Empowering Meta-Analysis: Leveraging Large Language Models for Scientific Synthesis

Jawad Ibn Ahad
Apurba-NSU R&D Lab, ECE
North South University
Dhaka, Bangladesh
jawad.ibn@northsouth.edu

Rafeed Mohammad Sultan
Apurba-NSU R&D Lab, ECE
North South University
Dhaka, Bangladesh
rafeed.sultan@northsouth.edu

Abraham Kaikobad
Apurba-NSU R&D Lab, ECE
North South University
Dhaka, Bangladesh
abraham.kaikobad@northsouth.edu

Fuad Rahman
Apurba Technologies
Sunnyvale, CA 94085, USA
fuad@apurbatech.com

Mohammad Ruhul Amin
Computer and Information Science
Fordham University
New York, USA
mamin17@fordham.edu

Nabeel Mohammed
Apurba-NSU R&D Lab, ECE
North South University
Dhaka, Bangladesh
nabeel.mohammed@northsouth.edu

Shafin Rahman
Apurba-NSU R&D Lab, ECE
North South University
Dhaka, Bangladesh
shafin.rahman@northsouth.edu

***Abstract*—This additional material offers supplementary details that bolster the conclusions outlined in the main manuscript.**

- **Section 1: Experimental Details (additional discussion in support of Section 4.A and 3.C of the main paper).**
- **Section 2: Evaluation (additional discussion in support of Section 4.A of the main paper).**

I. Experimental Details:

Table I provides a detailed breakdown of the hyperparameters we used in our experiments. The two models underwent ten epochs of fine-tuning, with a batch size of four and a learning rate of 2e-4. The Supervised Fine-tuning Trainer (SFTTrainer) with customized loss metrics, led the entire training process.

A. Set up for RAG

Setup for RAG: For the retrieval augmented generation approach, sentence transformers and embeddings handle large context lengths. The data is processed using the "Recursive TextSplitter" from *LangChain*[1] to break the text into manageable chunks, satisfying LLMs maximum context length. Subsequently, the *HuggingFaceEmbeddings* model "sentence-transformers/all-MiniLM-L6-v2" generates embeddings for the text chunks. Sentences are tokenized and converted into numerical tokens using the *AutoTokenizer* from *HuggingFace*[2]. The pre-trained Sentence Transformer model "all-MiniLM-L6-v2" obtains encoded inputs, including padding and truncation, as specified in the tokenizer. The normalized embeddings are stored in *Pinecone*[3], creating a searchable index. A question-answering chain, developed as a *RetrievalQA*, retrieves information based on the query. For the RAG approach, the dataset includes a third column containing specific queries regarding the context, effectively framing the query as a question about the contents of the meta-papers' abstracts.

Vector database: For the RAG approach, a vector database is necessary to store the embeddings of provided large contexts. The normalized embeddings are stored in *Pinecone*[3], creating a searchable index. *Pinecone* serves as the vector knowledge base, storing the embeddings from the sentence transformers.

Question Answering Chain: A question-answering chain, developed as a *RetrievalQA*, retrieves information based on the query. For the RAG approach, the dataset includes a third column containing specific queries regarding the context, effectively framing the query as a question about the contents of meta-papers' abstracts.

B. Set up for combining RAG with Fine-tuned LLM

Combining Retrieval Augmentation (RAG) with fine-tuned models enhances Language Model Models' (LLMs) performance, especially for tasks like generating meta-analysis abstracts. Fine-tuning LLMs improves task-specific capabilities but struggles with large contexts. RAG acts as an information bridge, retrieving relevant knowledge from external sources, enhancing accuracy, reducing fine-tuning needs, and mitigating irrelevant content generation. RAG ensures semantic alignment between retrieved information and user queries, improving the quality, readability, and clarity of meta-analysis abstracts. This integrated approach produces more accurate, focused, and readable content.

Furthermore, while fine-tuning large language models (LLMs) allows them to excel at specific tasks like generating meta-analysis abstracts, their inherent limitation in handling large contexts becomes a hurdle. This is where combining retrieval augmentation (RAG) with fine-tuning offers a powerful solution. RAG acts like an information bridge, retrieving relevant knowledge from external sources beyond the limited chunks provided. This expanded knowledge base empowers the LLM to process complex information from various sources, leading

[1] LangChain: https://www.langchain.com/
[2] HuggingFace: https://huggingface.co/
[3] Pinecone: https://www.pinecone.io/

TABLE I: Hyperparameter values used for our experiment

Hyperparameter	**Value**	**Hyperparameter**	**Value**
Batch Size	4	Learning Rate	2e-4
Epochs	10	Temperature	0.7
Loss Function	ICD	Optimizer	AdamW
Lora Alpha	4	Lora Dropout	0.1
Rank (r)	64	–	–

to several advantages. Firstly, the accuracy and factual grounding of the generated abstracts are significantly enhanced. Since factual accuracy is crucial for meta-analysis, RAG ensures the LLM incorporates reliable external knowledge. Secondly, leveraging external information through RAG has the potential to reduce the extensive fine-tuning required for LLMs, leading to faster training times. Finally, RAG helps mitigate the issue of irrelevant content generation that can sometimes plague fine-tuned models. In Retrieval-Augmented Generation (RAG) systems, semantic search ensures retrieved information is not just relevant, but also semantically aligns with the task. RAG uses semantic similarity to identify contexts that closely match the user's query and the expertise of the fine-tuned large language model (LLM). This injects relevant information, enhancing the LLM's understanding and synthesis of meta-analysis abstracts, leading to higher-quality outputs. Ultimately, RAG contributes to improved readability and clarity in the final abstracts. This combined approach, where fine-tuning grants the LLM the ability to extract relevant patterns for meta-analysis generation and RAG broadens the context with the necessary information, leads to more accurate, focused, and readable meta-analysis content.

This integrated process can reduce hallucinations and produce more accurate answers efficiently. The fine-tuned models are employed to perform RAG on unseen data. As aforementioned above, Figure 2 illustrates the complete workflow for this approach.

C. Data Processing and Context Analysis

The dataset, comprising main papers and their associated metadata, underwent preprocessing and context analysis which is shown in Fig1. The preprocessing steps involved chunking the metadata using a custom text splitter, designed to partition texts into manageable segments. Each segment was associated with its corresponding main paper label.

Subsequently, the length of each chunk, measured in characters, was computed for both the main paper and its metadata.

As previously mentioned, it will not be feasible to send all the information at once because the context size for abstracts will be greater than the context length of LLMs. Figure 2 illustrates the possibility of providing input from a single chunk for only fine-tuning inference. Thus, it goes without saying that the model will be unable to extract any data from the other chunked contexts. On the fine-tuned model, RAG is applied to solve this issue. This will make it possible to insert the entire context, including each abstract. This extensive context will be subjected to a semantic search process using queries, and the context that is found will be sent as input to LLM. LLM will acquire knowledge about each chunk in this manner. It is observed that through the fine-tuning process, models learned the patterns of constructing meta-analysis abstracts. Additionally, by applying RAG, they were able to learn from large unseen contexts, resulting in better performance.

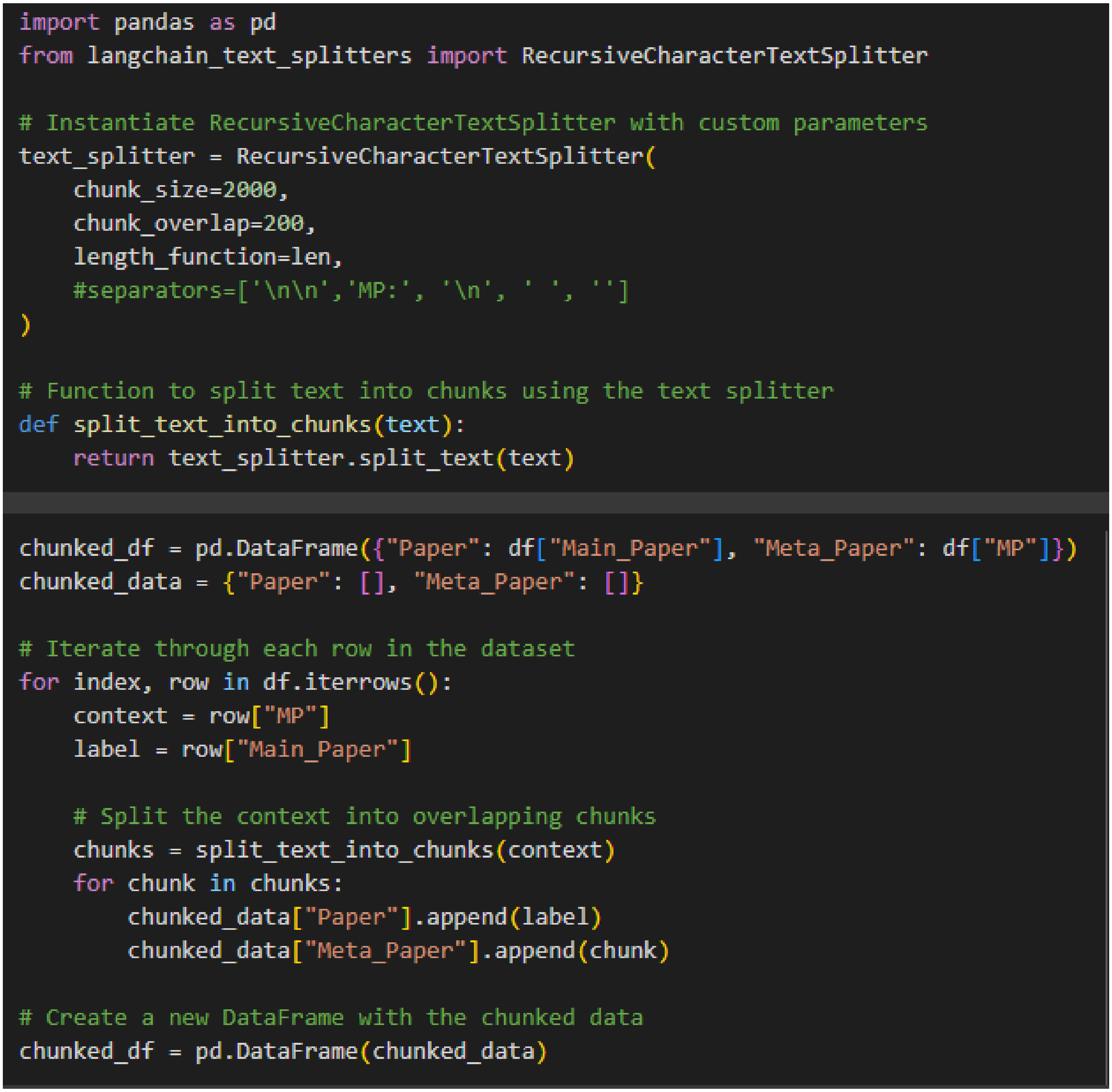
II. Evaluation

Human Evaluation. In this section, we outline the evaluation process employed in this study. Following the generation of responses by LLMs, a human evaluation process is conducted to align the generated text with human judgment. This process, adapted from a previous study, involves human judges categorizing LLM-generated text into three categories: ”Relevant” (2), indicating a close resemblance to the ground truth with high similarity and inclusion of important input information; ”Somewhat-Relevant” (1), suggesting an acceptable similarity with valuable information within an acceptable margin; and ”Irrelevant” (0), signifying a lack of important information or the presence of irrelevant contexts. This classification framework ensures a rigorous assessment of generated meta-analysis abstracts against expected standards.

The evaluation process involves three individuals independently assessing statements and responses from each model output as shown in Fig3, which was done in Google Sheets. These evaluators, instructed to evaluate relevancy, conduct their assessments without access to each other's evaluations. Final results are determined through hard voting by a designated final evaluator. If two of the evaluators select ”Relevant” (2), the final vote is considered as 2; the same applies for ”Irrelevant” (0). However, in cases where there are different evaluation labels, such as 0, 1, and 2, and there is no agreement among the evaluators, the final decision is assigned as 1. This is because ”Somewhat-Relevant” (1) is considered almost similar to ”Relevant” (2), indicating some level of relevancy in the processed output of models. Thus, if one evaluator labels it as 2 (relevant) and another as 1 (somewhat-relevant), the processed output is deemed somewhat-relevant, resulting in a final vote of 1. To mitigate biases, the human evaluation is conducted by university students who are not authors of the study.

Evaluation with RAG

Fig. 4 displays a sample for evaluation for the Retrieval Augmented Generation (RAG) approach. Cosine-similarity



```python
import pandas as pd
from langchain_text_splitters import RecursiveCharacterTextSplitter

# Instantiate RecursiveCharacterTextSplitter with custom parameters
text_splitter = RecursiveCharacterTextSplitter(
    chunk_size=2000,
    chunk_overlap=200,
    length_function=len,
    #separators=['\n\n','MP:', '\n', ' ', '']
)

# Function to split text into chunks using the text splitter
def split_text_into_chunks(text):
    return text_splitter.split_text(text)
```

```python
chunked_df = pd.DataFrame({"Paper": df["Main_Paper"], "Meta_Paper": df["MP"]})
chunked_data = {"Paper": [], "Meta_Paper": []}

# Iterate through each row in the dataset
for index, row in df.iterrows():
    context = row["MP"]
    label = row["Main_Paper"]

    # Split the context into overlapping chunks
    chunks = split_text_into_chunks(context)
    for chunk in chunks:
        chunked_data["Paper"].append(label)
        chunked_data["Meta_Paper"].append(chunk)

# Create a new DataFrame with the chunked data
chunked_df = pd.DataFrame(chunked_data)
```

Fig. 1: Illustration of the Chunking Process. The text splitter algorithm segments the metadata into overlapping chunks, facilitating efficient processing and analysis of the dataset

is used to evaluate the produced meta-analysis using RAG because the loss function for fine-tuning LLM was designed using dissimilarity. Again, due to the approach of RAG, LLMs are bound to produce contexts that are expected to be relevant to the input context. So, the generated outputs via RAG are relevant meta-analysis abstracts. For this reason, we need to calculate the similarity between two abstracts: one is an actual meta-analysis abstract, and the other is a generated abstract in the RAG approach which is the processed output. The average cosine similarity is listed in the paper.

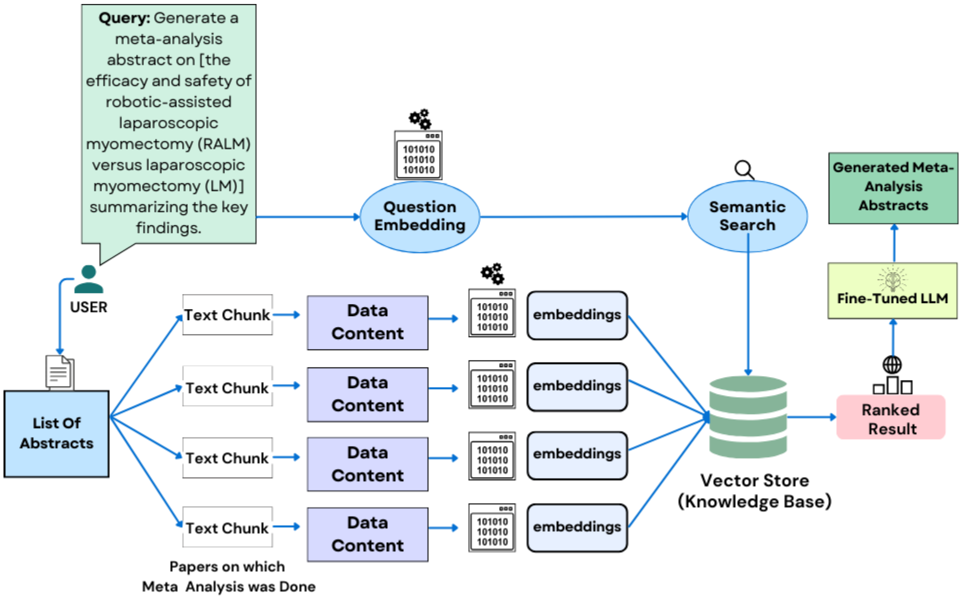
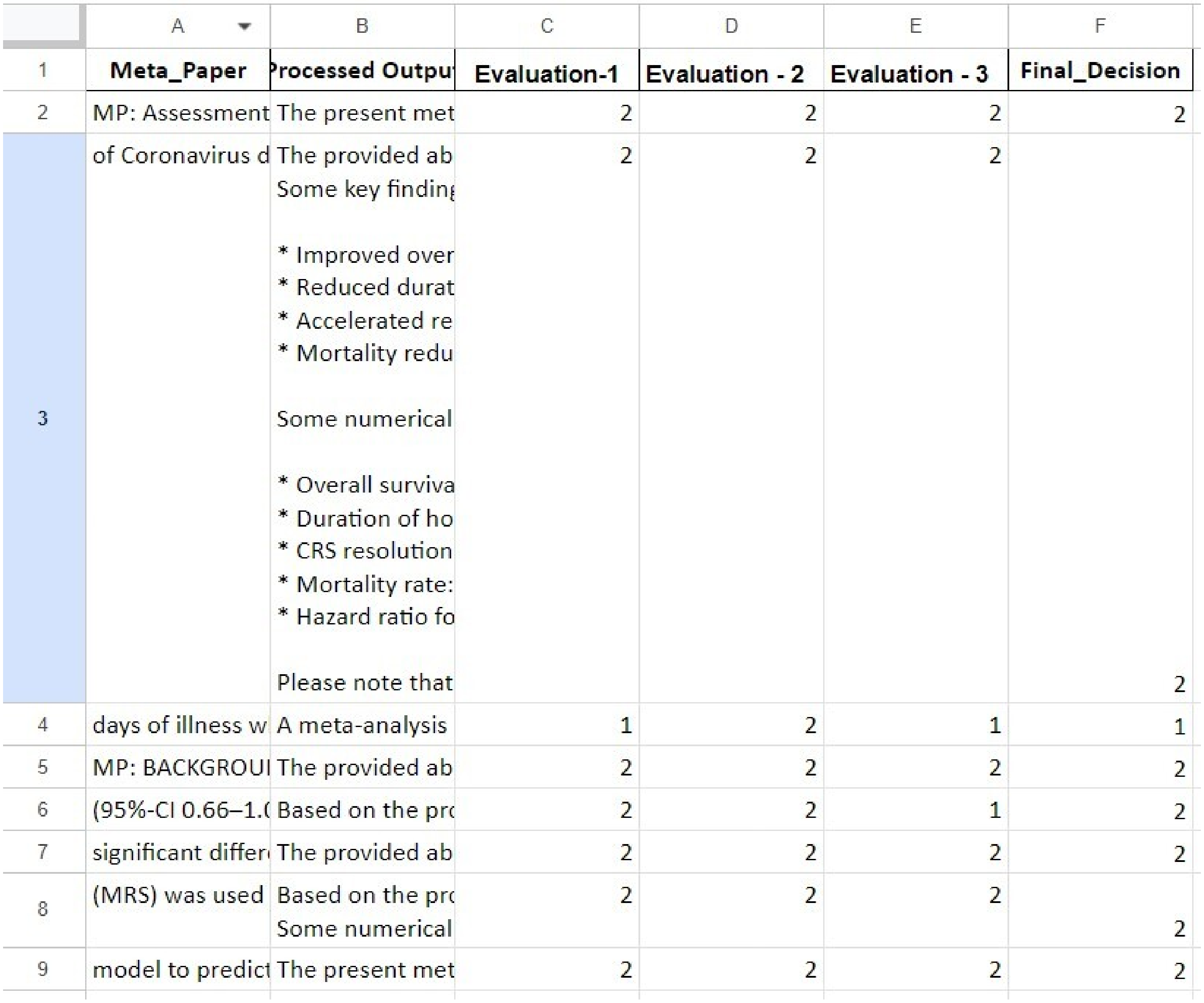
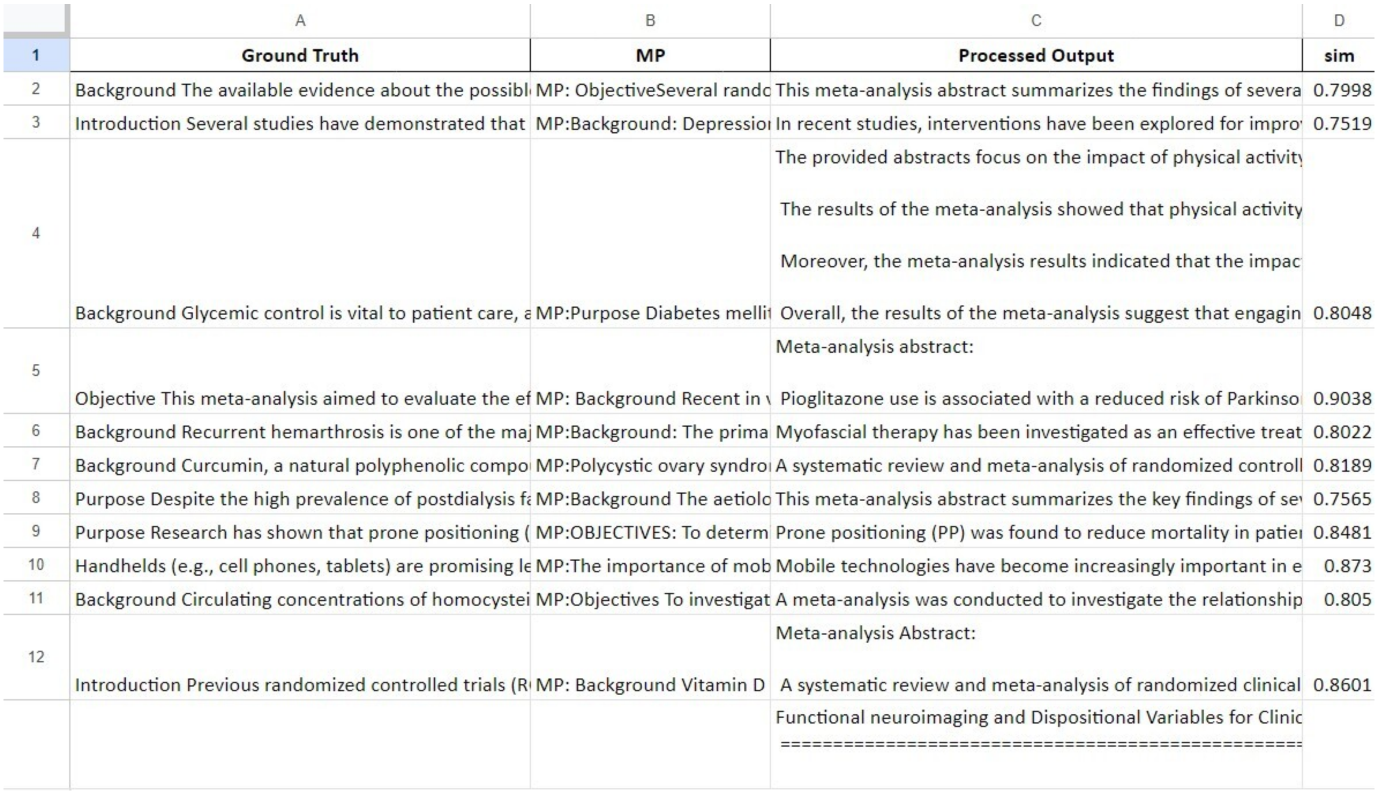

Query: Generate a meta-analysis abstract on [the efficacy and safety of robotic-assisted laparoscopic myomectomy (RALM) versus laparoscopic myomectomy (LM)] summarizing the key findings.
USER
List Of Abstracts
Text Chunk
Text Chunk
Text Chunk
Text Chunk
Papers on which Meta Analysis was Done
Data Content
Data Content
Data Content
Data Content
101010 101010 101010
embeddings
embeddings
embeddings
embeddings
Question Embedding
Semantic Search
Vector Store (Knowledge Base)
Ranked Result
Fine-Tuned LLM
Generated Meta-Analysis Abstracts

Fig. 2: The user will submit a query along with abstracts from several publications and a broad context. There will be several segments that are taken out of context. Every chunk's vector embeddings will be kept in a vector database. Using the embedded query in the vector base, a semantic search will be carried out. The context that is retrieved will be fed to the fine-tuned LLM that was trained on the conventions for generating meta-analysis abstracts. Lastly, the model produces an abstract for a meta-analysis.



	A	B	C	D	E	F
1	Meta_Paper	Processed Outpu	Evaluation-1	Evaluation - 2	Evaluation - 3	Final_Decision
2	MP: Assessment	The present met	2	2	2	2
3	of Coronavirus d	The provided ab Some key finding * Improved over * Reduced durat * Accelerated re * Mortality redu Some numerical * Overall surviva * Duration of ho * CRS resolution * Mortality rate: * Hazard ratio fo Please note that	2	2	2	2
4	days of illness w	A meta-analysis	1	2	1	1
5	MP: BACKGROUI	The provided ab	2	2	2	2
6	(95%-CI 0.66–1.(	Based on the prc	2	2	1	2
7	significant differ	The provided ab	2	2	2	2
8	(MRS) was used	Based on the prc Some numerical	2	2	2	2
9	model to predict	The present met	2	2	2	2

Fig. 3: Human evaluation spreadsheet showing meta-analysis abstract papers, LLM-generated outputs, evaluations of three individuals, and the overall evaluation.



	A	B	C	D
1	**Ground Truth**	**MP**	**Processed Output**	**sim**
2	Background The available evidence about the possibl	MP: ObjectiveSeveral rando	This meta-analysis abstract summarizes the findings of severa	0.7998
3	Introduction Several studies have demonstrated that	MP:Background: Depressior	In recent studies, interventions have been explored for impro	0.7519
4	Background Glycemic control is vital to patient care, a	MP:Purpose Diabetes mellit	The provided abstracts focus on the impact of physical activity The results of the meta-analysis showed that physical activity Moreover, the meta-analysis results indicated that the impac Overall, the results of the meta-analysis suggest that engagin	0.8048
5	Objective This meta-analysis aimed to evaluate the ef	MP: Background Recent in v	Meta-analysis abstract: Pioglitazone use is associated with a reduced risk of Parkinso	0.9038
6	Background Recurrent hemarthrosis is one of the maj	MP:Background: The prima	Myofascial therapy has been investigated as an effective treat	0.8022
7	Background Curcumin, a natural polyphenolic compo	MP:Polycystic ovary syndroı	A systematic review and meta-analysis of randomized controll	0.8189
8	Purpose Despite the high prevalence of postdialysis fa	MP:Background The aetiolo	This meta-analysis abstract summarizes the key findings of sev	0.7565
9	Purpose Research has shown that prone positioning (	MP:OBJECTIVES: To determ	Prone positioning (PP) was found to reduce mortality in patie	0.8481
10	Handhelds (e.g., cell phones, tablets) are promising le	MP:The importance of mob	Mobile technologies have become increasingly important in e	0.873
11	Background Circulating concentrations of homocystei	MP:Objectives To investigat	A meta-analysis was conducted to investigate the relationship	0.805
12	Introduction Previous randomized controlled trials (R	MP: Background Vitamin D	Meta-analysis Abstract: A systematic review and meta-analysis of randomized clinical	0.8601
			Functional neuroimaging and Dispositional Variables for Clinic ==	

Fig. 4: Evaluation of processed output using RAG on fine-tuned Mistral-v0.1 (7B) LLM is shown here. The last column ‘sim‘ means similarity with ground truth (SWGT), which refers to cosine-similarity between the generated meta-analysis abstract and the actual meta-analysis abstract.