TOPICAL: TOPIC Pages AutomagicaLly

John Giorgi[1,2,3*] **Amanpreet Singh**[4] **Doug Downey**[4,5]
Sergey Feldman[4] **Lucy Lu Wang**[4,6]

[1]University of Toronto [2]Terrence Donnelly Centre [3]Vector Institute for AI
[4]Allen Institute for AI [5]Northwestern University [6]University of Washington

john.giorgi@utoronto.ca, lucylw@uw.edu, {sergey, dougd}@allenai.org

Abstract

Topic pages aggregate useful information about an entity or concept into a single succinct and accessible article. Automated creation of topic pages would enable their rapid curation as information resources, providing an alternative to traditional web search. While most prior work has focused on generating topic pages about biographical entities, in this work, we develop a completely automated process to generate high-quality topic pages for scientific entities, with a focus on biomedical concepts. We release TOPICAL, a web app and associated open-source code, comprising a model pipeline combining retrieval, clustering, and prompting, that makes it easy for anyone to generate topic pages for a wide variety of biomedical entities on demand. In a human evaluation of 150 diverse topic pages generated using TOPICAL, we find that the vast majority were considered relevant, accurate, and coherent, with correct supporting citations. We make all code publicly available and host a free-to-use web app at: https://s2-topical.apps.allenai.org.

1 Introduction

The automatic generation of topic pages is a longstanding goal of the NLP community (Balasubramanian and Cucerzan, 2009, 2010a,b; Pochampally et al., 2021). In contrast to web search results—displayed as ranked lists of hyperlinks with short text snippets across many pages—topic pages aggregate useful information about various aspects of an entity or concept in a single, concise location. *Scientific* topic pages (Wodak et al., 2012; Azarbonyad et al., 2023) apply this thinking to scientific concepts by aggregating information from the primary literature to produce succinct and accessible summaries useful to both experts and non-experts alike (Figure 1). Among other things, high-quality scientific topic pages hold the promise of:

Obesity Paradox

The Obesity Paradox refers to the counterintuitive observation that overweight and obese individuals may have better survival rates in certain chronic diseases compared to their normal-weight counterparts (29852198).

The Obesity Paradox has been observed in a variety of chronic diseases including heart failure, coronary artery disease, atrial fibrillation, stroke, and even certain types of cancer (29852198, 32124408, 35087875, 27475805, 33160753). This phenomenon has been associated with improved survival rates, particularly in overweight and class I obesity, and less pronounced in more severe or morbidly obese populations (29981771). However, the Obesity Paradox remains controversial due to potential confounding factors such as the crudeness of Body Mass Index (BMI) as an obesity measure, retrospective nature of most studies, and differences in co-morbid conditions and disease characteristics (32124408, 27475805). Furthermore, recent studies suggest that cardiovascular fitness, rather than weight loss alone, influences the relationship between obesity and mortality in those with established cardiovascular diseases (36481212).

Future research should focus on addressing these methodological concerns and exploring the potential biological mechanisms underlying the Obesity Paradox (27475805).

Figure 1: Example of a scientific topic page generated by our system. Citations are provided as hyperlinks to PubMed articles and denoted by their PMID. The topic page is divided into the definition statement, main content, and future directions and open research questions.

1. **Helping manage the torrent of scientific literature**. A staggering amount of scientific information is published daily. In biomedicine alone, nearly 4,000 papers (>2 per minute) are deposited in PubMed or bioRxiv each day, leading to a general state of "information overload" (Landhuis, 2016; Hope et al., 2023). Automatically generated topic pages allow researchers to quickly familiarize themselves with an area and its most active research directions, while citations to source articles provide an entry-point into the literature for in-depth exploration.[1]
2. **Improving the accessibility of scientific texts**. Encyclopedic resources like Wikipedia contain

[*] Work performed during internship at AI2

[1] Topic pages generated by our system provide citations to highly relevant primary literature. See §3.2 for details.

descriptions for a small fraction of scientific concepts (King et al., 2020). Therefore, non-expert readers may turn to the primary literature for information (August et al., 2022), e.g., a patient or caregiver wishing to learn about a new drug or rare disease. However, most scientific text assumes extensive background knowledge that a non-expert reader is unlikely to possess (Portenoy et al., 2021; Murthy et al., 2022). Automatically generated topic pages hold the promise of improving the accessibility of scientific texts, both by providing an *alternative* to the primary literature and by serving as a *resource* to help fill in the gaps in a reader's background knowledge.

In this work, we develop a fully automated process leveraging large language models (LLMs) to generate high-quality scientific topic pages, with a focus on biomedical topics (§3). Our solution is available as an easy-to-use and publicly available web app (§4), and associated source code.[2] We validate the quality of TOPICAL via extensive human evaluation on 150 diverse biomedical terms from the MeSH[3] hierarchy (§5) and find that the vast majority of topic pages are rated as relevant, accurate, and coherent, with correct citations to primary sources (§6).

2 Related Work

Topic page generation Topic page generation is usually framed as a topic-focused, open-domain multi-document summarization (MDS) task (Giorgi et al., 2023). Most prior work is concerned with generating Wikipedia-like pages for general-domain entities and concepts (often biographical in nature). Early work clustered the web search query logs for an entity of interest to determine its various aspects, used each aspect cluster to retrieve and rank relevant sentences, and then re-organized the retrieved sentences for coherence to produce a bullet-list style topic page (Balasubramanian and Cucerzan, 2009, 2010a,b).

More recent work—also focused on biographical entities—first templates the topic page by copying common section headings from Wikipedia pages for related topics and trains a supervised model to select the text content for each section. An unsupervised component then creates topic-specific sections, and several post-processing steps are applied to reduce redundancy and improve coherence (Pochampally et al., 2021). In contrast, our work focuses on topics of scientific interest, does not try to match a Wikipedia-like structure, and generates topic pages in a *abstractive* fashion.

[2] https://github.com/allenai/TOPICAL

[3] Medical Subject Headings (MeSH) is a hierarchical vocabulary used to index articles and books in the life sciences.

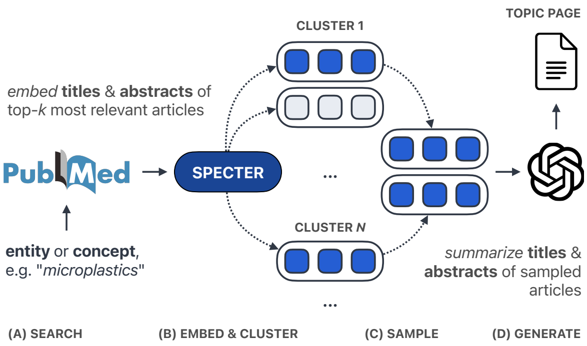

Figure 2: Overview of TOPICAL. Given a biomedical entity, we query PubMed for relevant literature (**A**). The titles and abstracts of the results are embedded with SPECTER (Singh et al., 2023) and clustered based on semantic similarity (**B**). We sample titles and abstracts from the clusters (**C**) and feed them to GPT-4 (OpenAI, 2023), alongside publication metadata and natural language instructions, to generate the topic page (**D**).

Scientific topic pages Azarbonyad et al. (2023) investigate generating scientific topic pages at scale; however, they do not synthesize a summary but focus rather on extracting a definition statement verbatim, alongside "mention snippets" and related concepts. In contrast, we attempt to synthesize more comprehensive topic pages, including a definition statement and content about the entity's main and future research directions. King et al. (2022) introduce a Scientific Concept Description task with similar motivation to our work, but focus on earlier, smaller generative models for describing computer science concepts, and find the systems to hallucinate relatively frequently. WikiCrow, based on PaperQA (L'ala et al., 2023), provides scientific topic pages generated by an LLM-based system for human protein-coding genes. In contrast to our approach, their publicly available demo is limited to 15,616 pre-generated topic pages and does not allow a user to generate topic pages for a new entity of interest on demand.[4]

3 Approach

Our approach follows a retrieval-augmented generation (RAG) setup (Guu et al., 2020; Lewis et al., 2020; Petroni et al., 2021; Izacard et al., 2022). A

[4] https://www.futurehouse.org/wikicrow

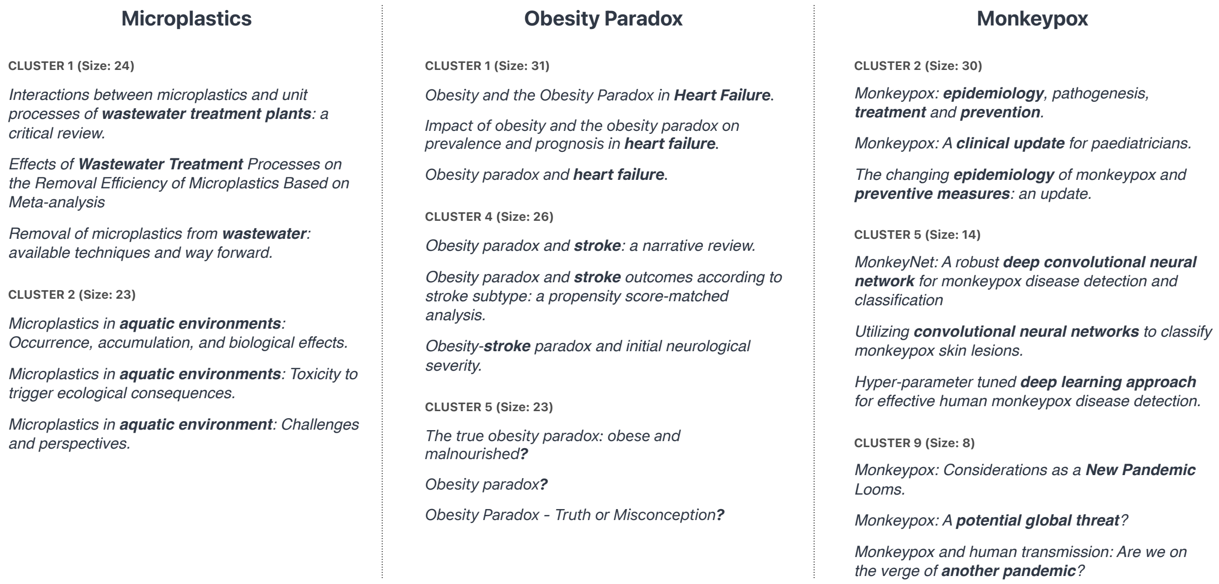

Figure 3: Example clusters. Three titles from a selection of clusters for each concept are shown. **Emphasis** ours.

large body of literature (up to 10k papers) is retrieved for a given entity (§3.1) and fed to a LLM alongside publication metadata and instructions (§3.3). Because the amount of retrieved literature is often many times larger than the LLM's maximum context size, we design a clustering step to loosely group the literature into areas of study and sample from these clusters for input (§3.2). During prompting, the model is instructed to provide in-line citations for all claims by outputting one or more PubMed IDs (PMIDs). See Figure 2 for an overview.

3.1 Querying PubMed

The generation of each topic page begins with a user-provided biomedical entity or concept. This entity is expected to be covered by papers indexed in PubMed,[5] a free search engine that indexes over 36 million papers on life science and biomedical topics. TOPICAL, our system, leverages the Entrez ESearch API (Kans, 2023) to query PubMed and supports the full syntax of the PubMed Advanced Search Builder; however, simply inputting the entity or concept verbatim is often sufficient, e.g. "*microplastic*," as the ESearch API will apply 'automatic term mapping' (ATM)[6] to this query to include, among other things, matching MeSH descriptors and pluralization (e.g. "*microplastics*"). We then download the titles and abstracts of the top 10,000 most relevant papers returned by ESearch.

3.2 Clustering and sampling the literature

The amount of retrieved literature is usually many times the maximum context size of the LLM. Therefore, we first cluster titles & abstracts by semantic similarity to identify major areas of study, then sample from these clusters to produce a diverse set of inputs. The steps are described below:

Embedding Titles and abstracts are jointly embedded using the SPECTER2 PRX model (Singh et al., 2023), a text encoder specifically designed for producing highly-quality representations of scientific text from a paper's title and abstract. We formatted each input as: "`{title} [SEP] {abstract}`".

Clustering We apply a clustering algorithm which identifies 'communities': clusters of embeddings of a minimum size with a pairwise cosine similarity greater than or equal to some threshold.[7] We set the similarity threshold to `0.96` and the minimum cluster size to 5. In degenerate cases where fewer than 2 clusters are identified, we iteratively reduce the similarity threshold by `0.02`, stopping when at least 2 clusters are identified or the threshold falls below `0.90`—in which case we skip the clustering step. See Figure 3 for examples of clusters produced by this process.

Sampling We sample as many titles and abstracts as will fit in the prompt to the LLM. If the number of papers returned in the search step is 100 or

[5] https://pubmed.ncbi.nlm.nih.gov/
[6] https://pubmed.ncbi.nlm.nih.gov/help/#automatic-term-mapping
[7] https://www.sbert.net/examples/applications/clustering/README.html#fast-clustering

Algorithm 1 Sampling Procedure for Papers

Require: Collection of clustered titles + abstracts, $\mathcal{C}$
Require: Maximum number of input tokens, $T_{\max}$
$\mathcal{C} \leftarrow \text{sorted}(\mathcal{C})$ ▷ By descending cluster size
Initialize $\mathcal{S} \leftarrow \emptyset$ ▷ Sampled papers
$t \leftarrow 0$ ▷ Current token count
for each C_i in $\mathcal{C}$ **do**
    $c \leftarrow$ centroid of C_i
    if $t + |c| \leq T_{\max}$ **then**
        Append c to $\mathcal{S}$
        $t \leftarrow t + |c|$ ▷ $|c|$ is the number of tokens in c
    end if
end for
while $t < T_{\max}$ and there exist unsampled papers in $\mathcal{C}$ **do**
    Sample a paper p from $\mathcal{C}$ with a probability $\propto \sqrt{|C_i|}$
    if $t + |p| \leq T_{\max}$ **then**
        Append p to $\mathcal{S}$
        $t \leftarrow t + |p|$ ▷ $|p|$ is the number of tokens in p
    end if
end while
Ensure: Return $\mathcal{S}$ as a list of lists (outer: unique clusters, inner: papers from the cluster)

less, or no clusters were identified in the clustering step, we randomly sample papers for inclusion. Otherwise, we do the following: first, sort clusters by decreasing size. Then, select each *centroid* for inclusion, starting with the largest cluster and continuing until the centroids of all clusters have been selected or the model's maximum input size has been reached. If all centroids have been selected and the model's maximum input tokens are not exhausted, we sample from the remaining clusters with a probability proportional to the square root of the cluster size (see Algorithm 1 for details).[8] This sampling strategy is motivated by the idea that we should aim to capture as many and as diverse areas of study for a concept as possible (hence the selection of centroids) while favouring more commonly studied subtopics (hence the weighted sampling).

3.3 Generating the topic page

We chose GPT-4[9] as the LLM due to its state-of-the-art performance across many text generation tasks (OpenAI, 2023). We designed a prompt including natural language instructions, publication metadata, and the sampled titles and abstracts. The prompt is broken into system and user roles (truncated example in Figure 4). In the system role, we provide instructions about the task and what constitutes a good topic page. The user role provides instruction about what the model will receive as input, followed by a description of how to cite its sources. We then provide information about the entity or concept, including the publications per year, total number of publications, and sampled titles and abstracts. These include a PMID and publication date and are sorted by decreasing cluster size. Finally, we provide instructions about the expected format of the topic page.

TOPICAL Prompt

System Role

You are a biomedical domain expert. Your job is to produce a high-quality, scientifically-orientated topic page for a given biomedical entity or concept grounded in the provided literature. [...]

A good scientific topic page is: [...]

Assume the target audience of this topic page will have basic scientific literacy (i.e. undergraduate-level biology). [...]

User Role

INSTRUCTIONS

I will provide you with a biomedical entity or concept, titles and abstracts that mention this entity. [...]

HOW TO CITE YOUR CLAIMS

Every scientific claim in the topic page should be followed by an in-line citation to PubMed using the provided PMIDs. [...]

ENTITY OR CONCEPT

Canonicalized entity name: Microplastics
Publications per year: 2006: 1, 2007: 1, [...] 2023: 2288
Total number of publications: 8217
Supporting literature:

Cluster 1
PMID: 37079238 **PubDate**: May 2023 **Title**: [...] **Abstract**: [...]
PMID: 35301580 **PubDate**: Mar. 2023 **Title**: [...] **Abstract**: [...]
[...]

Cluster N
PMID: 30036839 **PubDate**: Nov 2018 **Title**: [...] **Abstract**: [...]
[...]

TOPIC PAGE

Now, generate the scientific topic page section by section following the instructions below.

First, provide a short textbook or Wikipedia-like description of the entity that is easy to understand for a non-expert audience (1 sentence max).

Next, produce the main content of the topic page (6 sentences max). Summarize the main reasons for this entities notability and interest to science. [...]

Finish by commenting on any open questions or future research directions mentioned in the supporting literature [...]
(1 sentence max).

Figure 4: Truncated example prompt. The prompt is divided into system and user roles. In the user role, we provide instructions about the input, how to cite a claim, details about the entity or concept like publication metadata, the sampled literature, and guidance about the expected sections and lengths for the topic page. Emphasis is provided for visualization purposes only.

The model is instructed to produce three sections: a **definition statement** (1-2 sentences), **main content** (5-8 sentences) and a concluding remark about **future research directions and open questions** (1 sentence). We model the components of our target

[8] https://en.wikipedia.org/wiki/Square_root_biased_sampling
[9] Specifically, the 06/13/2023 snapshot, "gpt-4-0613"

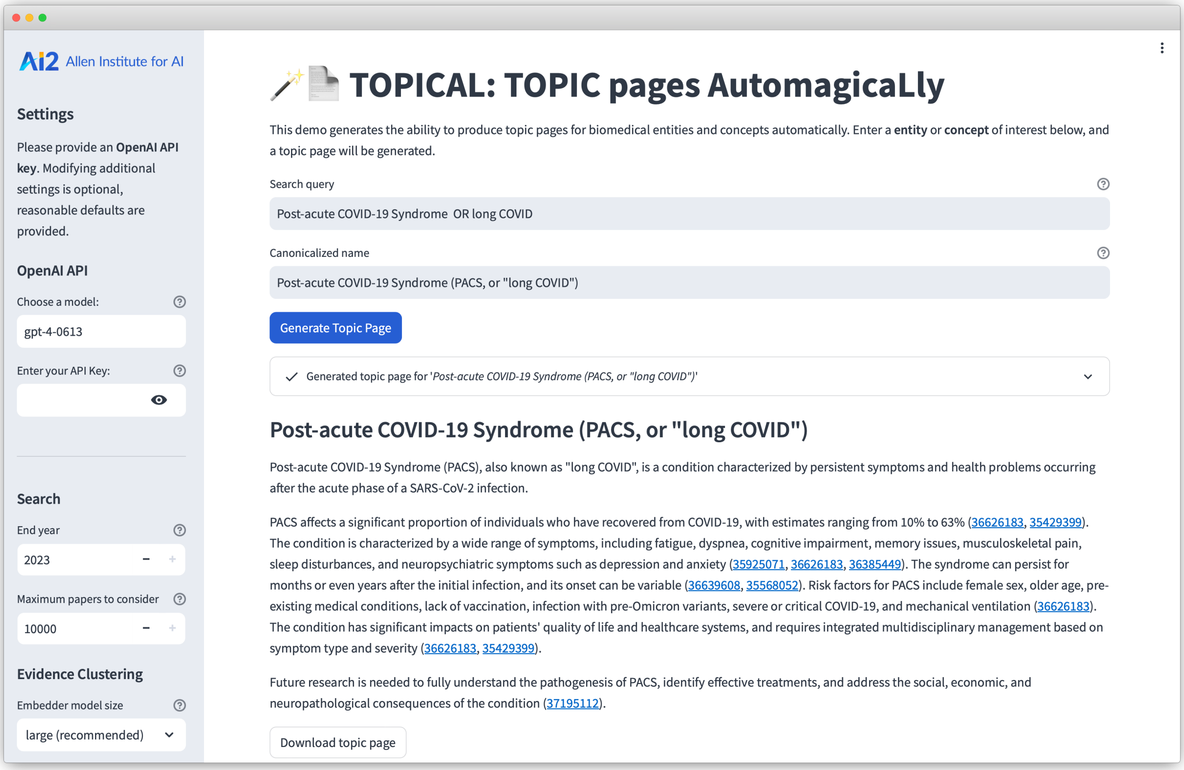

Figure 5: TOPICAL web app. Given a search query for a biomedical entity or concept of interest and a canonicalized name, it automatically generates a topic page for the concept. An expandable section provides additional information, like a histogram of publication dates for the query and the number of clusters identified.

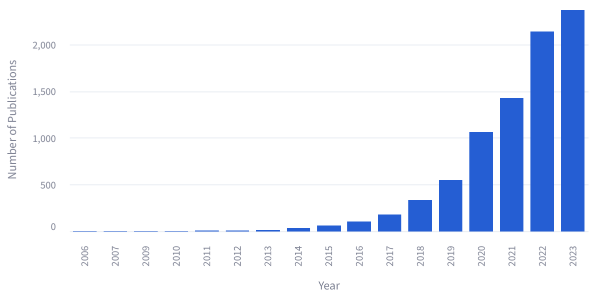

Figure 6: Example publications per year histogram displayed to users for the entity: "*Microplastics*".

topic pages based on the structure of existing scientific topic pages and the information researchers are likely to seek from a topical review. Curated topic pages typically begin with definitions,[10] so we also begin by generating a definition statement. Per the PRISMA guidelines for systematic reviews (Page et al., 2020), a primary goal of reviews is to provide "syntheses of the state of knowledge in a field, from which future research priorities can be identified"; from this goal, we derive the main content, which summarizes the main directions of research, and future research directions.

We set `temperature` to 0.0, `max_tokens` to 512, (the maximum tokens to generate for the topic page), and kept all other hyperparmeters of the OpenAI API at default values.[11] The model's maximum context size is 8,192 tokens, which is approximately enough for the prompt instructions and 16 abstracts. To fit more abstracts into the prompt, we take only the first three and last two sentences of each, joining them with a "`[TRUNCATE]`" token. These sentences tend to be rich in the type of content expected in a topic page, e.g., definition-like content, conclusions, major findings, and future directions.

4 TOPICAL Web App

TOPICAL is available as a web app (see Figure 5 for an overview). The web app can be run locally as a standalone python package but is also publicly available at https://s2-topical.apps.allenai.org. A user first inputs a PubMed search query, which supports the full syntax of the PubMed Advanced Search Builder (see Appendix A for details). However, in most cases, sim-

[10] e.g., https://www.sciencedirect.com/topics

[11] https://platform.openai.com/docs/api-reference/completions

ply inputting the entity or concept name directly and allowing ESearch to expand the query via automatic term mapping (ATM) works well, especially with respect to recall. A user can optionally provide alternative names for the entity, referred to as 'canonicalized names,' which are provided to the LLM as additional context. Once a user clicks the button to generate a topic page, the search, embedding, clustering, and generation steps are executed. An expandable section in the app displays progress, as well as additional information about the search, e.g., any query expansions made via ATM and a histogram of publications per year (see Figure 6). The generated topic pages can be downloaded as JSON files. A video demonstration of the system is available here: `https://youtu.be/hgnG7BnIeAY`.

5 Human evaluation

We conduct human evaluation to determine the quality of the automatically generated topic pages. The evaluation consists of two tasks, described below. All annotations were performed by three full-time, paid annotation specialists with undergraduate training spanning the biosciences, materials science, environmental science, and data science.

5.1 Annotation Task 1: Topic Page

In task 1, the goal was to evaluate the overall quality of the topic page along three facets, **relevance**, **accuracy**, and **coherence**, defined as:

- **Relevance**: whether the topic page covers only important aspects of the entity or concept; unimportant or excess information is penalized
- **Accuracy**: whether the topic page is free of obvious factual errors or contradictory information
- **Coherence**: whether sentences and sections fit together and sound natural, with little to no redundancy within or across sections

We adapt these facets and their definitions from the summarization evaluation facets used in Fabbri et al. (2021); we assess Accuracy instead of Consistency due to the infeasibility of comparing a generated topic page against all input documents. Relevance and accuracy were assessed per topic page *section* (definition statement, main content and future directions), while coherence was assessed globally. The annotation interface (see Appendix B) displayed each section of the topic page, and the annotators were provided instructions about how to evaluate topic pages along each facet. The annotation interface also provided a link to the PubMed query issued when building the topic page. Annotators were instructed to follow the link and skim a handful of abstracts to familiarize themselves with the entity or concept before evaluation.

For each facet, annotators selected from one of three options: 'not' {relevant, accurate, coherent}, 'somewhat' {relevant, accurate, coherent} or simply: {relevant, accurate, coherent}. We included a fourth option for relevance and accuracy: 'missing/invalid', in case the LLM failed to generate a particular topic page section.

5.2 Annotation Task 2: Citations

In task 2, the goal was to evaluate the relevance and sufficiency of model-provided citations. One citation from each topic page was sampled at random. Annotators were shown the citation in context and the cited article's title and abstract. They were instructed to annotate the citation as:

- **Correct**: citation is topically *relevant* (i.e., the cited article is about the target entity or concept) and provides sufficient evidence for the corresponding claim(s) in the topic page.
- **Incorrect (topically relevant)**: citation is topically relevant but does not provide sufficient evidence for the corresponding claim(s).
- **Incorrect (topically irrelevant)**: citation is topically *irrelevant*.
- **Incorrect (invalid)**: citation is not valid, e.g. the PMID does not exist or was truncated.

5.3 Choosing topics for evaluation

In order to choose a broad selection of topics for evaluation, we collected all terms added to the MeSH vocabulary in the last 10 years (01/01/2013–16/10/2023, inclusive). We only include terms with a maximum tree depth[12] of at least 7 as we found terms with a tree depth less than this tended to be overly broad and non-specific, e.g. "*Metadata*", "*Rural Nursing*", "*Infant Health*", and "*Missed Diagnosis*". The end result is 981 biomedical terms or concepts spanning a wide range of semantic types, including diseases (e.g. "*Charles Bonnet Syndrome*"), drugs (e.g. "*Modafinil*"), proteins (e.g. "*beta-Arrestin 1*), organisms (e.g "*Fallopia multiflora*"), cell types (e.g. "*Memory T Cells*") and broader concepts like "*Glycemic Load*". We subsampled from this set to produce the final list of entities for evaluation: 15 per annotator for the an-

[12] MeSH terms are organized in a polyhierarchical ontology, where more specific terms exist deeper in the tree.

Table 1: Results of human evaluation for annotation task 1. Total ratings for each facet and label are shown, along with agreement percentage. Two annotators rated 100 pages each (with 50% overlap). Each facet for each section was rated on an ordinal scale: "not", "somewhat", or "(yes)" relevant/accurate/coherent.

	Definition		Main content		Future directions		
Rating	relevant	accurate	relevant	accurate	relevant	accurate	coherent
missing/invalid	0	0	0	0	0	0	–
not	0	0	1	0	3	0	0
somewhat	4	1	7	0	15	0	15
yes	196	199	192	200	182	200	185
Percent agreement	94	98	94	100	88	100	82

Table 2: Results of human evaluation for annotation task 2. Total ratings per label are shown, along with agreement percentage. Each annotator rated 100 citations, with 50% overlap between annotators.

Rating	Number of Ratings
Incorrect (invalid)	0
Incorrect (topically irrelevant)	2
Incorrect (topically relevant)	32
Correct	166
Percent agreement	88

notation pilots (with 100% overlap) and 100 per annotator for the final evaluation (50% overlap).

6 Results

We find that the majority of topic pages are rated by our annotators as relevant, accurate, and coherent (Table 1), with high inter-annotator agreement ($\geq 82\%$).[13] We note that in no case did the model fail to output a topic page with the expected three-section structure. All sections received nearly perfect ratings for accuracy. The future direction section received the lowest rating for relevancy (18/200 ratings of 'not' or 'somewhat' relevant). Examining these instances reveals that the LLM often states vague or even obvious future directions, such as: "*[...] Future research is needed to further clarify the most effective use of this drug combination in the treatment of respiratory diseases*" or "*Future research directions include further investigation into the exact mechanisms of resveratrol's action in diseases such as cancer and diabetes [...]*." We believe this reflects the inherent difficulty of identifying future research directions and open questions about a given topic. Coherence was the next lowest-rated aspect, with 15/200 ratings of 'somewhat' coherent. The most common reason for this according to the annotators, by far, was extensive use of highly-specific jargon, making the topic page difficult to read as a non-expert.

Similarly, most model-provided citations were rated as correct (Table 2) with high inter-annotator agreement ($\geq 88\%$); in no case were the citations invalid, e.g., a hallucinated PMID. Most incorrect citations were marked as 'Incorrect (topically relevant)' (32/200), denoting cases where the citation was *on-topic*, but the cited article did not provide sufficient evidence for the corresponding claim(s).

[13] We report inter-annotator agreement as the percent agreement: (fraction of cases where annotators agree) / (total number of annotations).

7 Conclusion

In this paper, we present TOPICAL, a new approach for the automatic generation of high-quality scientific topic pages that leverages large language models (LLMs) and retrieval-augmented generation (RAG). We conducted an extensive human evaluation of 150 diverse topics from the biomedical literature and our annotators rated the vast majority of generated topic pages as relevant, accurate, and coherent; and model-provided citations as correct. Promising future directions include allowing users to provide custom instructions with respect to structure, focus and length of the automatically generated topic pages, and the investigation of open-source LLMs in place of the closed-source LLM we experimented with (GPT-4). We release a publicly available web app so that others can experiment with generating topic pages for entities or concepts of interest on demand.

Limitations

Context window Due to the limited context window of GPT-4 (8192 tokens), our system only ingests a small fraction of literature for most entities or concepts. We tried to partially alleviate this through our clustering and sampling procedure, which is designed to encourage diversity in the selected literature while maintaining the representation of common research threads. A promising future direction is to explore the use of language models with significantly larger context windows, such as the recently announced GPT-4-turbo (128,000 tokens).

Unclear provenance Our evaluation is not able to determine to what degree the information in the resulting topic pages is derived from the learned weights of the language model itself, versus the retrieved literature. This is partially alleviated by requiring the language model to provide citations for all scientific claims, allowing a user to verify the information.

Unit of retrieval We do not explore retrieving information other than titles or abstracts. It is possible that retrieving information on another level of granularity, e.g. sentences or "chunks", could improve the quality of the topic pages. It is also possible that extending retrieval to the full-content of a scientific paper could further improve quality. Determining the most performant granularity for the retrieval step is an exciting future direction.

Acknowledgements

This research was enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca) and Compute Ontario (www.computeontario.ca). We thank Jenna Sparks, Erin Bransom, and Bailey Kuehl for helping to conduct human evaluations. We thank the AI2 Reviz team for assistance with hosting the web application. We thank all internal and external reviewers for their thoughtful feedback, which improved earlier drafts of this manuscript.

Author Contributions

John Giorgi conducted data processing, engineered prompts, ran experiments, and implemented the evaluation. John also contributed to project scoping and ideation and wrote the paper with feedback from others. Sergey, Doug, and Lucy were project mentors, contributing equally to project scoping and experimental design and providing core ideas and direction throughout the course of the project and paper writing. Aman made technical contributions around scaling and hosting of the demo.

References

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2022. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction*, 30:1 – 38.

Hosein Azarbonyad, Zubair Afzal, and George Tsatsaronis. 2023. Generating topic pages for scientific concepts using scientific publications. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*, page 341–349, Berlin, Heidelberg. Springer-Verlag.

Niranjan Balasubramanian and Silviu Cucerzan. 2009. Automatic generation of topic pages using query-based aspect models. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 2049–2052, New York, NY, USA. Association for Computing Machinery.

Niranjan Balasubramanian and Silviu Cucerzan. 2010a. Beyond ranked lists in web search: Aggregating web content into topic pages. *Int. J. Semantic Comput.*, 4:509–534.

Niranjan Balasubramanian and Silviu Cucerzan. 2010b. Topic pages: An alternative to the ten blue links. In *2010 IEEE Fourth International Conference on Semantic Computing*, pages 353–360.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Wang, and Arman Cohan. 2023. Open domain multi-document summarization: A comprehensive study of model brittleness under retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8177–8199, Singapore. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *ArXiv*, abs/2002.08909.

Tom Hope, Doug Downey, Daniel S Weld, Oren Etzioni, and Eric Horvitz. 2023. A computational inflection for scientific discovery. *Communications of the ACM*, 66(8):62–73.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A. Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *ArXiv*, abs/2208.03299.

Jonathan Kans. 2023. Entrez direct: e-utilities on the unix command line. In *Entrez programming utilities help [Internet]*. National Center for Biotechnology Information (US).

Daniel King, Doug Downey, and Daniel S. Weld. 2020. High-precision extraction of emerging concepts from scientific literature. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy, and Doug Downey. 2022. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 555–571, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jakub L'ala, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *ArXiv*, abs/2312.07559.

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535:457 – 458.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Sonia Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl, Sophie Johnson, Jonathan Borchardt, Daniel Weld, Tom Hope, and Doug Downey. 2022. ACCoRD: A multi-document approach to generating diverse descriptions of scientific concepts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 200–213, Abu Dhabi, UAE. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Matthew J. Page, Joanne E. McKenzie, Patrick M. M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer Marie Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie May Glanville, Jeremy M. Grimshaw, Asbjørn Hrõbjartsson, Manoj Mathew Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A Stewart, James Thomas, Andrea C. Tricco, Vivian A Welch, Penny F. Whiting, and David Moher. 2020. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, 10.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Yashaswi Pochampally, Kamalakar Karlapalem, and Navya Yarrabelly. 2021. Semi-supervised automatic generation of wikipedia articles for named entities. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(2):72–79.

Jason Portenoy, Marissa Radensky, Jevin D. West, Eric Horvitz, Daniel S. Weld, and Tom Hope. 2021. Bursting scientific filter bubbles: Boosting innovation via novel author discovery. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Shoshana J. Wodak, Daniel Mietchen, Andrew M. Collings, Robert B. Russell, and Philip E. Bourne. 2012. Topic pages: Plos computational biology meets wikipedia. *PLoS Computational Biology*, 8.

A PubMed Advanced Search Builder

TOPICAL supports the full syntax of the PubMed Advanced Search Builder. For example, to search for mentions of an entity in the title only:

```
Post-acute COVID-19 Syndrome[Title]
```

or for papers with the corresponding MeSH term:

```
Post-acute COVID-19 Syndrome[MeSH Terms]
```

Search terms can be further combined with `AND`, `OR` and `NOT` operators:

```
Post-acute COVID-19 Syndrome[Title] AND
Post-acute COVID-19 Syndrome[MeSH Terms]
```

However, in most cases, we found that simply inputting the entity or concept name directly and allowing ESearch to expand the query via automatic term mapping (ATM) works best, especially with respect to recall.

B Annotation Interface

In Figure 7, we provide a screenshot of the annotation interface built in Google Sheets used for the human evaluation. Annotators were provided the contents of the topic page segmented into the three sections (definition statement, main content, and open research questions and future directions)

C Annotation pilots

Before the full evaluation, we ran 2 pilots with 3 annotators. The annotators evaluated the same 10 topic pages in the first pilot. We used their feedback to improve the annotation guidelines and identify the main sources of inter-annotator disagreement. Most notably, task 2 originally had annotators identify all unique claims in each section of the topic page and then annotate each following the guidelines. This turned out to be overly time-intensive, and determining the specific number of claims had a very low-inter-annotator agreement. Task 2 was therefore simplified by randomly sampling one citation in the topic page and having the annotators assess its relevance and sufficiency. We then ran a second pilot on a new set of 5 topic pages to finalize the annotation guidelines and identify any remaining sources of significant annotator disagreement.

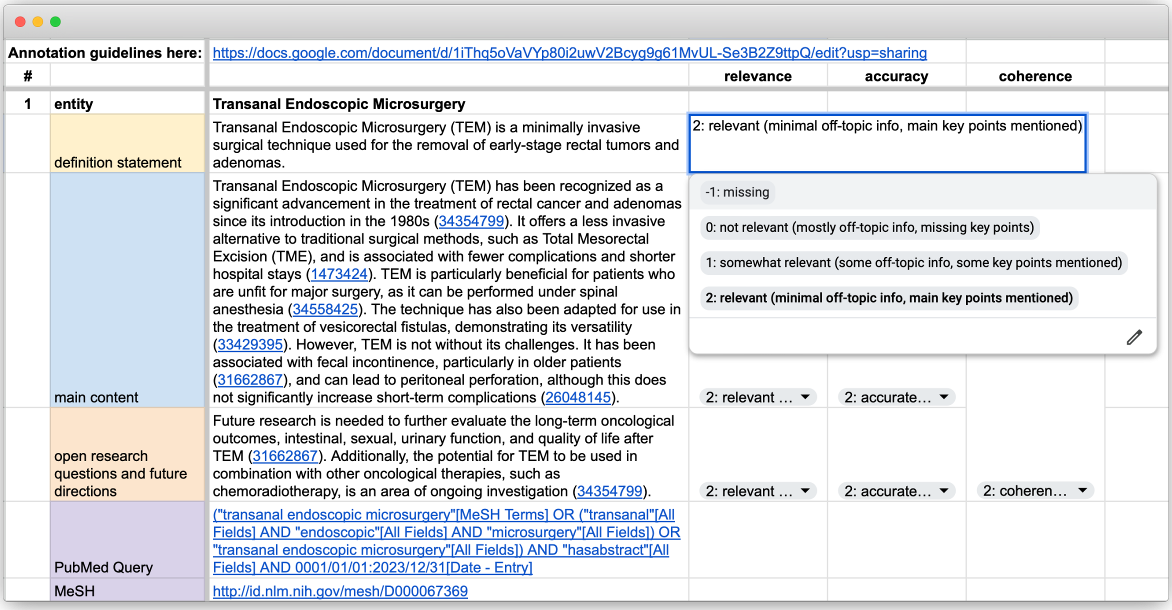

Figure 7: Annotation interface for annotation task 1.