

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

Jawad Ibn Ahad<sup>1</sup>, Rafeed Mohammad Sultan<sup>1</sup>, Abraham Kaikobad<sup>1</sup>, Fuad Rahman<sup>2</sup>, Mohammad Ruhul Amin<sup>3</sup>, Nabeel Mohammed<sup>1</sup>, and Shafin Rahman<sup>1</sup>

<sup>1</sup> Apurba-NSU R&D Lab, Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

<sup>2</sup> Apurba Technologies, Sunnyvale, CA 94085, USA

<sup>3</sup> Fordham University, USA

{jawad.ibn, rafeed.sultan, abraham.kaikobad, nabeel.mohammed, shafin.rahman}@northsouth.edu

fuad@apurbatech.com, mamin17@fordham.edu

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

- Section 1: Data Collection (additional discussion in support of Section 3.1 of the main paper).
- Section 2: Experimental Details (additional discussion in support of Section 3.2 and 4.1 of the main paper).
- Section 3: Evaluation (additional discussion in support of Section 4.1 of the main paper).

## 1 Data Collection

Here, in this section, we delve into the methodology of us authors have curated this meta-analysis dataset. For this research, data collection played a crucial

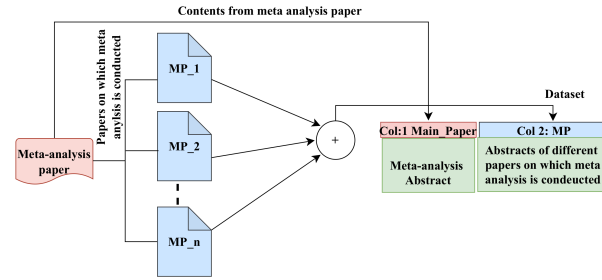


Fig. 1: Procedure for gathering data. From the meta-analysis paper, a number of papers are gathered on which meta analysis is performed.

role. The dataset for this study comprises two columns: one for meta-analysis

Table 1: Hyperparameter values used for our experiment

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

abstracts, and another for abstracts of the individual research papers included in the meta-analysis. Figure 1 depicts the process of data collection. For our research endeavor, we meticulously curated a corpus of 625 meta-analysis papers sourced from Science Direct. From this comprehensive dataset, we conducted a rigorous process of manual extraction to identify papers on meta-analysis. This extraction involved meticulous scrutiny of the information presented in tables and figures within the meta-analysis main papers, along with reference cross-referencing. Additionally, we augmented our paper collection by sourcing pertinent references from the meta-analysis section of identified meta-analysis papers. These references provided invaluable guidance in identifying and extracting papers and their corresponding abstracts, which were subsequently organized and cataloged in an adjacent column alongside their corresponding meta-analysis papers. In total, 6344 papers’ abstracts were listed alongside 625 meta-analysis papers’ abstracts. Abstracts from recently published meta-analysis studies were scraped as the human-written ground truth, while the abstracts of the other studies included in the meta-analysis were treated as meta-paper abstracts. To collect data, the authors selected a recently published meta-analysis study. The abstract of this paper was scraped as the human-written ground truth, while the abstracts of the other studies included in the meta-analysis were treated as meta-paper abstracts.

When a meta-analysis is conducted on a set of papers, the abstract of the meta-analysis paper is expected to encapsulate the information present in the abstracts of those constituent papers. Given the prevalence of meta-analyses in the medical field, the focus of this research was primarily within this domain.

## 2 Experimental Details:

Table 1 provides a detailed breakdown of the hyperparameters we used in our experiments. The two models underwent four 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.

## 2.1 Prompt Selection:

The selection of prompts significantly impacts model performance by providing instructions on how to approach a task. Through rigorous exploration, the efficacy of various prompts was assessed to determine the most effective one. For instance, when analyzing a meta-paper’s abstract regarding the assessment of therapeutic plasma exchange (TPE) in treating life-threatening COVID-19, Prompt 1 produced more informative outputs compared to Prompt 2. Furthermore, text generated using Prompt 1 by the fine-tuned Llama-2 (7B) model demonstrated higher similarity (ROUGE = 0.39) to the ground truth compared to the non-fine-tuned model (ROUGE = 0.29). As a result, Prompt 1 was selected for subsequent experiments, highlighting its superior performance in guiding model outputs.

## 2.2 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*<sup>4</sup> 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*<sup>5</sup>. 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*<sup>6</sup>, 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*<sup>6</sup>, 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.

## 2.3 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 gener-

<sup>4</sup> LangChain: <https://www.langchain.com/>

<sup>5</sup> HuggingFace: <https://huggingface.co/>

<sup>6</sup> Pinecone: <https://www.pinecone.io/>

Table 2: Comparison of Outputs for Prompt 1 and Prompt 2

| Prompt 1 and generated context  | Prompt 2 and generated context   |
|---|--|
| <b>Prompt:</b> Given a collection of abstracts from papers used in various medical fields for meta-analysis, generate a meta-analysis abstract. Summarize the key findings and provide numerical values or statistical information for specific observations that are commonly reported in the provided abstracts.  | <b>Prompt:</b> There are given some abstracts of papers that are used for meta-analysis in different medical fields. Generate a meta-analysis abstract based on the given abstracts of papers. Please try to provide numerical values for any specific findings that were used in most of the abstracts.   |
| <b>Output:</b> The present meta-analysis aimed to summarize the key findings of the provided abstracts related to the efficacy and safety of therapeutic plasma exchange (TPE) in critically ill COVID-19 patients. A total of 87 abstracts were included in the analysis. Key Findings: TPE was associated with improved clinical recovery (as measured by SOFA score) and reduced days on MV and ICU LOS. There was no significant difference in 35-day mortality between the TPE and standard treatment groups. 35-day mortality: TPE, 20.9% (95% CI, 15.5–28.3%); standard treatment, 34.1% (95% CI, 27.2–41.3%). Days on MV: TPE, 10.3 (95% CI, 7.6–13.0); standard treatment, 14.3 (95% CI, 10.9–17.7). ICU LOS: TPE, 8.0 (95% CI, 5.9–10.1); standard treatment, 12.7... | <b>Output:</b> The aim of this meta-analysis was to evaluate the efficacy and safety of therapeutic plasma exchange (TPE) in critically ill patients with COVID-19. TPE was administered at different time points after disease onset. The pooled data showed that TPE was associated with a significant reduction in 35-day mortality (odds ratio (OR) = 0.62, 95% confidence interval (CI) = 0.45–0.85, P = 0.003). Subgroup analyses revealed that TPE was more effective in patients with lower PaO <sub>2</sub> /FiO <sub>2</sub> ratios (OR = 0.49, 95% CI = 0.28–0.87, P = 0.01) and those receiving TPE within 72 hours of disease onset (OR = 0.43, 95% CI = 0.23–0.82, P = 0.008). TPE was also associated with improvement in SOFA score (P < 0.001) and reduction in inflammatory biomarkers (P < 0.001).... |

ating 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 ex-

ternal sources beyond the limited chunks provided. This expanded knowledge base empowers the LLM to process complex information from various sources, leading 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 3 illustrates the complete workflow for this approach.

## 2.4 Data Processing and Context Analysis

The dataset, comprising main papers and their associated metadata, underwent preprocessing and context analysis which is shown in Fig2. 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 3 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.

```

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)

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

### 3 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.

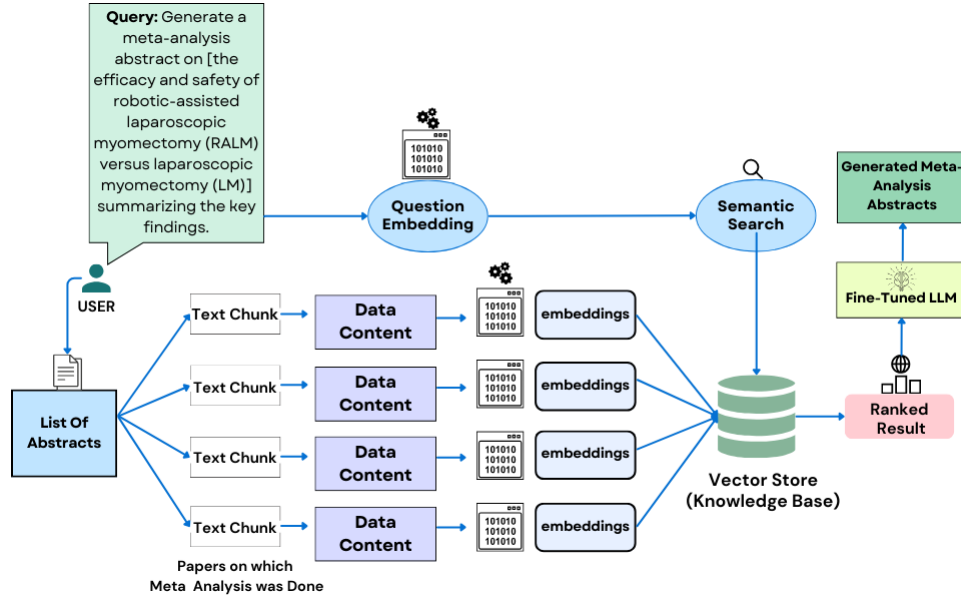


Fig. 3: The user will submit a query along with abstracts from several publications and a broad context. There will be several segments that are taken out of context. Every chunk's vector embeddings will be kept in a vector database. Using the embedded query in the vector base, a semantic search will 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.

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 various model outputs as shown in Fig4, 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.

|   | A ▼                | B   | C            | D              | E              | F              |
|---|--------------------|---|--------------|----------------|----------------|----------------|
| 1 | Meta_Paper         | Processed Output  | 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   | 2            | 2              | 2              |                |
|   |                    | Some key finding  |              |                |                |                |
|   |                    | <ul style="list-style-type: none"> <li>* Improved over</li> <li>* Reduced durat</li> <li>* Accelerated re</li> <li>* Mortality redu</li> </ul>                                |              |                |                |                |
| 3 |                    | Some numerical  |              |                |                |                |
|   |                    | <ul style="list-style-type: none"> <li>* Overall surviva</li> <li>* Duration of ho</li> <li>* CRS resolution</li> <li>* Mortality rate:</li> <li>* Hazard ratio fo</li> </ul> |              |                |                |                |
|   |                    | Please note that  |              |                |                | 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 pro  | 2            | 2              | 1              | 2              |
| 7 | significant differ | The provided ab   | 2            | 2              | 2              | 2              |
| 8 | (MRS) was used     | Based on the pro  | 2            | 2              | 2              |                |
|   |                    | Some numerical  |              |                |                | 2              |
| 9 | model to predict   | The present met   | 2            | 2              | 2              | 2              |

Fig. 4: Human evaluation spreadsheet showing meta-analysis abstract papers, LLM-generated outputs, evaluations of three individuals, and the overall evaluation.

### Evaluation with RAG

Fig. 5 displays a sample for evaluation for the Retrieval Augmented Generation (RAG) approach. Cosine-similarity 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. It is observed that fine-tuned models generated at most 84% similar meta-analysis than non-fine-tuned models.



|    | A  | B                            | C   | D      |
|----|--|------------------------------|---|--------|
| 1  | Ground Truth   | MP                           | Processed Output  | sim    |
| 2  | Background The available evidence about the possible             | MP: ObjectiveSeveral random  | This meta-analysis abstract summarizes the findings of several        | 0.7998 |
| 3  | Introduction Several studies have demonstrated that              | MP:Background: Depression    | In recent studies, interventions have been explored for improving     | 0.7519 |
| 4  |  |                              | 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 impact         |        |
|    | Background Glycemic control is vital to patient care, and        | MP:Purpose Diabetes mellitus | Overall, the results of the meta-analysis suggest that engaging       | 0.8048 |
| 5  |  |                              | Meta-analysis abstract:   |        |
|    | Objective This meta-analysis aimed to evaluate the effectiveness | MP: Background Recent in     | Pioglitazone use is associated with a reduced risk of Parkinson's     | 0.9038 |
| 6  | Background Recurrent hemarthrosis is one of the major            | MP:Background: The primary   | Myofascial therapy has been investigated as an effective treatment    | 0.8022 |
| 7  | Background Curcumin, a natural polyphenolic compound             | MP:Polycystic ovary syndrome | A systematic review and meta-analysis of randomized controlled        | 0.8189 |
| 8  | Purpose Despite the high prevalence of postdialysis fluid        | MP:Background The aetiology  | This meta-analysis abstract summarizes the key findings of several    | 0.7565 |
| 9  | Purpose Research has shown that prone positioning (PP)           | MP:OBJECTIVES: To determine  | Prone positioning (PP) was found to reduce mortality in patients      | 0.8481 |
| 10 | Handhelds (e.g., cell phones, tablets) are promising for         | MP:The importance of mobile  | Mobile technologies have become increasingly important in healthcare  | 0.873  |
| 11 | Background Circulating concentrations of homocysteine            | MP:Objectives To investigate | A meta-analysis was conducted to investigate the relationship between | 0.805  |
| 12 |  |                              | Meta-analysis Abstract:   |        |
|    | Introduction Previous randomized controlled trials (RCTs)        | MP: Background Vitamin D     | A systematic review and meta-analysis of randomized clinical          | 0.8601 |
|    |  |                              | Functional neuroimaging and Dispositional Variables for Clinical      |        |
|    |  |                              | =====   |        |

Fig. 5: 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.