National Research University Higher School of Economics
Faculty of Computer Science
School of Data Analysis and Artificial Intelligence

TOPICAL: TOPIC Pages AutomagicaLly

Presenters: M. Kirdin, K. Ostudina

Authors: John Giorgi, Amanpreet Singh, Doug Downey, Sergey Feldman, Lucy Lu Wang

Conference: NAACL 2024, Mexico City, Mexico.

Outline

- Problem Statement
- Methods
- Human evaluation
- Conclusion
- References

Problem Statement

The article addresses challenge of creating a tool that solves the following problems:

- First of all it is helping manage the torrent of scientific literature.
- The second one is improving the accessibility of scientific texts.

Methods

Methods

The authors propose **TOPICAL** – an open-source tool, which is also available as a web application. It, unlike many of its predecessors, generates topic pages in an *abstractive* fashion, rather than Wikipedia-like articles.

Methods

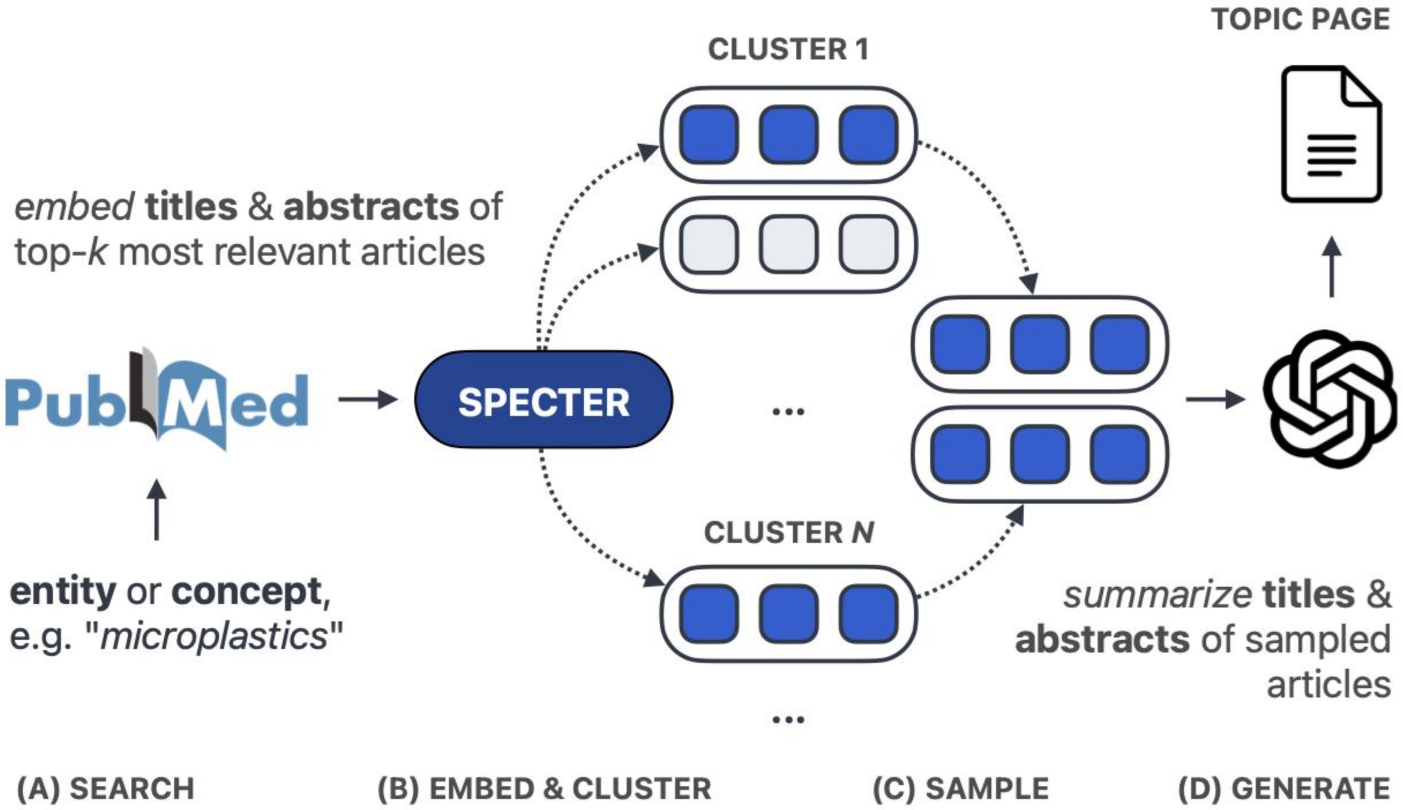
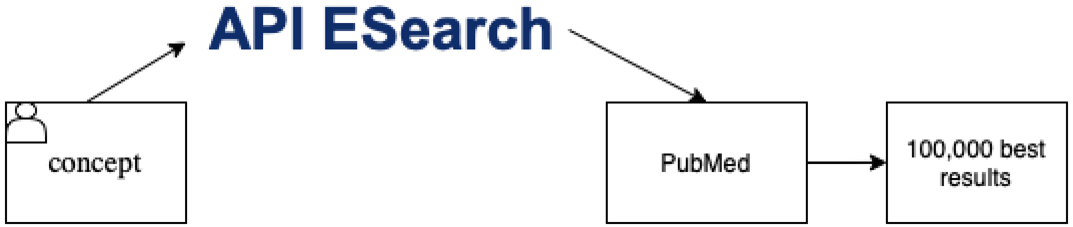

TOPIC PAGE
CLUSTER 1
embed **titles** & **abstracts** of top-*k* most relevant articles
PubMed
SPECTER
...
CLUSTER *N*
entity or **concept**, e.g. "*microplastics*"
summarize **titles** & **abstracts** of sampled articles
...
(A) SEARCH
(B) EMBED & CLUSTER
(C) SAMPLE
(D) GENERATE

Figure 2: Overview of TOPICAL

Querying PubMed

Algorithm:

1. Receive a user-provided biomedical entity or concept
2. System leverages the Entrez ESearch API to query PubMed
3. ESearch API applies automatic term mapping (ATM) to the query
4. Return 10,000 most relevant papers



API ESearch

concept

PubMed

100,000 best results

Embedding and clustering

Embedding

Titles and abstracts are jointly embedded using the SPECTER2 PRX model

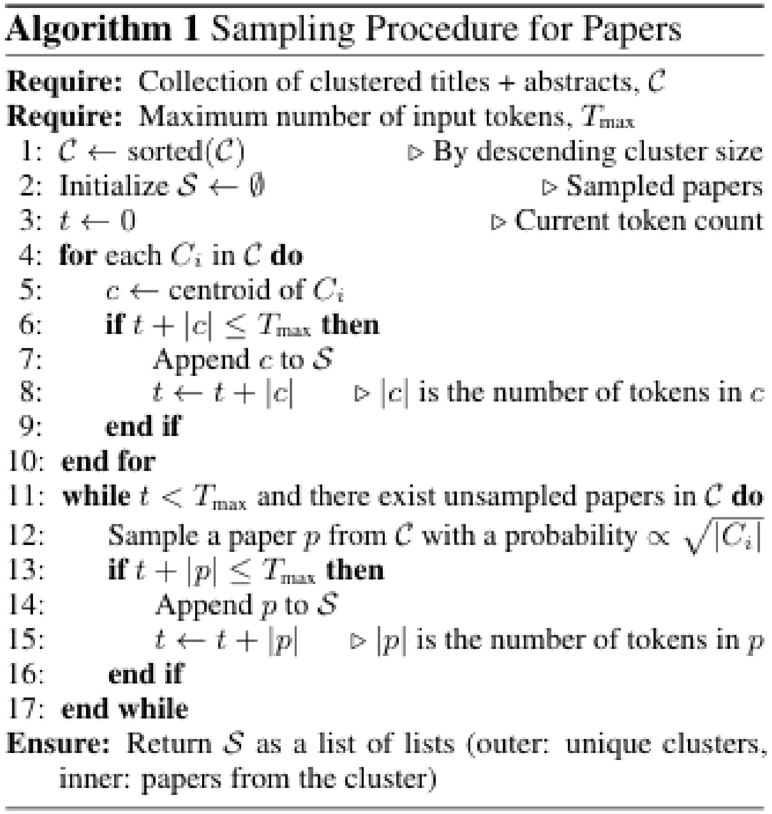
Each input are formatted as: “{title}[SEP]{abstract}”

Clustering

Clustering starts with the highest similarity threshold at $t = 0.96$. Then embeddings are clustered using the current threshold. If number of clusters is fewer than $k = 2$ clusters, the threshold is reduced by 0.02 and the algorithm is applied again. If the threshold falls below $t_{min} = 0.90$ and at least 2 clusters are found, return clusters and finish the clustering. Otherwise, skip clustering.

Sampling

Sample as many title-abstract pairs as fits in a prompt. The sample itself is comprised of cluster centroids and, if all centroids have been selected and the model’s maximum input tokens are not exhausted, samples from the remaining clusters are drawn with probabilities proportional to the square root of their cluster size.



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)
```

Generating the topic page

LLM model: GPT-4

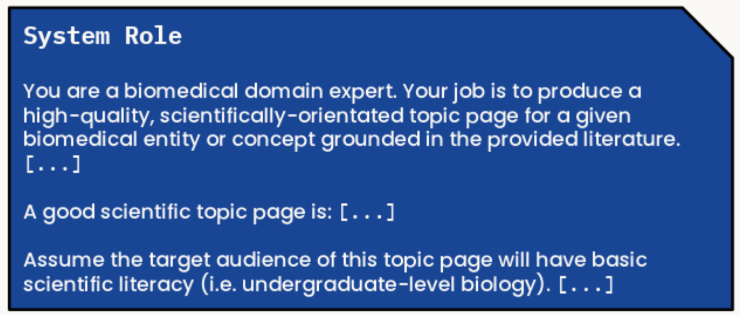
Model's input:

- natural language instructions, which contains user role
- publication metadata
- the sampled titles and abstracts

Customization:

- temperature to 0.0
- max_tokens to 512

All other hyperparameters of the OpenAI API at default values

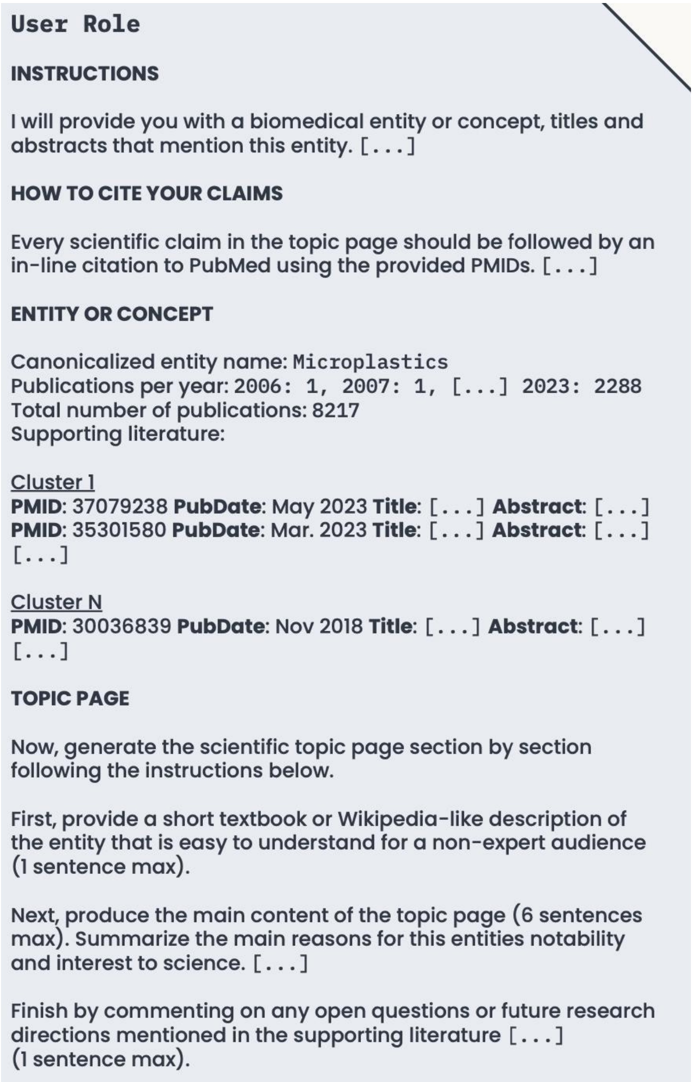


`System Role`

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

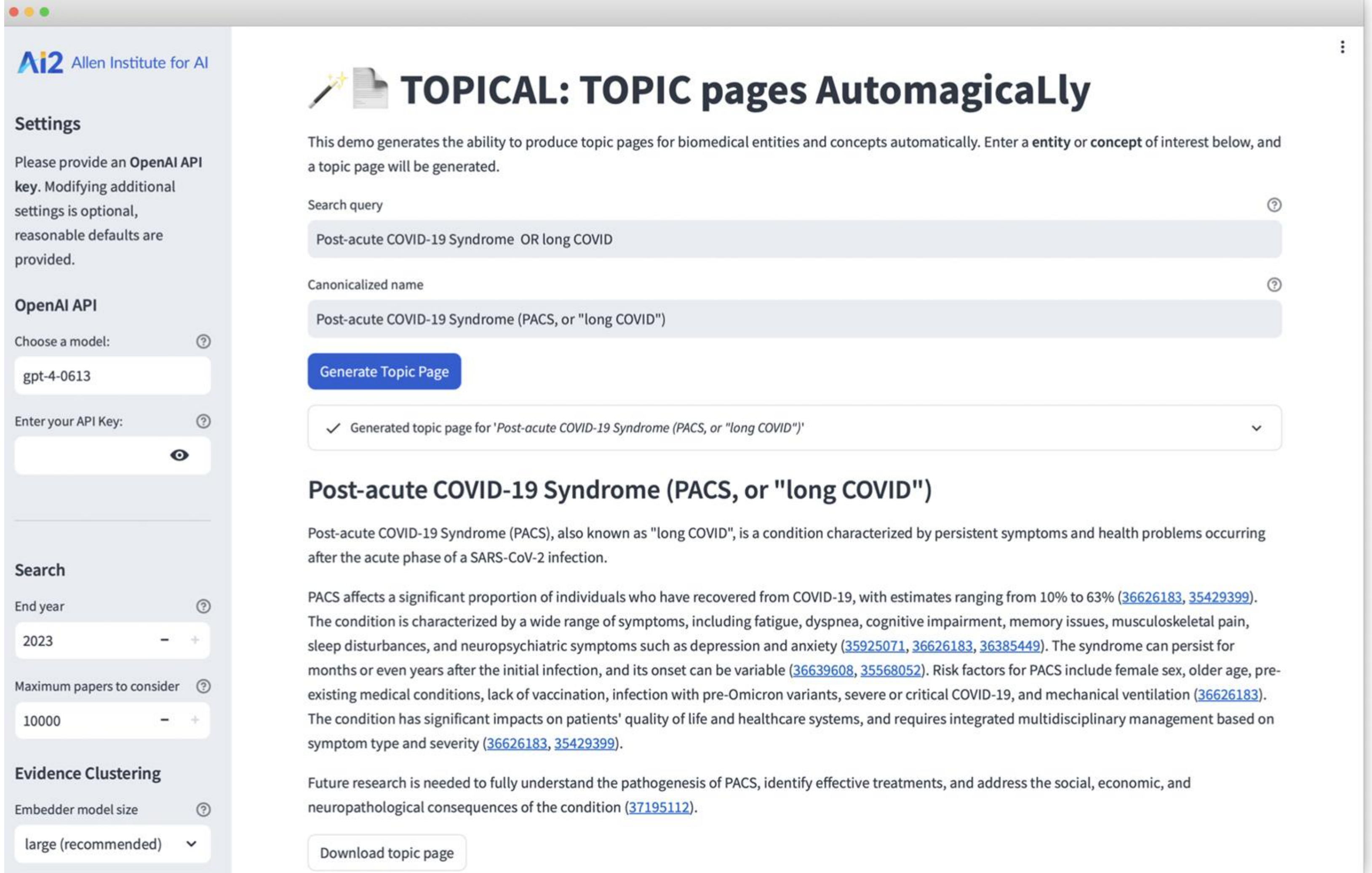
Now, generate the scientific topic page section by section following the instructions below.

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).

TOPICAL web app



Settings

Please provide an **OpenAI API key**. Modifying additional settings is optional, reasonable defaults are provided.

OpenAI API

Choose a model: gpt-4-0613

Enter your API Key:

Search

End year: 2023

Maximum papers to consider: 10000

Evidence Clustering

Embedder model size: large (recommended)

🪄📄 TOPICAL: TOPIC pages AutomagicaLly

This demo generates the ability to produce topic pages for biomedical entities and concepts automatically. Enter a **entity** or **concept** of interest below, and a topic page will be generated.

Search query: Post-acute COVID-19 Syndrome OR long COVID

Canonicalized name: Post-acute COVID-19 Syndrome (PACS, or "long COVID")

✓ Generated topic page for '*Post-acute COVID-19 Syndrome (PACS, or "long COVID")*'

Post-acute COVID-19 Syndrome (PACS, or "long COVID")

Post-acute COVID-19 Syndrome (PACS), also known as "long COVID", is a condition characterized by persistent symptoms and health problems occurring after the acute phase of a SARS-CoV-2 infection.

PACS affects a significant proportion of individuals who have recovered from COVID-19, with estimates ranging from 10% to 63% (36626183, 35429399). The condition is characterized by a wide range of symptoms, including fatigue, dyspnea, cognitive impairment, memory issues, musculoskeletal pain, sleep disturbances, and neuropsychiatric symptoms such as depression and anxiety (35925071, 36626183, 36385449). The syndrome can persist for months or even years after the initial infection, and its onset can be variable (36639608, 35568052). Risk factors for PACS include female sex, older age, pre-existing medical conditions, lack of vaccination, infection with pre-Omicron variants, severe or critical COVID-19, and mechanical ventilation (36626183). The condition has significant impacts on patients' quality of life and healthcare systems, and requires integrated multidisciplinary management based on symptom type and severity (36626183, 35429399).

Future research is needed to fully understand the pathogenesis of PACS, identify effective treatments, and address the social, economic, and neuropathological consequences of the condition (37195112).

Human evaluation

The quality of generated topic pages was evaluated by human specialists. The evaluation itself consisted of two tasks:

- overall quality of a topic page;
- relevance and sufficiency of generated citations.

Human evaluation

The quality of the topic page was represented by three facets *{relevance, accuracy, coherence}* each of which had three options:

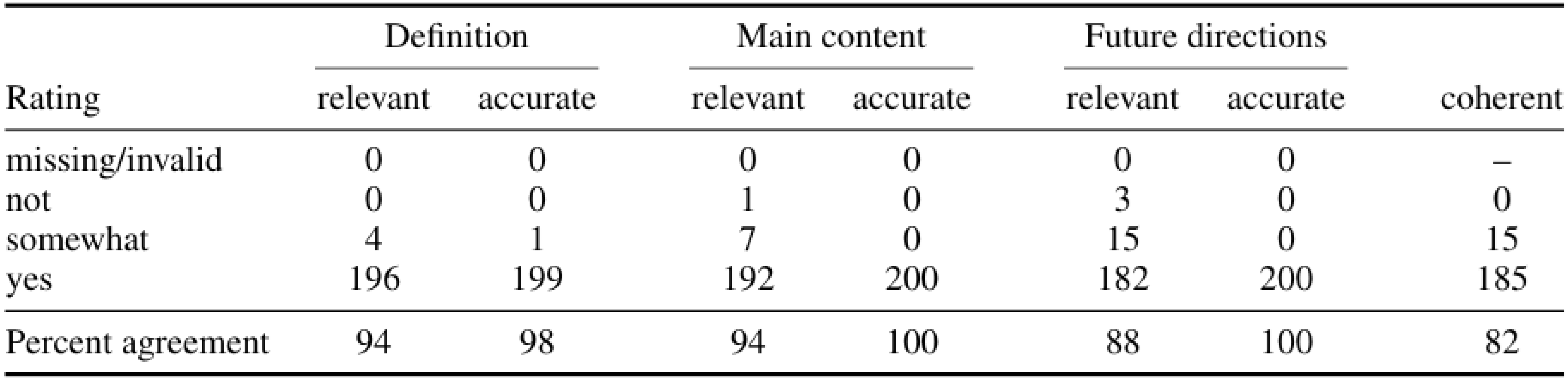
- "not *{relevant, accurate, coherent}*";
- "somewhat *{relevant, accurate, coherent}*";
- "*{relevant, accurate, coherent}*".

Human evaluation

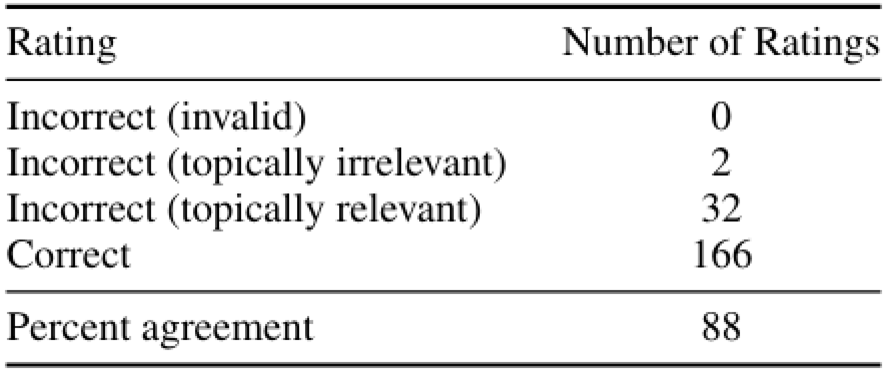
Citations were annotated as:

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

Human evaluation



	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



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

Results

Most topic pages are rated as relevant, accurate, and coherent, with high annotator agreement (≥82%). The model always produces topic pages 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).

Conclusion

The paper introduces TOPICAL, a method for automatically generating high-quality scientific topic pages using LLMs and retrieval-augmented generation. Through human evaluation of 150 biomedical topics, most generated pages were rated as relevant, accurate, and coherent, with correct citations. A public web app is released for on-demand topic page generation.

References

- John Giorgi, Amanpreet Singh, Doug Downey, Sergey Feldman, and Lucy Wang. 2024. TOPICAL: TOPIC Pages AutomagicaLly. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations), pp.1–11, Mexico City, Mexico. Association for Computational Linguistics.
- https://github.com/allenai/TOPICAL
- https://s2-topical.apps.allenai.org/

Thanks!