

# SurveySum: A Dataset for Summarizing Multiple Scientific Articles into a Survey Section

Leandro Carísio Fernandes, Gustavo Bartz Guedes, Thiago Soares Laitz,  
Thales Sales Almeida, Rodrigo Nogueira, Roberto Lotufo, and **Jayr Pereira**



# SurveySum Authors



**Leandro Carísio**



**Gustavo Bartz**



**Thiago Laitz**



**Thales Sales**



**Rodrigo Nogueira**



**Roberto Lotufo**



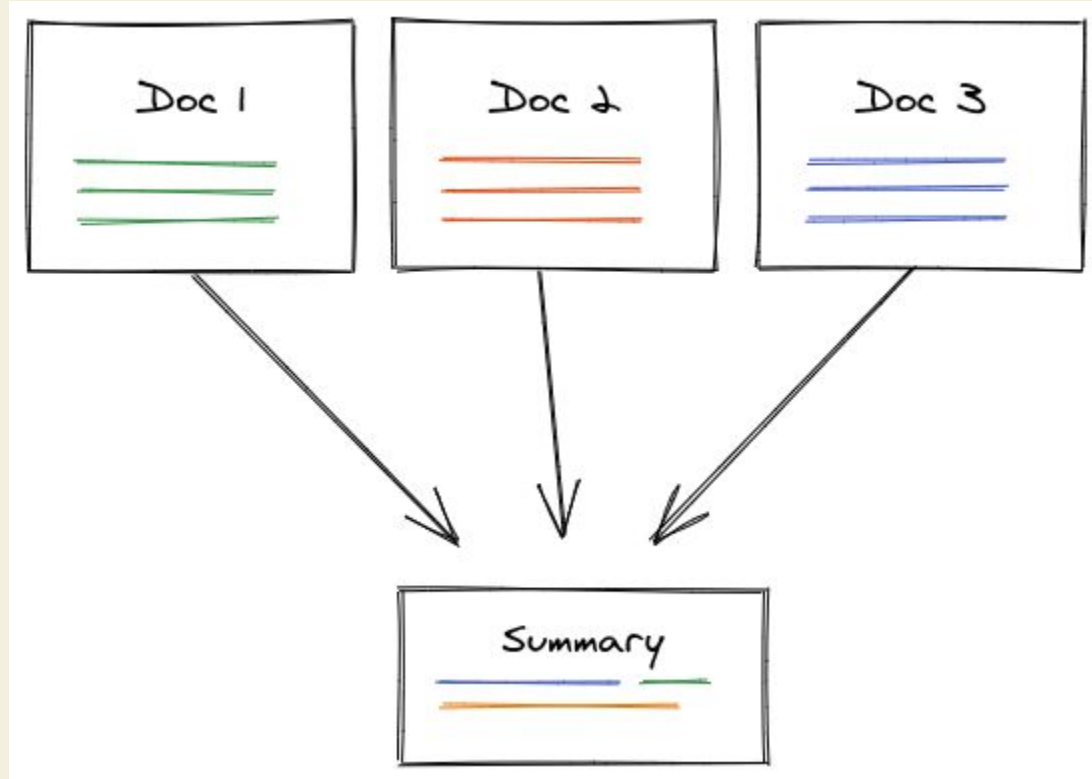
**Jayr Pereira**

# Our contributions

<b>SurveySum</b>
A dataset for Multi-document summarization.

<b>Survey generation pipelines</b>
Two pipelines for generating comprehensive surveys automatically.
Evaluation of both pipelines and some variations, using SurveySum.

# Multi-Document Summarization (MDS)



# A Comprehensive Survey

- A literature review
- Provides a broad overview of a specific topic
- Cover the most relevant and recent research in the field
- Is generally divided into sections
- Each section cover a specific topic

Feng et al. (2021). A Survey of Data Augmentation Approaches for NLP.

## 5.2 Question Answering (QA)

Longpre et al. (2019) investigate various DA and sampling techniques for domain-agnostic QA including paraphrasing by *backtranslation*. Yang et al. (2019) propose a DA method using distant supervision to improve BERT finetuning for open-domain QA. Riabi et al. (2020) leverage Question Generation models to produce augmented examples for zero-shot cross-lingual QA. Singh et al. (2019) propose XLDA, or CROSS-LINGUAL DA, which substitutes a portion of the input text with its translation in another language, improving performance across multiple languages on NLI tasks including the SQuAD QA task. Asai and Hajishirzi (2020) use logical and linguistic knowledge to generate additional training data to improve the accuracy and consistency of QA responses by models. Yu et al. (2018) introduce a new QA architecture called QANet that shows improved performance on SQuAD when combined with augmented data generated using backtranslation.

A comprehensive survey section is a multi-document summary.

The sources (i.e., the cited papers) are  
(generally) long documents.

SurveySum is a dataset for the summarization of multiple and potentially long documents!



# SurveySum: Survey collection criteria

## **1. Comprehensive survey in Fields of interest**

Fields: artificial intelligence, natural language processing, or machine learning.

## **2. Survey must be divided into sections**

Each of the sessions is also required to have at least one citation.

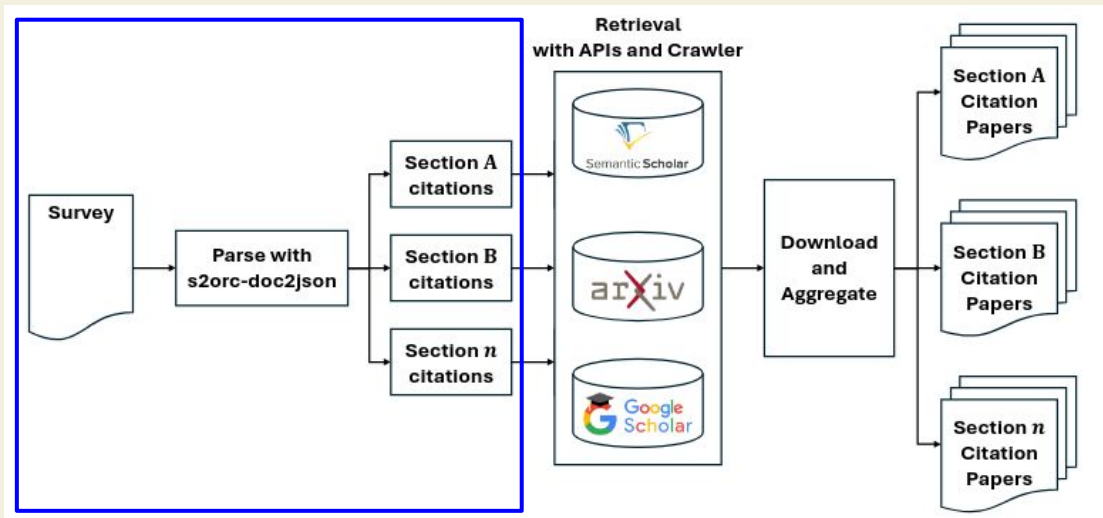
## **3. Survey Must be Freely available online**

## **4. The survey must be written in english**

# SurveySum: Dataset creation

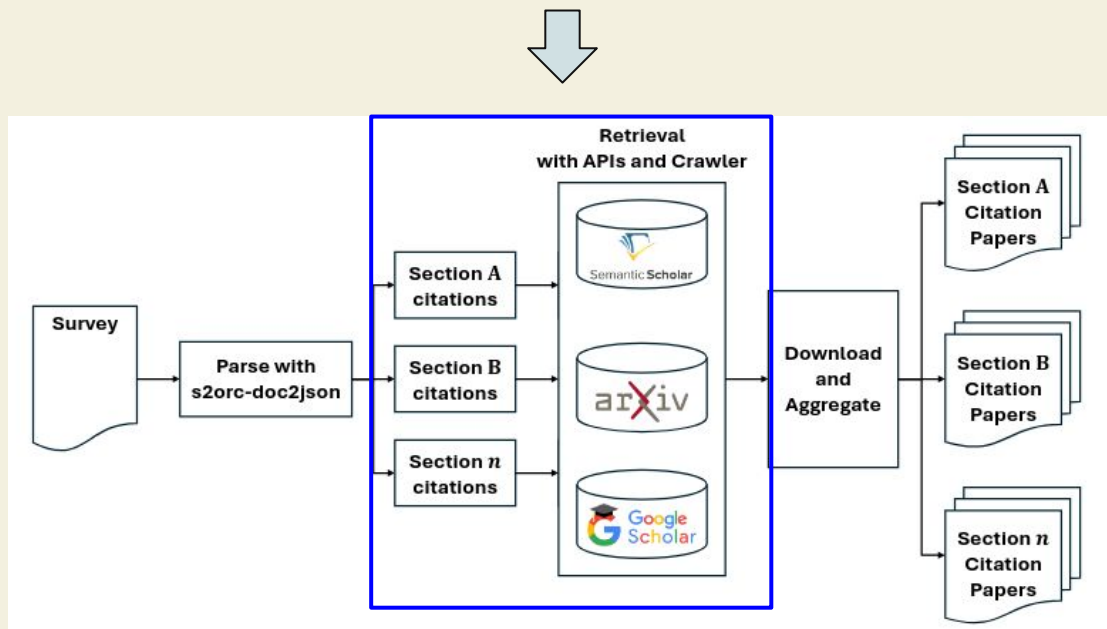


- Following the selection criteria, we gathered 6 total surveys.
- Each of the surveys was parsed with s2orc-doc2json, a library to parse scientific articles to json format.
- With the surveys parsed, we retrieved the name of each paper cited in each section of the surveys.



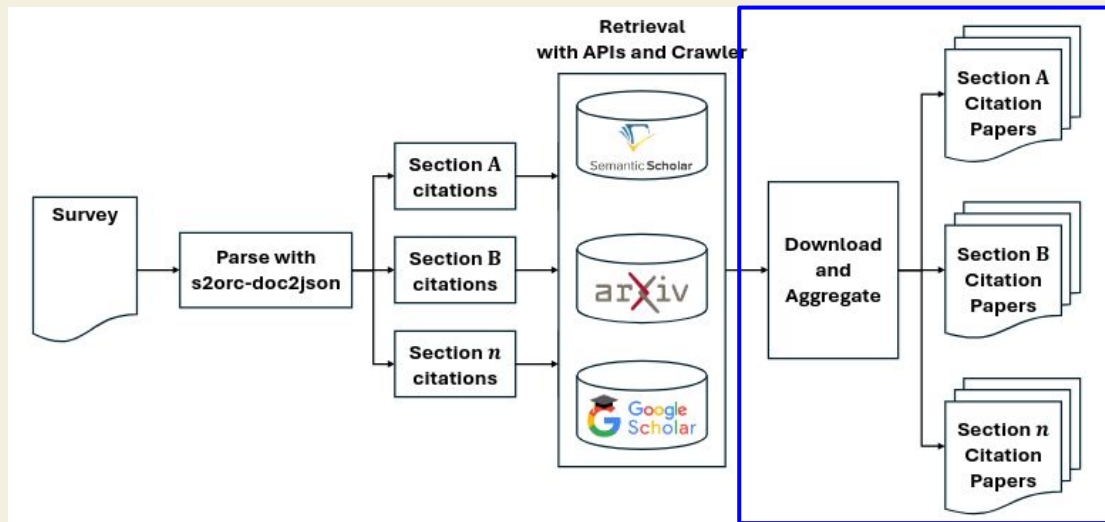
# SurveySum: Dataset creation

- Using various APIs, we retrieved the original paper for each citation present in the original surveys.
- Some papers were not found automatically, in which case the authors manually retrieved the paper.




# SurveySum: Dataset creation

- With all citation papers retrieved, we aggregate them with the original survey section.



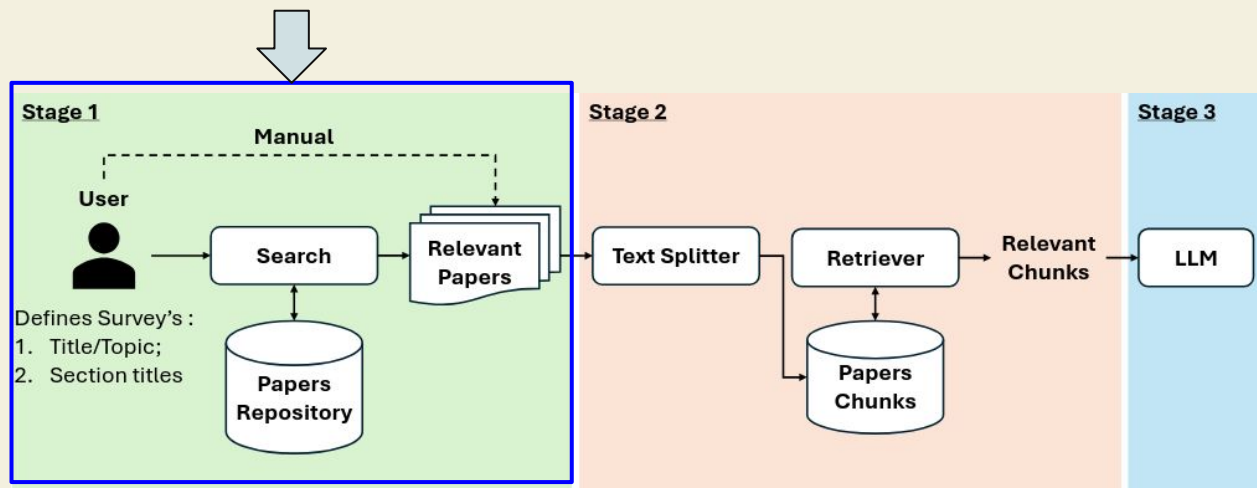
# SurveySum: the Dataset

- 6 surveys
- 79 sections (i.e., summaries)
- 583 cited papers (i.e., source documents)
- An average of ~7 papers per section
- Publicly available at huggingface: <https://huggingface.co/datasets/unicamp-dl/SurveySum>

<b>survey_title</b> string · <i>classes</i>	<b>section_title</b> string · <i>lengths</i>	<b>generated_section_text</b> dict	<b>citations</b> dict
 6 values	 14 128		
"A Comprehensive Survey on Deep Music...	Datasets::MIDI	<code>{"autosurvey_t5_3b_10_chunks": {"references_sent_to_gpt":...</code>	<code>{"BIBREF0":null,"BIBREF1": REF100":null,"BIBREF101":r</code>
A Survey of Chain of Thought Reasoning:...	Discussion::Comparison between...	<code>{"autosurvey_t5_3b_10_chunks": {"references_sent_to_gpt":...</code>	<code>{"BIBREF0":null,"BIBREF1": REF100":null,"BIBREF101":r</code>
A Survey of Chain of Thought Reasoning:...	Methods::XoT Structural Variants::Tree Structure	<code>{"autosurvey_t5_3b_10_chunks": {"references_sent_to_gpt":...</code>	<code>{"BIBREF0":null,"BIBREF1": REF100":null,"BIBREF101":r</code>

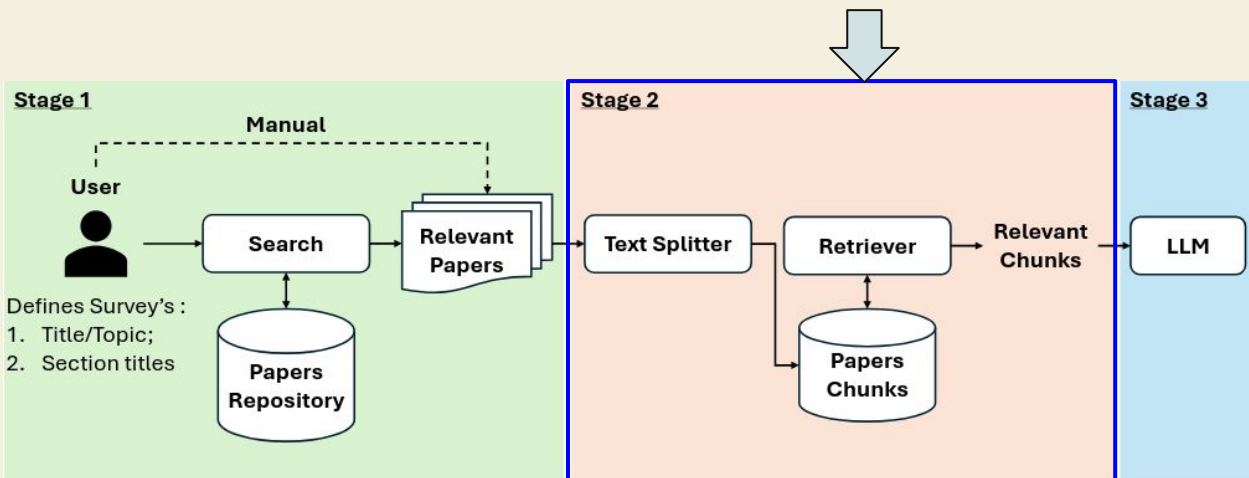
# Automatic Survey generation

- In the first stage the user can define the title of the survey, and the respective sections.
- We also need to define the source of the papers that will be used, this can be a couple of pre selected documents, or the entirety of the papers publicly available.



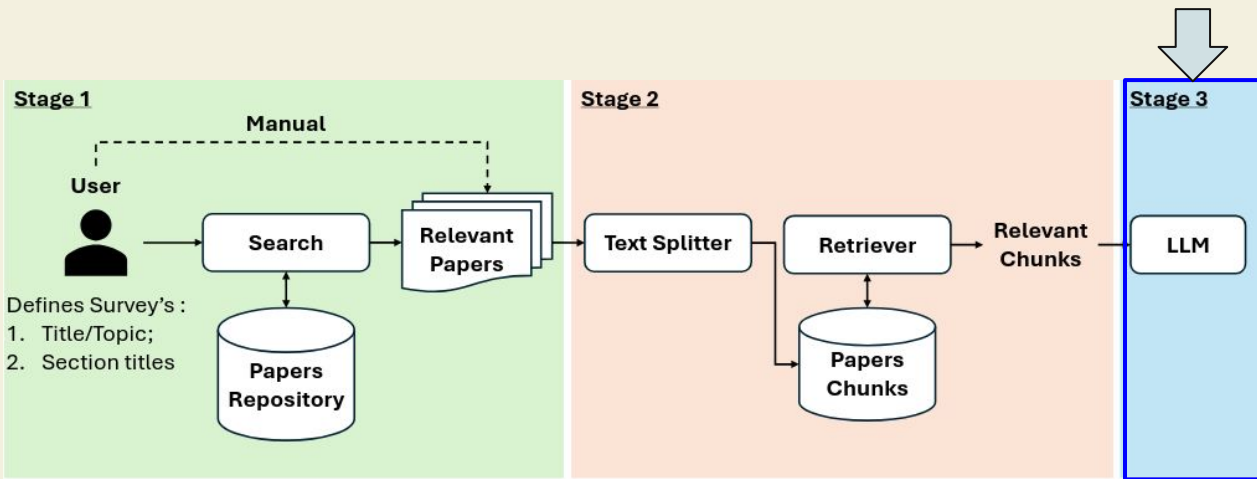
# Automatic Survey generation

- Each relevant paper has its text extracted and splitted in chunks.
- We then apply a retriever, using the section title as the query, to retrieve the most relevant chunks for that section.



# Automatic Survey generation

- With the relevant chunks, we ask a LLM to generate the final survey text.
- In this work we choose to limit the LLMs used to gpt-3.5-turbo and gpt-4-preview.





# Automatic Survey generation

Retrieved  
chunks

**[System]**

You are a renowned scientist who is writing a survey entitled '{TITLE}'.

**[User]**

Your task is to write the text of the section '{SECTION\_TITLE}' of the survey. To complete this task, I will give you a list of documents that should be used as references. Each document has a text and an alphanumeric ID.

When writing the section, you MUST follow this rules:

- be aware of plagiarism, i.e., you should not copy the text, but use them as inspiration.
- when using some reference, you must cite it right after its use. You should use the IEEE citing style (write the id of the text between square brackets).
- you are writing the paragraphs of the section. You MUST write only this section.
- you MUST NOT split the section in subsections, nor create introduction and conclusion for it.
- DO NOT write any conclusion in any form for the subsection.
- DO NOT write a references section.
- DO NOT begin the text writing that the context is 'context', as this is obvious from the title of the survey.

Do you understand your task?

**[AI]**

Sure, send me a list of text and I will write a section about '{SECTION\_TITLE}' using them as references. I am aware that I should use the IEEE citing style.

**[User]**

ID: {REF0}

Text: {CHUNK\_0}

...

ID: {REFN}

Text: {CHUNK\_N}

# Automatic Survey generation – pipelines

## Pipeline 1

Uses a MonoT5 **reranker** model as a retriever.

Uses a **few-shot** prompt for text generation

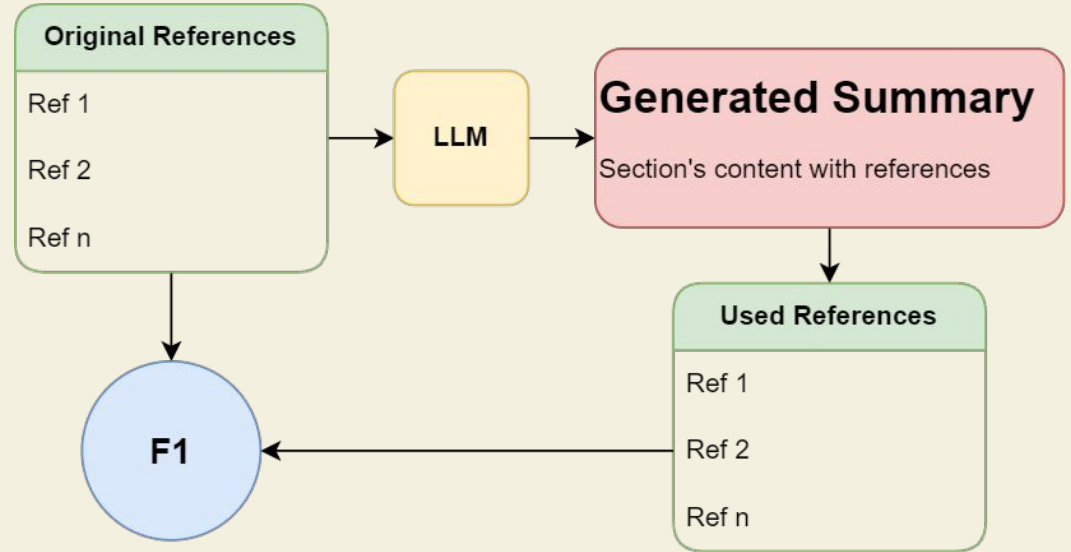
## Pipeline 2

Uses the SPECTER2 model to generate **embeddings** for each chunk, and uses these embeddings for retrieval.

Uses a **single shot** prompt with more complex instructions.

# Evaluation metrics – Reference f1

F1 score is used to evaluate the ability of LLMs to use the same references in their summaries as those found in the original text written by humans.



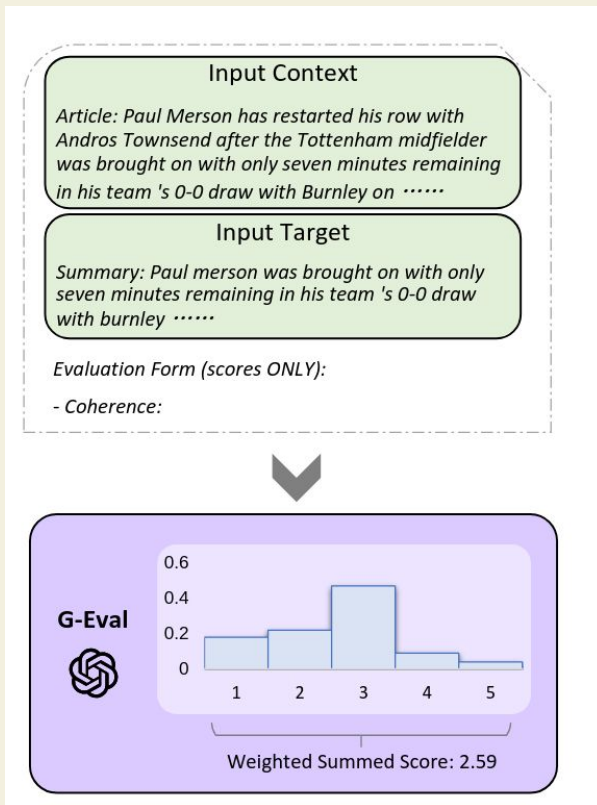
# Evaluation metrics – G-eval

G-eval is a metric for evaluation content generated by LLMs that showed a high correlation with human preference.

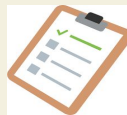
G-eval generates a score for the provided text with an LLM using a high temperature, the score is sampled 20 times.

The final G-eval score is the weighted sum of the scores, weighted by the probability of the judge model to give score.

G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment



# Evaluation metrics – CheckEval



- The Check-Eval metric evaluates the consistency between the candidate text and the reference text using an evaluation checklist.
- The checklist is generated by a LLM model, which is then used to evaluate the candidate text.
- if the evaluation checklist consists of ten elements and the candidate text includes eight of these, the Check-Eval score will be 0.8.

# Results

- Check-Eval and G-eval show a high correlation.
- G-eval presents a much lower variability

Pipeline	LLM	Number of Chunks	Source	IR Pipeline	CHECK-EVAL Score	F1 Score	G-Eval Score
1.1	gpt-3.5-turbo-0125	5	SURVEYSUM	MonoT5	<b>0.415</b>	0.64	3.8
1.2	gpt-3.5-turbo-0125	10	SURVEYSUM	MonoT5	0.244	0.59	3.74
1.3	gpt-3.5-turbo-0125	10	Semantic Scholar	MonoT5	0.141	-	2.84
2.1	gpt-3.5-turbo-0125	1	SURVEYSUM	Embeddings	0.188	0.36	3.42
2.2	gpt-3.5-turbo-0125	5	SURVEYSUM	Embeddings	0.257	0.66	3.9
2.3	gpt-3.5-turbo-0125	10	SURVEYSUM	Embeddings	0.284	0.7	3.9
2.4	gpt-4-0125-preview	1	SURVEYSUM	Embeddings	0.271	0.36	3.82
2.5	gpt-4-0125-preview	5	SURVEYSUM	Embeddings	0.323	0.67	3.99
2.6	gpt-4-0125-preview	10	SURVEYSUM	Embeddings	0.352	<b>0.72</b>	<b>4.01</b>

# Results – More chunks is always better?

- Increasing the number of chunks did not always bring better results.

Pipeline	LLM	Number of Chunks	Source	IR Pipeline	CHECK-EVAL Score	F1 Score	G-Eval Score
1.1	gpt-3.5-turbo-0125	5	SURVEYSUM	MonoT5	<b>0.415</b>	0.64	3.8
1.2	gpt-3.5-turbo-0125	10	SURVEYSUM	MonoT5	0.244	0.59	3.74
1.3	gpt-3.5-turbo-0125	10	Semantic Scholar	MonoT5	0.141	-	2.84
2.1	gpt-3.5-turbo-0125	1	SURVEYSUM	Embeddings	0.188	0.36	3.42
2.2	gpt-3.5-turbo-0125	5	SURVEYSUM	Embeddings	0.257	0.66	3.9
2.3	gpt-3.5-turbo-0125	10	SURVEYSUM	Embeddings	0.284	0.7	3.9
2.4	gpt-4-0125-preview	1	SURVEYSUM	Embeddings	0.271	0.36	3.82
2.5	gpt-4-0125-preview	5	SURVEYSUM	Embeddings	0.323	0.67	3.99
2.6	gpt-4-0125-preview	10	SURVEYSUM	Embeddings	0.352	<b>0.72</b>	<b>4.01</b>

Worse with more chunks

Better with more chunks

# Results - The importance of the LLM

- When changing only the LLM used for writing the section, the result always improved when using a more capable LLM.

Pipeline	LLM	Number of Chunks	Source	IR Pipeline	CHECK-EVAL Score	F1 Score	G-Eval Score
1.1	gpt-3.5-turbo-0125	5	SURVEYSUM	MonoT5	<b>0.415</b>	0.64	3.8
1.2	gpt-3.5-turbo-0125	10	SURVEYSUM	MonoT5	0.244	0.59	3.74
1.3	gpt-3.5-turbo-0125	10	Semantic Scholar	MonoT5	0.141	-	2.84
2.1	gpt-3.5-turbo-0125	1	SURVEYSUM	Embeddings	0.188	0.36	3.42
2.2	gpt-3.5-turbo-0125	5	SURVEYSUM	Embeddings	0.257	0.66	3.9
2.3	gpt-3.5-turbo-0125	10	SURVEYSUM	Embeddings	0.284	0.7	3.9
2.4	gpt-4-0125-preview	1	SURVEYSUM	Embeddings	0.271	0.36	3.82
2.5	gpt-4-0125-preview	5	SURVEYSUM	Embeddings	0.323	0.67	3.99
2.6	gpt-4-0125-preview	10	SURVEYSUM	Embeddings	0.352	<b>0.72</b>	<b>4.01</b>



# Results - The importance of the source

- When using all documents available in Semantic Scholar, instead of the pre selected documents from SurveySum. The result was much worse.
- The quality of documents automatically retrieved from SS, was much worse, resulting in a worse generated section.

Pipeline	LLM	Number of Chunks	Source	IR Pipeline	CHECK-EVAL Score	F1 Score	G-Eval Score
1.1	gpt-3.5-turbo-0125	5	SURVEYSUM	MonoT5	<b>0.415</b>	0.64	3.8
1.2	gpt-3.5-turbo-0125	10	SURVEYSUM	MonoT5	0.244	0.59	3.74
1.3	gpt-3.5-turbo-0125	10	Semantic Scholar	MonoT5	0.141	-	2.84

# Conclusions

- **Importance of Source Papers:** High-quality and relevant source papers are crucial for accurate and coherent summaries.
- **Role of LLM Quality:** Advanced models like GPT-4 significantly outperform earlier versions, highlighting their impact on summary quality.
- **Pipeline Insights:** Effective retrieval and selection of text chunks enhance summarization outcomes but depend on robust retrievers.
- **Benchmark Contribution:** SurveySum provides a benchmark for developing and evaluating multi-document summarization models in scientific contexts.
- **Github:** <https://github.com/unicamp-dl/surveysum>

# Obrigado!

SurveySum: A Dataset for Summarizing Multiple Scientific Articles into a Survey Section

