to filter out the subtopic cluster. This filtering process eliminates clusters deemed explicitly unrelated to topic T, addressing the noise often present in large retrieved document collections on broad topics. The performance of this filtering function is evaluated by a domain expert in §5.2.

We chose to implement the three steps (naming, describing, and filtering) of the Cluster Reader as a single LLM call to induce a scratchpad reasoning process (Nye et al., 2021), where each step builds upon the previous stages. This also motivated the ordering of the generated content, from a detailed description to a concise name, to relevance scoring, and finally to relevance judgment.

3.4 Thematic Organization

The set of subtopics resulting from the cluster reader are diverse and fine-grained. The Thematic Organizer component takes all of the subtopic names and descriptions as inputs and groups them into meaningful thematic groups. For example, *"Dinosaur Thermoregulation and Metabolism"* and

"Dinosaur Musculature and Biomechanics" would be grouped under *"Physiology and Functional Morphology"*, while *"Evolutionary Transition from Dinosaurs to Birds"* and *"Origins and Ascent of Dinosaurs"* grouped under *"Evolution and Phylogeny"*. Such an organization greatly helps in browsing the list of results.

The cluster names and descriptions are sufficiently short for their entire set to fit in a single prompt, which is how the thematic organizer is implemented. The prompt contains the full set of topics and an instruction to organize them into higher level groups.

From a technical standpoint, we found it essential to associate each input topic with an explicit numeric ID and task the LLM with listing the IDs for each of its generated high-level themes, rather than to replicate the cluster names in its output. Using the IDs greatly reduced hallucinations, ensured consistent output, and increased coverage of the input topics. See C.2 for the exact prompt.

3.5 Subtopic Expander

The Subtopic Expander is an LLM-based component designed to enable searchers to automatically retrieve additional scientific documents relevant to fine-grained subtopics, allowing for deeper exploration of a specific subtopic without the need for manual query curation. Given the content of a subtopic cluster (i.e., its title, description, and assigned papers), the Subtopic Expander generates a list of terms directly related to the subtopic and specifically extracted from the scientific terminology present in the cluster's papers. These extracted terms are then concatenated into a single query, which is used to retrieve additional results that will populate the subtopic cluster. The expanded subtopic cluster can subsequently serve as an initial topical corpus for a secondary organization into a two-level thematic structure of subtopics. We conduct an isolated evaluation of the Subtopic Expander on CLUSTREC-COVID in 5.1, and an end-to-end system evaluation on SCITOC in 5.3. See C.3 for the exact prompt.

4 Structured Scientific Literature Benchmarks

An ideal dataset for evaluating the system end-to-end would necessitate exhaustive annotation of thousands of scientific documents, classifying each for relevancy, subtopics, and thematic groups. Unfortunately, no existing dataset supports this task comprehensively. To this end, we introduce two novel benchmarks, CLUSTREC-COVID and SCITOC, that facilitate a robust evaluation of Knowledge Navigator, both on an individual component level and end-to-end.[3]

CLUSTREC-COVID To assess the construction and naming of subtopic clusters, we modified TREC-COVID, which was initially designed as an information retrieval (IR) benchmark for evaluating search performance on scientific literature related to COVID-19 (Voorhees et al., 2021). It includes 50 expert-curated queries, each representing a subtopic of COVID-19 research across multiple fields, with each query serving as a concise subtopic title, such as 'coronavirus heart impacts'. For each query, medical experts judged hundreds of documents, annotating each for relevance to the topic, resulting in an average of 300 highly relevant documents per query. These characteristics—expert curation, diverse subtopics, and detailed relevance annotations—make the TREC-COVID benchmark particularly suitable for evaluating the Knowledge Navigator system's components. We transformed the TREC-COVID benchmark into a clustering benchmark by forming clusters from documents annotated as "highly relevant" to specific topics. We randomly sampled up to 50 documents from each topic, ensuring documents appear in only one cluster. This resulted in a dataset with 2,284 documents assigned to 50 subtopic clusters labeled with expert-curated titles.

SCITOC To assess the system's ability to handle complex scientific topics and produce well-organized outputs, we constructed a novel benchmark of 50 tables of contents (TOC) of scientific reviews sourced from 15 diverse peer-reviewed journals published by Annual Reviews[4]. These reviews span a wide array of scientific fields, including biology, medicine, food industry, environmental studies, and more. Reviews were selected based on predefined criteria: explicit and scientifically described tables of contents and subtopics, excluding metaphorical or abstract language. For instance, the review paper *The Effects of Psychedelics on Neuronal Physiology* (Hatzipantelis and Olson, 2024) included headers such as "Effects of Psychedelics on Gene and Protein Expression" and

[3] Benchmarks are publicly available in our GitHub project.
[4] annualreviews.org

"Effects of Psychedelics on Neuronal Survival and Neurogenesis"—which we keep—as well as structural headers like "Introduction," "Background," and "Discussion,"—which we discard. See an example from SciTOC in D.1.

For each review paper, we transformed the main title into a query focused on scientific terms to retrieve documents relevant to the review topic. Queries included only keywords appearing in the review title. For instance, the aforementioned title was modified to the Boolean query `"Psychedelic" AND "Neuronal Physiology"` to enable accurate retrieval of documents related to the main topic of the review.

5 Experiment

The Knowledge Navigator system comprises several components that have not been extensively explored, particularly within the domain of scientific literature. We present an evaluation of these components, both individually and as part of the integrated system, specifically focusing on the organization of topical scientific corpora.

5.1 CLUSTREC-COVID Experiments

Does clustering effectively discover subtopics in an already topical corpora? Clustering algorithms aim to group similar instances, but their success in accurately clustering scientific documents into coherent subtopics is uncertain due to the complex nature of scientific concepts, which may not align with the algorithms' similarity measures. Our goal is to assess whether the clustering component effectively organizes CLUSTREC-COVID documents into subtopic clusters that align with human annotations..

For the evaluation of the subtopic clustering component, we experiment with GMM clustering using various text representation methods. Following literature on document clustering in information retrieval (Yuan et al., 2022), we report relevance-based measures. We use "Clusters per topic by relevance" ($R_c@p$) metric which indicates the number of clusters needed to observe p% of relevant documents, while "coverage per topic by relevance" ($R_v@p$) shows the % of minimum number of documents needed to cover p% of relevant documents.

Table 1 shows that all models reconstructed topic groups with a fair degree of correlation between the gold mapping and clustered map. The best results were achieved by *text-embedding-3-large* and *SFR-mistral*. The $R_c@80$ metric indicates that, on average, a searcher needs to evaluate 2.3-3.18 clusters out of 50 to cover 80% of the documents in a topic, compared to an average of 20.4 clusters in a random cluster assignment, and 5-7% of the entire collection (See $R_v@80$) compared to 41% of the corpus documents. Interestingly, the *SPECTER2* model, despite being trained for scientific documentation, lagged behind general text embedding models. For the rest of the evaluations, we used the *text-embedding-3-large* model for ease of use via an API, but the results indicate that an open-source version would achieve similar performance.

Can LLMs accurately identify and name underlying subtopic in a document cluster? We evaluate the cluster reader's ability to replicate the cluster names assigned to CLUSTREC-COVID clusters by expert annotators. The Cluster Reader received the titles and abstracts of all the papers in a cluster as input and was tasked with generating a representative title for that cluster. Each generated title was then judged against the original subtopic query for a match, with the evaluation conducted by annotators with academic backgrounds in biomedical research.

Table 2 shows that the Cluster Reader (using *GPT-4o* or *Mixtral-8x7B*) successfully generated subtopics that matched 88% of the subtopics in CLUSTREC-COVID. Most unmatched subtopics were closely related to the benchmark's subtopics but not identical (e.g., "COVID-19 in African-Americans" (expert) vs. "Racial and Ethnic Disparities in COVID-19 Outcomes" (cluster-reader)). When papers were assigned to random clusters, the generated subtopic titles were mostly broad and general descriptions of COVID-19 research, resulting in very low coverage (6%). None of the subtopics were filtered by the Cluster Reader, as desired.

Can LLMs generate effective queries from clusters of scientific documents? To evaluate the effectiveness of the Subtopic Expander in generating queries from scientific document clusters, we created 50 subtopic clusters, each containing 20 randomly sampled relevant papers from one of the 50 topics in CLUSTREC-COVID. The Subtopic Expander then used the subtopic title, description, and associated papers to produce a specialized query. We assessed retrieval performance using the BM25 retriever against the entire TREC-COVID corpus (192K documents). We compared

	Adjusted Rand Index ↑	NMI↑	R_c@80 / R_v@80 ↓
text-embedding-3-large	**0.516**	**0.732**	2.4 / 5.1%
text-embedding-3-small	0.496	0.719	2.5 / 5.7%
SFR-Mistral 7b	0.513	0.736	**2.3 /5.1%**
SPECTER2	0.435	0.674	3.18 / 7%
Random	0.00	0.153	20.4 / 41%

Table 1: Comparison of different vector representations for clustering scientific documents to subtopics in the CLUSTREC-COVID benchmark

Experiment	% Subtopic Match
Random clusters + GPT-4o	$6\% \pm 3$
GPT-4o	$88\% \pm 4$
Mixtral-8x7B	$88\% \pm 4$

Table 2: Comparison of Generated Subtopic Titles to Ground Truth in CLUSTREC-COVID

the Subtopic Expander's performance against two baselines: (1) generating a new query based solely on the subtopic title without considering the papers in the cluster, and (2) using the original, unmodified TREC-COVID topic query. Remarkably, the queries generated by the Subtopic Expander significantly outperformed both baselines (see table 3), with improvements of up to 7.4% in precision@K and 14.2% in recall@K (see Appendix for details on recall@K) compared to the original TREC topic queries. These results demonstrate that this method enables searchers to achieve superior retrieval capabilities in fine-grained scientific subtopics without requiring prior knowledge of specific terms or jargon embedded in the subtopic cluster's papers.

5.2 SCITOC Experiments

We now present our evaluation based on the Scientific Reviews Table-of-Contents benchmark. Unless otherwise specified, the evaluations are based on a complete human annotation of 20 reviews, annotated by a hired academic researcher with a PhD in Biology. The annotator evaluated each generated subtopic in these 20 scientific review papers, resulting in a total of 1,471 relevant subtopics and 261 filtered subtopics. Each subtopic was assessed through multiple questions, which will be detailed in the following subsections. We performed an inter-annotator agreement assessment on a subset of reviews to evaluate the reliability of the annotation process. The results indicated a high level of agreement for the tasks evaluated. Further details are provided in B.1.

Can LLMs effectively filter irrelevant subtopic clusters? To assess the Cluster Reader's ability to filter non-relevant subtopics, the expert annotator evaluated the relevance of each generated subtopic title and summary to the original query topic. Relevance was defined as the subtopic having a direct and clear connection to the original topic. The annotation process covered both filtered and non-filtered subtopics to assess filter performance. The Subtopic Filter flagged 261 subtopic clusters as not relevant to the initial topic. Of these, 87.7% (229) were confirmed as non-relevant by the annotator. Conversely, only 0.14% (2 subtopics) of the 1471 non-filtered subtopics were marked as non-relevant. This indicates that the filter has a high precision and a very low false negative rate, resulting in an overall accuracy of 98.8%. effectively removing irrelevant subtopics while retaining those that are relevant.

Can LLMs organize subtopics into coherent thematic categories? To assess the thematic organization component, we evaluated the Knowledge Navigator's output for each topic in the Scientific Reviews benchmark. For each of the 1,471 generated subtopics, an expert annotator determined whether the subtopic was assigned to a relevant theme. For example, in the topic "Gut Microbiota in Colorectal Cancer," the subtopic *"Role of Probiotics in Colorectal Cancer Prevention and Treatment"* was correctly assigned to the theme "Therapeutic and Preventive Approaches Targeting Gut Microbiota," while the subtopic *"Impact and Modulation of Gut Microbiota in Colorectal Cancer"* was marked as a false assignment. Overall, we found that **95.2%** of subtopics were assigned correctly to themes within their respective topics.

5.3 End-to-End Subtopic Coverage Evaluation

We evaluate the Knowledge Navigator's overall performance to identify meaningful subtopics within diverse scientific domains by comparing its gener-

Query type	P	@20	@70	@100	@200
Original Query		0.43	0.36	0.33	0.27
Subtopic Expander					
Title		0.45	0.38	0.36	0.30
Title + Cluster		**0.50**	**0.43**	**0.41**	**0.33**

Table 3: Precision@K on TREC-COVID retrieval using different query generation methods for subtopic expansion

	Soft Heading Recall
GPT-4o Direct Gen	82.6%
GPT-4o + KN	**87.1%**
Mixtral-8x7B Direct Gen	75.3%
Mixtral-8x7B + KN	**88.3%**

Table 4: Results for automatic evaluation of subtopic coverage of SciTOC tables of contents for Knowledge Navigator (KN) and direct generation by *GPT-4o* and *Mixtral-8x7B*.

ated output to the table of contents (TOC) of corresponding human-authored review articles. Our assessment encompasses two key aspects: (a) the extent to which the generated subtopic titles match and cover the headers in the reviews TOC (subtopic coverage), and (b) the extent to which the generated subtopics introduce additional topics not present in the human-authored review (novelty of subtopics).

Overall Statistics The human-authored reviews have 10.5±1 valid subtopics on average, while the Knowledge Navigator produces an average of 73.5±3 relevant subtopics per topic.

Automatic Evaluation To assess system capabilities across the entire benchmark of 50 review papers for both *GPT-4o* and *Mixtral-8x7B* (Jiang et al., 2024), we employed a heading soft recall (Fränti and Mariescu-Istodor, 2023) evaluation method, as suggested in (Shao et al., 2024), to compare the recall of generated subtopics against human-authored review outlines. This method is suitable due to the comparison of subtopic title lists against an existing review table of contents.

As a baseline, we prompted the LLM directly to generate subtopic lists for the same topics given to Knowledge Navigator, assessing its ability to generate meaningful scientific subtopics based on its parametric knowledge. It's important to note that the evaluated reviews are publicly available and may have been encountered during the LLM's pretraining.

Table 4 shows that Knowledge Navigator outperforms the Direct Generation (Direct Gen) setup in both models[5], with a 13% improvement in coverage for *Mixtral*. This suggests that Knowledge Navigator can compensate for limitations in model scale, enhancing the identification of relevant subtopics. As we see in the next section, the knowledge navigator also produces many novel topics, not covered by either the review or the LLM.

[5]Both pairs show statistically significant differences in the paired t-test with $p < 0.01$

Notably, *Mixtral-8x7B* and *GPT-4o* achieve similar results for Knowledge Navigator (KN in the table), suggesting open-source models can be viable alternatives.

Domain Expert Evaluation We also grounded the automatic evaluation with full human expert evaluation over 20 of the reviews. To assess Knowledge Navigator's ability to identify meaningful scientific subtopics within a given topic, we compared its generated output to the corresponding review paper's Tables of Contents (TOCs). The review paper's TOC serves as an indicator of the foundational subtopics a reader should encounter. The annotator first identified all relevant headers in the review TOC, excluding headers of general background information unrelated to the topic or subjective headers that do not represent explicit subtopics (see D.3). Next, for all valid headers, the annotator matched generated subtopics to TOC headers only if they explicitly addressed the same subtopic. Unmatched yet relevant subtopics were classified as novel.

On average, Knowledge Navigator explicitly covered $71.6\% \pm 3$ of the review headers and generated 35 ± 4 novel subtopics per review (on top of the 10 topics in the review). This result demonstrates the system's ability to identify scientifically meaningful subtopics considered foundational by the experts who authored the reviews. Ultimately, the system aims to organize corpora of scientific literature into a structure that mirrors how an expert would approach the task, while being more exhaustive in its inclusion of relevant subtopics and themes for a comprehensive overview of the broader topic.

5.4 Expanding Subtopics by Retrieving Additional Relevant Papers

We evaluated the Subtopic Expander as part of the entire Knowledge Navigator on the expert-

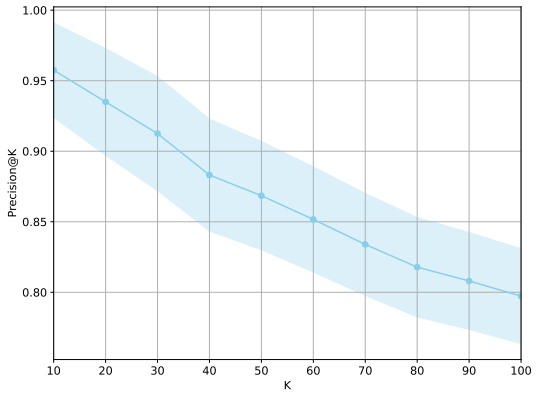


Figure 3: Average Precision@K of the K documents retrieved using a query generated by the Subtopic Expander for the SciTOC reviews.

annotated reviews to assess its capabilities in generating queries for the expansion of fine-grained scientific subtopics. We expanded 40 random subtopics, 2 from each of the 20 annotated topic reviews. Using the generated query in a Boolean form where each keyword is concatenated with an "OR" clause, we retrieved 100 documents from the search API. Overall, we collected 4,000 scientific papers for 40 different subtopics.

In order to conduct a relevant judgment evaluation over thousands of retrieved documents, we constructed and validated an LLM-judge capable of assessing a retrieved document's relevance to the subtopic title. The LLM-judge achieved a high level of agreement with the human expert. See E.1 for a detailed description.

Subtopic Expander Evaluation Using the validated LLM judge, we assessed the relevancy of retrieved papers to evaluate Precision@K. The order of the retrieved papers reflects the original sequence from the search API output. As shown in Figure 3, the results indicate that the retrieved documents achieve high precision, enabling the accurate retrieval of up to 100 new scientific papers within fine-grained subtopics without requiring the searcher to formulate a new query while maintaining reasonable precision. This demonstrates the system's potential for scientific exploration.

6 Related Works

Cluster-based Browsing notably the Scatter/-Gather paradigm (Cutting et al., 1992; Pirolli et al., 1996), was proposed to enhance exploratory search (Gong et al., 2012) by grouping retrieved documents into clusters and allowing iterative refinement. However, limitations in representing cluster content with extracted keywords hindered its adoption (Zhang et al., 2014). Since then, other notable methods have proposed different cluster-based approaches aimed at improving the computational efficiency of clustering algorithms or the selection of representative keywords (Zamir and Etzioni, 1999; Osinski and Weiss, 2005). Our work draws inspiration from this approach but leverages advancements in LLMs to structure and interpret documents, enhancing interpretability. Additionally, our focus on enabling exploration through multi-level subtopic hierarchies differentiates our system from Scatter/-Gather's document retrieval focus.

Information Organization with LLMs It was shown that LLMs are capable of clustering items (Viswanathan et al., 2023; Zhang et al., 2023; Wang et al., 2023) and uncovering latent topics in text collections (Pham et al., 2023). However, these methodologies do not integrate those capabilities within practical applications, focusing instead on evaluating the capabilities of LLMs in isolation. In our work, we showed how using LLMs as a component within a framework can transform large corpora of scientific literature into a thematic organization of subtopics. Each subtopic can then be expanded by using an automatically constructed query to retrieve additional relevant documents.

7 Discussion

We demonstrate how the challenge of navigating the scientific literature when embarking on a new field can be facilitated by an LLM-aided process, which we call Knowledge Navigator.

The Knowledge Navigator operates on a corpus of documents which enables a more holistic understanding and organization of knowledge within the domain. The effectiveness of our framework demonstrates the potential of the bottom-up approach in other settings where LLMs are tasked with extracting insight from a collection of items.

In addition, we believe that future work could use outputs from frameworks like Knowledge Navigator in prompts for other systems or in planning tasks for agents. For example in the retrieval-augmented generation (RAG) settings, where structured data boosts the performance and utility of LLMs in various applications.

Limitations

Knowledge Navigator demonstrates promising results in organizing and structuring scientific knowledge, offering a potential solution to the challenges of information overload in exploratory search. However, like any system, it has limitations that can be addressed in future work:

Corpus Quality and Recall. The system's performance is inherently dependent on the quality of the retrieved corpus. Suboptimal retrieval can still impact the system's output, even with the subtopic filtering mechanism in place. This limitation highlights the importance of further refining the retrieval process to improve recall and ensure the inclusion of all relevant documents.

Document Assignment. Although Knowledge Navigator utilizes soft clustering to potentially assign documents to multiple subtopics, this approach is not exhaustive due to the limitations of clustering algorithms and the representation space. Exploring alternative assignment strategies could enhance the system's ability to represent complex relationships between documents and subtopics.

User Interface and Experience. Our work primarily focuses on the technological and system design aspects of information organization. The development of a user interface (UI) that leverages Knowledge Navigator's capabilities and optimizes the user experience is crucial for its practical application.

Acknowledgements

We would like to thank SerpAPI for generously granting us credits for our evaluations.
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 802774 (iEXTRACT).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*, pages 420–434. Springer.

Douglas R. Cutting, Jan O. Pedersen, David R Karger, and John W. Tukey. 1992. Scatter/gather: a cluster-based approach to browsing large document collections. *ACM SIGIR Forum*, 51:148 – 159.

Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50.

Pasi Fränti and Radu Mariescu-Istodor. 2023. Soft precision and recall. *Pattern Recognition Letters*, 167:115–121.

Xuemei Gong, Weimao Ke, and Ritu Khare. 2012. Studying scatter/gather browsing for web search. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–4.

Cassandra J Hatzipantelis and David E Olson. 2024. The effects of psychedelics on neuronal physiology. *Annual Review of Physiology*, 86(1):27–47.

Marti A Hearst. 1999. The use of categories and clusters for organizing retrieval results. In *Natural language information retrieval*, pages 333–374. Springer.

Marti A. Hearst. 2006. Clustering versus faceted categories for information exploration. *Commun. ACM*, 49(4):59–61.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Mika Käki. 2005. Findex: search result categories help users when document ranking fails. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 131–140.

Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46.

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

Lokman I Meho and Helen R Tibbo. 2003. Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American society for Information Science and Technology*, 54(6):570–587.

Y. R. Nedumov and S. D. Kuznetsov. 2019. Exploratory search for scientific articles. *Programming and Computer Software*, 45(7):405–416.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Stanislaw Osinski and Dawid Weiss. 2005. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20:48–54.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

Peter Pirolli, Patricia Schank, Marti Hearst, and Christine Diehl. 1996. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 213–220.

Peter Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.

Ayah Soufan, Ian Ruthven, and Leif Azzopardi. 2022. Searching the literature: An analysis of an exploratory search task. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, page 146–157, Regensburg Germany. ACM.

Chyrine Tahri, Aurore Bochnakian, Patrick Haouat, and Xavier Tannier. 2023. Transitioning from benchmarks to a real-world case of information-seeking in scientific publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1066–1076, Toronto, Canada. Association for Computational Linguistics.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large language models enable few-shot clustering. *arXiv preprint arXiv:2307.00524*.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1).

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-driven explainable clustering via language descriptions. *arXiv preprint arXiv:2305.13749*.

Ryen W White and Resa A Roth. 2009. *Exploratory search: Beyond the query-response paradigm*. 3. Morgan & Claypool Publishers.

Meng Yuan, Justin Zobel, and Pauline Lin. 2022. Measurement of clustering effectiveness for document collections. *Information Retrieval Journal*, 25(3):239–268.

Oren Zamir and Oren Etzioni. 1999. Grouper: A dynamic clustering interface to web search results. *Comput. Networks*, 31:1361–1374.

Yan Zhang, Ramona Broussard, Weimao Ke, and Xuemei Gong. 2014. Evaluation of a scatter/gather interface for supporting distinct health information search tasks. *Journal of the Association for Information Science and Technology*, 65(5):1028–1041.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. Clusterllm: Large language models as a guide for text clustering. *ArXiv*, abs/2305.14871.

A Knowledge Navigator interface simulation

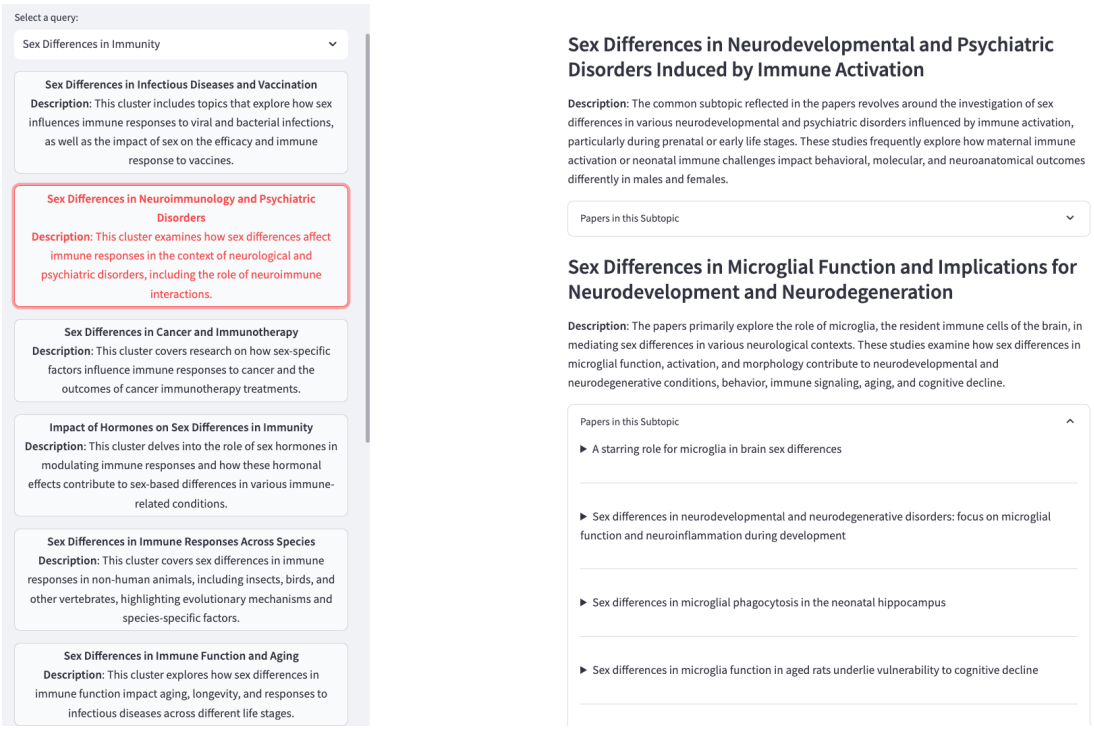
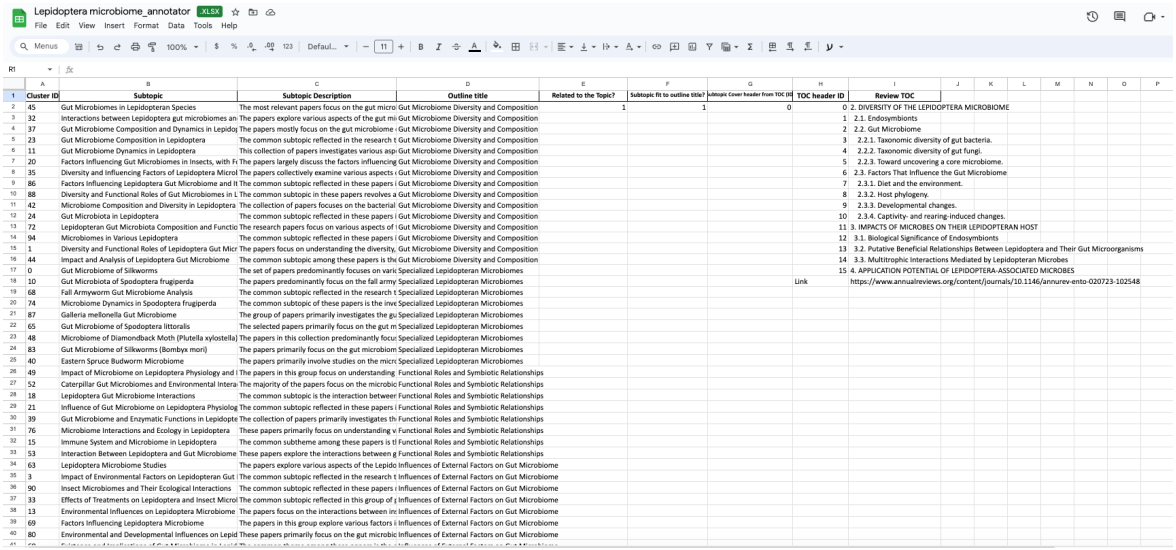


Figure 4: Knowledge Navigator implementation over a Streamlit web application for demonstration

B Expert Annotation



Cluster ID	Subtopic	Subtopic Description	Outline title	Related to the Topic?	Subtopic fit to outline title?	Subtopic Cover header from TOC (0/	TOC header ID	Review TOC
45	Gut Microbiomes in Lepidopteran Species	The most relevant papers focus on the gut microl	Gut Microbiome Diversity and Composition	1	1	0	0	2. DIVERSITY OF THE LEPIDOPTERA MICROBIOME
32	Interactions between Lepidoptera gut microbiomes an	The papers explore various aspects of the gut mi	Gut Microbiome Diversity and Composition				1	2.1. Endosymbionts
37	Gut Microbiome Composition and Dynamics in Lepidop	The papers mostly focus on the gut microbiome i	Gut Microbiome Diversity and Composition				2	2.2. Gut Microbiome
23	Gut Microbiome Composition in Lepidoptera	The common subtopic reflected in the research t	Gut Microbiome Diversity and Composition				3	2.2.1. Taxonomic diversity of gut bacteria.
11	Gut Microbiome Dynamics in Lepidoptera	This collection of papers investigates various asp	Gut Microbiome Diversity and Composition				4	2.2.2. Taxonomic diversity of gut fungi.
20	Factors Influencing Gut Microbiomes in Insects, with Fo	The papers largely discuss the factors influencing	Gut Microbiome Diversity and Composition				5	2.2.3. Toward uncovering a core microbiome.
35	Diversity and Influencing Factors of Lepidoptera Microl	The papers collectively examine various aspects	Gut Microbiome Diversity and Composition				6	2.3. Factors That Influence the Gut Microbiome
86	Factors Influencing Lepidoptera Gut Microbiome and It	The common subtopic reflected in these papers i	Gut Microbiome Diversity and Composition				7	2.3.1. Diet and the environment.
88	Diversity and Functional Roles of Gut Microbiomes in L	The common subtopic in these papers revolves a	Gut Microbiome Diversity and Composition				8	2.3.2. Host phylogeny.
42	Microbiome Composition and Diversity in Lepidoptera	The collection of papers focuses on the bacterial	Gut Microbiome Diversity and Composition				9	2.3.3. Developmental changes.
24	Gut Microbiota in Lepidoptera	The common subtopic reflected in these papers i	Gut Microbiome Diversity and Composition				10	2.3.4. Captivity- and rearing-induced changes.
72	Lepidopteran Gut Microbiota Composition and Functio	The research papers focus on various aspects of t	Gut Microbiome Diversity and Composition				11	3. IMPACTS OF MICROBES ON THEIR LEPIDOPTERAN HOST
94	Microbiomes in Various Lepidoptera	The common subtopic reflected in these papers i	Gut Microbiome Diversity and Composition				12	3.1. Biological Significance of Endosymbionts
1	Diversity and Functional Roles of Lepidoptera Gut Micr	The papers focus on understanding the diversity,	Gut Microbiome Diversity and Composition				13	3.2. Putative Beneficial Relationships Between Lepidoptera and Their Gut Microorganisms
44	Impact and Analysis of Lepidoptera Gut Microbiome	The common subtopic among these papers is the	Gut Microbiome Diversity and Composition				14	3.3. Multitrophic Interactions Mediated by Lepidopteran Microbes
0	Gut Microbiome of Silkworms	The set of papers predominantly focuses on varic	Specialized Lepidopteran Microbiomes				15	4. APPLICATION POTENTIAL OF LEPIDOPTERA-ASSOCIATED MICROBES
10	Gut Microbiota of Spodoptera frugiperda	The papers predominantly focus on the fall army	Specialized Lepidopteran Microbiomes				Link	https://www.annualreviews.org/content/journals/10.1146/annurev-ento-020723-102548
68	Fall Armyworm Gut Microbiome Analysis	The common subtopic reflected in the research t	Specialized Lepidopteran Microbiomes					
74	Microbiome Dynamics in Spodoptera Frugiperda	The common subtopic of these papers is the inve	Specialized Lepidopteran Microbiomes					
87	Galleria mellonella Gut Microbiome	The group of papers primarily investigates the gu	Specialized Lepidopteran Microbiomes					
65	Gut Microbiome of Spodoptera littoralis	The selected papers primarily focus on the gut m	Specialized Lepidopteran Microbiomes					
48	Microbiome of Diamondback Moth (Plutella xylostella)	The papers in this collection predominantly focus	Specialized Lepidopteran Microbiomes					
83	Gut Microbiome of Silkworms (Bombyx mori)	The papers primarily focus on the gut microbiom	Specialized Lepidopteran Microbiomes					
40	Eastern Spruce Budworm Microbiome	The papers primarily involve studies on the micr	Specialized Lepidopteran Microbiomes					
49	Impact of Microbiome on Lepidoptera Physiology and I	The papers in this group focus on understanding	Functional Roles and Symbiotic Relationships					
52	Caterpillar Gut Microbiomes and Environmental Intera	The majority of the papers focus on the microbic	Functional Roles and Symbiotic Relationships					
18	Lepidoptera Gut Microbiome Interactions	The common subtopic is the interaction between	Functional Roles and Symbiotic Relationships					
21	Influence of Gut Microbiome on Lepidoptera Physiolog	The common subtopic reflected in these papers i	Functional Roles and Symbiotic Relationships					
39	Gut Microbiome and Enzymatic Functions in Lepidopte	The collection of papers primarily investigates th	Functional Roles and Symbiotic Relationships					
76	Microbiome Interactions and Ecology in Lepidoptera	These papers primarily focus on understanding v	Functional Roles and Symbiotic Relationships					
15	Immune System and Microbiome in Lepidoptera	The common subtheme among these papers is t	Functional Roles and Symbiotic Relationships					
53	Interaction Between Lepidoptera and Gut Microbiome	These papers explore the interactions between g	Functional Roles and Symbiotic Relationships					
63	Lepidoptera Microbiome Studies	The papers explore various aspects of the Lepido	Influences of External Factors on Gut Microbiome					
3	Impact of Environmental Factors on Lepidopteran Gut	The common subtopic reflected in the research t	Influences of External Factors on Gut Microbiome					
90	Insect Microbiomes and Their Ecological Interactions	The common subtopic reflected in these papers i	Influences of External Factors on Gut Microbiome					
33	Effects of Treatments on Lepidoptera and Insect Micro	The common subtopic reflected in this group of p	Influences of External Factors on Gut Microbiome					
13	Environmental Influences on Lepidoptera Microbiome	The papers focus on the interactions between in	Influences of External Factors on Gut Microbiome					
69	Factors Influencing Lepidoptera Microbiome	The papers in this group explore various factors i	Influences of External Factors on Gut Microbiome					
80	Environmental and Developmental Influences on Lepid	These papers primarily focus on the gut microbic	Influences of External Factors on Gut Microbiome					

Figure 5: Annotation interface (Google sheet) for the expert annotator. After the training session and overview of the instructions, the annotator evaluated each topic in a separate file.

B.1 Annotator agreement

To assess the reliability of the annotation process, a second expert annotator was employed to independently annotate a subset of the data. Specifically, four out of the 20 reviews previously annotated by the primary expert were selected, encompassing a total of 307 subtopics. The annotation process for both annotators was identical, ensuring consistency and allowing for a direct comparison of their results. Agreement was assessed on three key tasks:

B.1.1 Agreement for experiments in section 5.2

Can LLMs effectively filter irrelevant subtopic clusters? The agreement between annotators regarding the relevancy of a subtopic to the review topic (title) achieved a 96% percent agreement, where agreement is defined as the percentage of times raters agree. Annotators were asked to mark (1 or 0) if a subtopic is highly relevant and explicitly discuss all the terms in the topic.

Can LLMs organize subtopics into coherent thematic categories? Thematic organization evaluation yielded another high percentage agreement of 96.5%. The annotators were asked to mark (1 or 0) if a subtopic fit the assigned theme or not. A "fit" is defined if the theme correctly describes the membership of that subtopic. For example, a theme about brain pathologies would not fit a subtopic about brain control on the motor system, even if some of the papers in it deal with pathologies, while a subtopic about brain cancer will fit. Subtopics should be fully dedicated to the theme they belong to.

B.1.2 Agreement for experiments in section 5.3

End-to-End Subtopic Coverage Evaluation For subtopic coverage, we took the set of covered headers in the original table of contents in each review and measured the Jaccard similarity, meaning that we measured the number of agreed covered headers divided by the union of all annotated covered headers. We got an average of 95.4% overlap coverage. This means that, on average, annotators agreed on most of the covered headers in the reviews.

These results imply that, unlike subjective annotation, this type of annotation is based on a common understanding of scientific terms and therefore achieves high agreement.

C Prompts and system implementation details

C.1 Cluster Reader prompt

```
# Task Overview:
You are provided with a general topic and a set of scientific papers retrieved
by a lexical search system using this topic as a query. Your task is to analyze
how the papers relate to the topic and categorize their relevance.

# Instructions:

Evaluate Relevance: Determine if the papers are directly related to the research
topic.

- If they are not related to the research domain or do not address the topic
directly, mark them as "NOT RELATED."
- If they are a genuine subtopic of the main topic, mark them as "RELATED."
- If the papers would not be relevant to a user searching for the main topic,
consider them not related.
- IF the papers do not address an explicit relation to the topic, consider them
not related.

# Output Requirements:
Output should be a json with the following fields:
Description: Write a summary describing the common subtopic reflected in the
research theme of the papers in the group in relation to the Topic.
Subtopic: Give a title for the group of papers that represents a meaningful
subtopic of the Topic.
Relatedness: Rate the relatedness on a scale from 1 to 5, where 1 means not
relevant at all, and 5 indicates the papers deal directly with the topic.
Is Related: State whether the papers are "RELATED" or "NOT RELATED" based on
their relevance to the original topic.
- Write nothing else

Topic: {query}
Papers: {papers_list}
```

C.2 Thematic Organization

```
You are given a nested dictionary where each key is a subtopic_id and the value
is a dictionary of subtopics of the topic {query}. Reflect on the subtopics
and their descriptions and define clusters of topics that group the subtopics
into meaningful research clusters. Create the clusters as an outline where
each cluster is a foundational chapter about {query}. Those clusters will be
used by a user to navigate between different domains of his research topic.
Give each topic a clear label and describe the subtopics that the cluster is
dealing with. Output must be in json. Do not leave any subtopic without a cluster.

## Output
- Output a json object with:
- clusters: list of dictionaries with digits from '1' to 'N' of "cluster_ids",
"cluster_title" and "description"
- subtopics: dictionary with the subtopic_id as a field and the appropriate
cluster id as a key for each subtopic in the input.

Subtopic dictionary: {subtopic_and_cluster_ids}
```

C.3 Subtopic Expander

> You are tasked with identifying keywords or terms from a list of scientific paper titles and abstracts specifically relevant to the subtopic {subtopic_title}. These keywords will be used to generate a query to retrieve documents about this specific subtopic. It's crucial to focus only on this particular aspect of {query} research and not on the general topic.
> Here is the list of titles and abstracts:
> Query Description: {subtopic_description}
> <titles_and_abstracts>{papers}</titles_and_abstracts>
> Please follow these steps to complete the task: 1. Carefully read through each title and abstract.
> 2. As you read, identify words or short phrases that are specifically related to subtopic.
> 3. Pay special attention to recurring terms or concepts across multiple papers, as these are likely to be particularly relevant.
> 4. Avoid selecting overly general terms related to {query}. Focus on terms that specifically relate to the {subtopic_title} aspect.
> 5. After analyzing all the titles and abstracts, compile a list of the most relevant and frequently occurring keywords or phrases.
> 6. Present your final list of keywords in order of relevance, with the most important or frequently occurring terms first. Include this list within <keywords> tags.
> 7. Provide a brief explanation for why you chose each keyword, highlighting its relevance to {subtopic_title}. Include this explanation within <justification> tags.
> Remember, the goal is to identify terms that will help retrieve documents specifically about {subtopic_title}, not about {query} in general. Your selected keywords should reflect this focused approach. "

Output:

> Subtopic Query : [Subtopic title + term_1,+...+,term_n]

D SciTOC Examples and error analysis

D.1 SciTOC example

Title: Visual Dysfunction in Diabetes

2. VISUAL DYSFUNCTION IN EARLY DIABETES
3. RETINAL NEURONAL DYSFUNCTION AND DEATH IN EARLY DIABETES
3.1. Changes in the Retinal Electroretinogram in Diabetes
3.2. Changes in Retinal Neuronal Structure in Early Diabetes
4. MECHANISMS OF RETINAL NEURONAL DYSFUNCTION IN EARLY DIABETES
4.1. Changes in Retinal Neuronal Inhibition in Early Diabetes
4.2. Changes in Retinal Neuronal Glutamate Signaling in Early Diabetes
4.3. Changes in Retinal Dopaminergic Signaling in Early Diabetes
4.4. Diabetes May Have Distinct Effects on Retinal Pathways
4.5. Potential Treatments Related to Neuronal Dysfunction

Figure 6: An example of the table of contents from the review 'Visual Dysfunction in Diabetes'. The header for the '1. Introduction' section has been removed from the benchmark.

D.2 Error analysis

To better understand cases where Knowledge Navigator did not match TOC headers, we conducted an error analysis on the 20 annotated reviews, categorizing a total of 58 errors into two groups:
***Grouping strategy mismatch* (50% of errors)**: These cases represent discrepancies in how subtopic boundaries are defined. The Knowledge Navigator might include a subtopic within a broader category, while the review paper presents it as a standalone header. This highlights nuanced differences in how our system and human reviewers conceptualize and organize information.
***Misses* (50% of errors)**: These are relevant subtopics expected to appear in the system's output but were not found. This can occur for several reasons: relevant papers might be dispersed across multiple clusters, diminishing their individual impact; the initial corpus might lack sufficient coverage of the subtopic; or, in one isolated instance, a potentially matching subtopic was incorrectly filtered out by the subtopic filtering process.

D.3 Header types removed from evaluation

Header Type	Query	Header Example
Introduction type headers	Tissue Immunity in the Bladder	2.1. Anatomy
Introduction type headers	Bacteriophages in the Dairy Industry	3. THE PHAGE PROBLEM
Subjective header	Schwann Cells	4.1. Pathology Due to Defects in Canonical Functions
Subjective header	APOBEC3 in Human Papillomavirus Infection and Oncogenesis	4. CRITICAL GAPS IN OUR UNDERSTANDING OF APOBEC3 IN VIRUS-INDUCED CANCERS

Table 5: Non Valid headers from SciTOC removed from the evaluation

Table 6: Example of Knowledge Navigator output for the review Endocrine Disorders and COVID-19

Themes	Subtopic Title and Description
Impact of COVID-19 on General Endocrine Health	**Impact of COVID-19 on Metabolic and Endocrine Health** The common subtopic in these papers revolves around the intersection of COVID-19 and endocrine/metabolic disorders. The papers specifically address how metabolic syndrome, diabetes, and other related endocrine dysfunctions affect susceptibility to COVID-19, the severity of the disease, and the clinical management of these patients during the pandemic.
	Impact of COVID-19 on Adrenal Insufficiency and Glucocorticoid-related Endocrine Disorders The papers collectively discuss the intersection of endocrine disorders, particularly adrenal insufficiency and glucocorticoid-related diseases, with COVID-19. They examine the outcomes, management strategies, risk factors, and potential new onset of endocrine disorders in COVID-19 patients.
Thyroid Disorders and COVID-19	**Thyroid Disorders Post COVID-19 Vaccination** The papers predominantly discuss various thyroid disorders such as thyroiditis, thyrotoxicosis, and Graves' disease occurring after COVID-19 vaccination. They present case reports, studies, and reviews examining the potential link between COVID-19 vaccines and the onset or exacerbation of these endocrine conditions.
	Thyroid Dysfunction in COVID-19 The common subtopic in these papers is the prevalence, impact, and outcomes associated with thyroid dysfunctions in patients who have contracted COVID-19. They explore various dimensions including the changes in thyroid hormone levels, the association of thyroid disorders with COVID-19 severity and outcomes, and potential mechanisms linking thyroid function with the disease.

E Subtopic Expander experiments

E.1 Expert Relevancy Annotation and LLM-Judge

We randomly sampled 13 subtopics out of the pool, and for each, we sampled 10 papers out of 100 to represent papers from the entire rank distribution, ending with 130 scientific papers. For each paper, the expert judged a score from 0 to 2, where "0" means the paper's relevancy to the subtopic is marginal, and "2" means the paper is focused on the subtopic. The annotation instructions were identical to those given to the LLM (see C.3). We then let an LLM judge each paper in the same manner and evaluate their agreement. Since in some cases the degree of relevancy can be subjective between "2" and "1," we binarized the scores for "relevant" [2,1] and "not relevant" [0], similar to other recent studies on LLM relevancy judgment (Faggioli et al., 2023; Thomas et al., 2024). We found that the binary agreement between the expert and the LLM reaches 87% with a Cohen's kappa of 0.63, which is on par with other strong LLM relevancy judgment methods on TREC benchmarks (Thomas et al., 2024).