Zero-Shot Information Extraction for Clinical Meta-Analysis using Large Language Models

David Kartchner and **Irfan Al-Hussaini** and **Olivia Kronick**
Georgia Institute of Technology
{david.kartchner, alhussaini.irfan, okronick3}@gatech.edu

Selvi Ramalingam
Emory University
sramal3@emory.edu

Cassie Mitchell
Georgia Institute of Technology
cassie.mitchell@bme.gatech.edu

Abstract

Meta-analysis of randomized clinical trials (RCTs) plays a crucial role in evidence-based medicine but can be labor-intensive and error-prone. This study explores the use of large language models to enhance the efficiency of aggregating results from randomized clinical trials (RCTs) at scale. We perform a detailed comparison of the performance of these models in zero-shot prompt-based information extraction from a diverse set of RCTs to traditional manual annotation methods. We analyze the results for two different meta-analyses aimed at drug repurposing in cancer therapy and pharmacovigilance in chronic myeloid leukemia. Our findings reveal that the best model for the two demonstrated tasks, ChatGPT, can generally extract correct information and identify when the desired information is missing from an article. We additionally conduct a systematic error analysis, documenting the prevalence of diverse error types encountered during the process of prompt-based information extraction.

1 Introduction

Meta-analysis is a statistical method that combines and analyzes data from multiple studies to obtain an overall effect size or estimate of treatment effect. It is widely used in healthcare research, particularly in clinical trials, to provide a comprehensive and robust summary of the available evidence (Gopalakrishnan and Ganeshkumar, 2013).

The importance of meta-analysis lies in its ability to increase statistical power and reduce bias, thereby improving the accuracy and reliability of the findings. Meta-analysis also allows for the identification of important subgroups of patients and provides insights into the potential sources of heterogeneity in the results of different studies (Sedgwick, 2013; Song et al., 2001).

In the context of clinical trials, meta-analysis plays a crucial role in the evaluation of new treatments and interventions (Heys et al., 1999; Al-Karawi and Jubair, 2016; Henna et al., 2004; Boulé et al., 2001). By combining data from multiple studies, researchers can obtain a more precise estimate of the effectiveness of treatment and identify any potential adverse effects.

While clinical meta-analysis is essential to establishing guidelines for clinical best-practice (Stangl and Berry, 2000), curating data is time-consuming for medical professionals. A recent survey found that most clinical meta-analyses require 6-10 months of data gathering and analysis for 5 individuals (Borah et al., 2017), which does not account for research development time spent before registering meta-analyses on PROSPERO registry of systematic reviews (Booth et al., 2012). Moreover, most meta-analyses seek to answer very targeted questions about specific diseases or drugs, making it difficult to adapt existing datasets for the automatic or semi-automatic extraction of needed data.

A recent review (Wornow et al., 2023) examined founation models/large language models (LLMs), such as ChatGPT, and opined that while there is evidence that clinical foundation models improve accuracy, there has been minimal work to validate other potential benefits, such as reducing data labeling burden, enabling new clinical applications, and offering novel human-AI interfaces. However, foundation models also present significant risks, including data privacy and security concerns, interpretability challenges, high up-front costs, and biases (Wornow et al., 2023).

This paper advocates for the development of new evaluation tasks, metrics, and datasets to better understand how foundation models perform on clinical tasks. To this end, it contributes the following to the development and use of foundation models in clinical and biomedical research:

- We present, to our knowledge, the first detailed evaluation of how well generative foundation models perform for extracting informa-

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 396–405

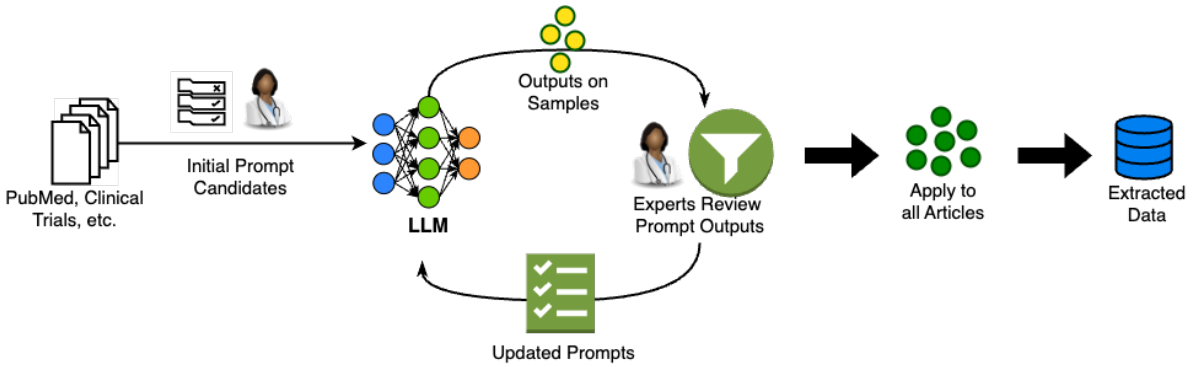

Figure 1: Pipeline for zero-shot data extraction for Meta-Analysis using LLMs.

tion for clinical meta-analysis.

- We detail prompting and post-processing strategies that can improve the performance of prompt-based extraction and normalization of clinical information.
- We present an error analysis systematically cataloging the prevalence of various types of errors encountered during prompt-based information extraction and normalization.

2 Related Work

Recent surveys and case studies have shown that natural language processing can powerfully accelerate the identification of candidates for drug repurposing (Subramanian et al., 2020) and the identification of adverse effects in pharmacovigilance efforts (Bhatnagar et al., 2022). Substantial work in clinical evidence extraction has been in the collection of patients, interventions, comparators, and outcomes (PICO) from clinical trials. Nye et al. (2018) and Zlabinger et al. (2020) both present a corpora with PICO annotations aimed at tagging the high-level characteristics of clinical trials. Biomedical transformer models are currently state-of-the-art on PICO tagging (Alrowili and Shanker, 2021; Yasunaga et al., 2022; Tinn et al., 2021). Additional improvements can be made by augmenting PICO data with weak supervision from external dictionaries and hand-crafted rules (Dhrangadhariya and Müller, 2023).

A number of works have built models to improve the screening and extracting evidence from clinical trials for meta-analysis and systematic review. Wallace et al. (2021) automatically constructs a dataset focused on summarizing systematic reviews using the "Conclusion" section of structured abstracts, and this work was extended in a recent shared task on multi-document summarization for scientific literature reviews (Wang et al., 2022). Al-Hussaini et al. (2022) takes a different approach by producing an end-to-end system for screening and summarizing individual clinical cohort studies based on PICO elements and clinically-relevant details annotated by epidemiologists during data curation. Kang et al. (2023) creates a novel, hierarchical framework for structuring evidence in clinical trials for better usability and understanding.

While existing work offers many valuable resources for evidence extraction across a diverse array of tasks, it is limited to specific, narrowly-scoped tasks and requires substantial annotation efforts to work effectively. In contrast, many critical clinical meta-analyses answer very specific clinical questions which are out of the scope of existing data and require manual curation. This work differs from existing work by evaluating the effectiveness of foundation NLP models for zero-shot extraction of information for highly-specific clinical questions, enabling a much broader range of applications.

3 Problem setup and methodology

Two prior works facilitated the development of this zero-shot information extraction work, the Remedy Database (Emory, 2023) and CML dataset registered on PROSPERO for an ongoing meta-analysis (Kronick et al., 2023). Each was created through manual annotations of clinical trial and cohort study articles. Lists of the data columns included in each dataset are included in Tables 1 and 2.

3.1 Remedy Database

Repurposing drugs is the process of identifying new therapeutic uses for existing drugs that have

Data Field	Definition
Study Type	Study characteristics such as Phase II, open-label, placebo-controlled, etc.
Cancer Type	What type of cancer did the study focus on?
Cancer Stage	What stage(s) of cancer did the study focus on?
Drug Name	What are the (non-cancer) drugs studied in this clinical trial?
Treatment Timing	When was treatment given relative to other standard-of-care (SOC) treatment?
Dosage	What is the dosage of the drug(s) used in the study?
Concurrent SOC	What is the standard of care (SOC) treatment is given to cancer patients in this study?
Sample Size	How many patients were enrolled in the study?
Summary	Template-baased summary capturing information on intervention and comparator groups, study outcomes, and the authors' conclusions.

Table 1: Data fields of ReMedy dataset

already been approved for other indications. Clinical trial articles provide valuable information on the safety and efficacy of new treatments and interventions. However, identifying and extracting critical information from clinical trial articles is a time-consuming and tedious process. ReMedy cancer database was built to address this issue (Emory, 2023). It provides a comprehensive and standardized source of information for clinical trials, observational studies, and case reports data for repurposed drugs for cancer, thereby making the repurposed drug data easily accessible to physicians, patients, and potential investigators. In this article, we have focused mainly on extracting information from clinical trial articles.

The current data extraction process is mostly manual. The data extraction is divided into three distinct categories: (i) Article identification, (ii) Extraction of clinical data, and (iii) Creation of structured article summary. Identification of the articles is done manually by searching for repurposed drug articles in databases such as PubMed. The screening process ensures that the identified articles fit the inclusion criteria, such as study design, intervention, and outcome measures.

Articles were included based on study type, interventional drug, and outcome measures. All articles were required to evaluate the effect of a non-cancer drug on cancer patient outcomes to be included. Every clinical trial article entry in the Remedy database is reviewed at least twice, and changes are made to ensure accuracy.

The data elements included in the ReMedy database are Pubmed ID (PMID), year of publication, study type, cancer stage, cancer type, cancer sub-type, drug name, drug category, treatment timing, the dosage of the drug, the concurrent standard of care, number of patients enrolled in the study, clinical trial outcomes (primary and secondary) and author's conclusion. Data was curated by students pursuing a masters of public health (MPH) degree and takes between 60-90 minutes for an average abstract.

3.2 CML Database

Chronic myeloid leukemia, or CML, is a relatively rare form of leukemia characterized by the presence of a Philadelphia chromosome, which results from the fusion of BCR and ABL1 genes (Jabbour and Kantarjian, 2018). Though once highly lethal, the development of a tyrosine-kinase inhibitor (TKI) drugs dramatically improved long-term CML survival (Minciacchi et al., 2021). This dataset was developed for a meta-analysis to examine the hematological adverse events (HAEs) associated with TKI use. Examined HAEs include anemia, thrombocytopenia, neutropenia, aplastic anemia, pancytopenia, and myelosuppression. Data extraction followed a procedure close to the one described in (Mohanavelu et al., 2021). All articles were selected to include at least one TKI and were filtered to exclude combination therapies with non-TKI drugs. Articles were identified using searches in *Pubmed* and *ClinicalTrials.gov*.

The following data fields are included for each article: Source (PMID or National Clinical Trials (NCT) number), TKI name, number of patients under treatment, number of patients experiencing included HAE, condition grade (if available), and

Data Field	Definition
All TKIs	List of tyrosine kinase inhibitors administered to patients in the study/trial
All Phases	List of phases of chronic myeloid leukemia of patients at time of treatment.
All HAEs	List of all hematological adverse events experienced by patients in the study/trial.
HAE Grade	Whether or not the trial lists all HAEs or only severe HAEs (i.e. grade 3 or 4).
Num. Treated	For each combination of (TKI, CML phase) in the trial, how many total patients were treated?
HAE Counts	For each combination of (TKI, CML phase) in the trial, how many patients experienced a particular HAE?

Table 2: Data columns of CML dataset

CML phase (if available). Quality control was conducted by an independent team on the extracted data. The total number of articles listing each HAE as follows: aplastic anemia:3; anemia: 50; neutropenia: 55; thrombocytopenia: 61; myelosuppression: 13; and pancytopenia:16.

4 Methods

4.1 Models

Our selection of models sought to balance performance and cost for applicability to an extensive range of researchers. Balancing these criteria, we selected two models, GPT 3.5 Turbo (Ouyang et al., 2022) (also known as ChatGPT) and Together's GPT-JT (Together, 2023) for use. ChatGPT was chosen due to its well-known human-like performance across a wide range of tasks and its ability to follow detailed instructions. GPT-JT, an open, fine-tuned extension of GPT-J (Wang and Komatsuzaki, 2021) has also shown state-of-the-art performance across a range of prompt-based tasks despite using up to two orders of magnitude fewer parameters (6B vs. up to 530B) (Together, 2022). All outputs for GPT-JT were obtained via the Manifest wrapper for the API (Orr, 2022).

4.2 Prompt Creation and Tuning

Prompts were created and refined using an iterative, human-in-the-loop creation process. Initial prompts were created for each dataset in collaboration with one of the original curators of the database. Prompts for multiple-choice or select-all-that-apply study characteristics were often (but not always) provided with a list of possible values to guide model outputs.

The "article summary" column from ReMedy follows a template to ensure that human curators gather all relevant epidemiological data for a drug-repurposing meta-analysis. To produce these summaries, we provided models with a similar template, augmented with instructions on what data to put in each field. The full text of each summary was generated in a single shot for each article.

The CML dataset required the extraction of quantitative information about each hematological adverse event (HAE) listed in each paper. This extraction was needed for each stage of CML and each TKI analyzed in the study. We extracted this information by first prompting the model to provide a list of all TKIs, CML phases, and HAE listed in each study. After normalizing the outputs for each category (see below), we iterated through each (TKI, phase, HAE) model identified in the paper and extracted the count and/or percentage of each adverse event. We additionally attempted to extract this information using a one-shot templated summary but found the model output too inconsistent to reliably parse the desired quantitative data.

A list of prompts used for ReMedy are given in Table 3 and prompts for CML are given in Table 4.

4.3 Data Extraction and Postprocessing

Outputs from generative LLMs were often noisy and required significant normalization to be usable. For example, outputs of GPT-JT consistently began generating additional questions after each output was extracted, so outputs had to be split by newline character, and everything after the first newline was thrown away. Outputs of categorical columns (single and multi-label) were normalized to acceptable values using fuzzy string matching with the acceptable values. All matches within an appropriate threshold (usually 80% or higher) were replaced with the matched canonical value, while those be-

Data Field	Prompt
Study Type	What type of study is this? I define study type as "phase 0", "phase 1", "phase 2", "phase 3", "phase 4", "randomized", "double blind", "open label", "placebo controlled", "pilot studies". Please list all applicable study types.
Cancer Type	What type of cancer(disease) did the study focus on?
Cancer Stage	What stage of cancer did the study focus on? I define cancer stage as "Early (stages 0 - 3)", "Advanced (stages 3-4)", "Metastatic (Stage 4)", "All stages (0-4)"
Drug Name	What is/are the name(s) of the drug(s) used in the study?
Treatment Timing	What is the treatment timing? I define treatment timing as "adjuvant", "continuous", "maintenance", "neo-adjuvant", "palliative", "peri-operative", "post-diagnosis", "primary therapy", and "secondary therapy".
Dosage	What is the dosage of the drug(s) used in the study? Please list values in a dictionary as {drug_name: dosage}
Concurrent SOC	What is the standard of care? I define standard of care as "anti-angiogenic", "check point inhibitors", "chemo-radiation", "immune therapy", "radiation", and "targeted therapy"
Sample Size	How many patients were enrolled in the study? Please give only the number and no other words.
Summary	Write a summary the study using the following guidelines/template. Fill out a template that includes the following features: Include the disease name, drug name, type of clinical trial for the first sentence. Then summarize the abstract by including PICO (Population, Intervention,comparison (control & intervention) and outcome. Also include the standard of care that is used. Finally include author's conclusion. Below is a sample template, with values you should insert in brackets []. "A [clinical trial phase] clinical trial evaluating the effects of [drug(s)] in patients with [cancer type(s)] cancer. Disease: [cancer types] cancer Population: [n] patients with histopathology in [location] Intervention (n=[num patients in intervention group]): [dosage details] Control (n=[n patients in control/standard of care group]): [standard of care treatment. should not include non-cancer drug being investigated] Concurrent treatment: [treatment type, e.g. chemotherapy (detailed or not?)] Primary outcomes: a)[study endpoint/outcome metric], [value (intervention vs control)], [confidence interval], [p-value] b) Secondary outcomes: a) [study endpoint/outcome metric], [value (intervention vs control)], [confidence interval], [p-value] b) The authors conclude: [single sentence direct quote from author conclusions]

Table 3: Prompts used for ReMedy data extraction

Data Field	Prompt
All TKIs	Please list all tyrosine kinase inhibitors studied. List only the TKI name(s) delimited by commas. Do not return anything other than the TKI name(s).\nChoices: ['imatinib', 'nilotinib', 'dasatinib', 'radotinib', 'ruxolitinib', 'bosutinib', 'tipifarnib', 'asciminib', 'ponatinib', 'bosutonib']
All Phases	What phase(s) of chronic myeloid leukemia (CML) did patients in the study have? List all that apply. Do not return anything other than the CML phase(s).\nChoices: ["chronic", "accelerated", "blast"]\nIf not specified, reply "n/a".
All HAEs	What hematological adverse events were experienced by patients in the study? Select all that apply. Do not return anything other than the adverse events.\nChoices: ["anemia", "pancytopenia", "myelosuppression", "aplastic anemia", "neutropenia", "thrombocytopenia"]\nIf none mentioned, reply "n/a"\n
HAE Grade	Does this study list all grades of adverse events or only severe (i.e. grade 3 or 4) adverse events? If all grades, please respond "all" and if severe only, please respond "severe". If no adverse events are mentioned, reply "n/a"
Num. Treated	How many patients with phase phase CML were treated with tki? Please return a single integer and nothing else. If no patients with phase phase CML were treated with tki were specified, return "n/a"
HAE Counts	How many {tki} treated patients with {phase} phase CML experienced {ade}? Please return a single integer. If the number of patients with {ade} is not listed, reply "n/a"

Table 4: Prompts used for CML data extraction

	Unfiltered		Filtered		
	ChatGPT	**GPT-JT-6B**	**ChatGPT**	**GPT-JT-6B**	**Metric**
Concurrent SOC	0.135	**0.192**	0.149	**0.201**	Accuracy
Cancer Type	**0.897**	0.510	**0.904**	0.518	Accuracy
Treatment Type	0.235	**0.564**	0.248	**0.549**	Accuracy
Num. Patients	**0.719**	0.638	**0.720**	0.634	Accuracy
Cancer Stage	**0.403**	0.224	**0.408**	0.215	Accuracy
Dosage	**0.461**	0.083	**0.472**	0.077	Fuzzy Acc
Study Subtype	**0.584**	0.375	**0.604**	0.382	Jaccard
Summary (p)	**0.469**	0.197	**0.496**	0.205	Precision
Summary (r)	**0.420**	0.201	**0.431**	0.201	Recall
Summary (f1)	**0.421**	0.188	**0.441**	0.193	F1

Table 5: Results of LLM information extraction on Remedy database

low the threshold were marked as incorrect.

"Cancer Type" and "Concurent Standard-of-Care Treatment" columns in Remedy were not given candidate values in the multiple choice list during initial LLM labeling, so we asked each LLM to generate a mapping from the free-text values to their normalized forms. ChatGPT was able to to map the free-form values in these columns to the candidates in the multiple choice list for over 95% of values, while GPT-JT correctly normalized outputs less than 10% of the time. When comparing the outputs of these columns, we compare the self-normalized version of ChatGPT's output and the unnormalized version of GPT-JT's output, giving each model the highest respective performance.

4.4 Hyperparameters

ChatGPT was run with `p=0` and maximum tokens set to 30 tokens or less for all fields except the summary. All other ChatGPT parameters were set to default values. Manifest parameters were set to `top_p=0.9`. We additionally set `top_k=40` for ReMedy non-summary fields, and `top_k=1` for ReMedy summary + all CML fields. All summaries were generated using a maximum of 256 tokens, occasionally resulting in summary truncation.

4.5 Evaluation Metrics

4.5.1 Classification

Each dataset in question produced diverse data types that require different evaluation strategies. All single-label classification columns were evaluated using accuracy. For multi-label classification, where each model outputs a set of characteristics (e.g., *Study Subtype* in Remedy), we compare the set of characteristics labeled by the foundation model with human-annotated set using Jaccard similarity:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

4.5.2 Quantitative Information Extraction

Integer columns (e.g., number of patients, side-effect counts) were evaluated using accuracy, measured as the mean frequency of an exact match between the human-annotated integer and the output of the foundation model. For columns that measure percentages, we considered an answer "correct" if the output of the foundation model and the human annotation differ by no more than 1% to account for rounding error.

4.5.3 Short Answer

Some fields list a short, free-form text answer (e.g., dosage). For these fields, multiple equivalent answers could be acceptable. For example, the dosages "40mg BID" and "40 mg twice daily" are semantically equivalent. These differences are common as clinicians prefer shorter acronyms such as "BID" whereas LLMs prefer more colloquial explanations such as "twice daily". We compare LLM outputs' similarity with gold answers in these fields using fuzzy string matching, with cutoffs chosen based on human inspection of included/excluded matches. We then calculate "fuzzy accuracy" as the mean of how often gold and LLM responses were above the required similarity threshold.

4.5.4 Summarization

We compare generated and human summaries using Rouge (Lin, 2004). Specifically, we use the Rouge-L metric, which analyzes the length of the longest common subsequence (LCS) of the generated and reference summaries. Let m and n be

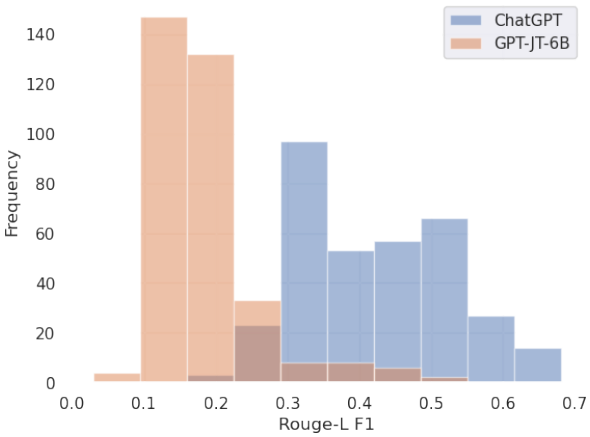

Figure 2: ReMedy Summary Rouge-L F1 score distribution for each model.

the lengths of the reference and generated summaries, respectively. Then the Rouge-L precision (P_{lcs}), recall (R_{lcs}), and F-measure (F_{lcs}) scores are computed as follows:

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$
$$P_{lcs} = \frac{LCS(X,Y)}{n}$$
$$F_{lcs} = \frac{\left(1+\beta^2\right) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

where $\beta = \frac{P_{lcs}}{R_{lcs}}$.

5 Results

5.1 Remedy

The results of the models on ReMedy are shown in Table 5. These results show a wide range in accuracy, with the extraction of cancer type and sample size being the most accurate and identification of the concurrent standard-of-care treatment being the least accurate.

In general, we observe that ChatGPT dramatically outperforms GPT-JT for most data fields. The largest gains occur in the "Dosage" and "Summary" fields. This likely occurs because ChatGPT's training objective provides more explicit optimization for summarization and formatting, both of which are critical to performing these tasks correctly. GPT-JT's ability to summarize is notably poor, only performing 3% above the score given by the template given in the prompt, $(P, R, F1) = (0.206, 0.158, 0.170)$. A comparison of the distribution of Rouge-L F1 scores for ChatGPT and GPT-JT is shown in Figure 2.

	ChatGPT	GPT-JT-6B	Metric
TKIs	**0.870**	0.448	Jaccard
CML Phases	0.919	**0.932**	Jaccard
HAEs	**0.500**	0.466	Jaccard
HAE Grade	0.044	**0.128**	Accuracy
Num. Treated	**0.444**	0.154	Accuracy
Anemia	0.244	**0.308**	+/-1 Accuracy
Neutropenia	0.133	**0.179**	+/-1 Accuracy
Thrombocytopenia	0.133	**0.128**	+/-1 Accuracy
Pancytopeia	0.956	**0.974**	+/-1 Accuracy
Myelosuppression	1.000	1.000	+/-1 Accuracy
Aplastic Anemia	1.000	1.000	+/-1 Accuracy

Table 6: Results of LLM information extraction on CML dataset. +\- Accuracy counts numerical matches that are off-by-1 to account for rounding error.

5.2 CML

The results of the models on the CML dataset are shown in Table 6. We observe that ChatGPT outperforms GPT-JT on identifying the number of patients treated and the TKIs used in the study.

One interesting result is that accuracy on pancytopenia, myelosuppression, and aplastic anemia are high. This is because these HAEs are very rarely observed in the data, and the models correctly predict them as N/A. This is a promising result showing that models are capable of refusing to make up values when the data is not present for extraction.

5.3 Error Analysis

We perform an error analysis of the mistakes made by models when generatively extracting data. We found that errors extracting specific data elements generally fell into the following categories:

Excessive Verbosity Despite prompts explicitly designed to limit verbosity (e.g., "...Please list only the tki name(s) delimited by commas", "...Please return a single integer. If the answer is not found, return 'n/a' "), models frequently output extraneous text trying to explain their answers, particularly ChatGPT. Excess explanations turned single-word answers into complete sentences, added undesired (but grammatically correct) punctuation, or explained that the data was not found in the quoted text rather than simply returning "n/a". This verbosity was the leading cause of postprocessing needed to make model output usable.

Data not in Abstract Each of the databases used in this study were curated the full text of research articles. However, our models only examined data present in the abstracts of such articles. While all

needed information was sometimes present in the abstract alone, specific quantitative results were often present only in the full text of the research article, particularly for the CML dataset. In these cases, failure to agree with the curated data reflects a correct assessment by the model that information is lacking in the text it was provided. In the Remedy dataset, the study subtype was sometimes not present in the abstract, which ChatGPT correctly detected and returned a response indicating its absence. For 33% of the articles, ChatGPT detected all study subtypes accurately from the abstract.

Incomplete Gold-standard Data In select cases, LLM output revealed human errors in the curation of the original databases. An analysis of the 30 lowest-scoring abstracts for document summarization revealed that 20 had incomplete human summaries, 3 of which had no human summary. Two additional articles had been removed from ReMedy and/or referenced by a different PMID in the database. This indicates that text summarization results are overly conservative in estimating model performance, especially on the lowest-scoring portion of the data. We report the results after removing the subset with errors under the "Filtered" tab of the ReMedy results.

Hallucinations Models occasionally hallucinate false information from studies. The most common hallucinations from ChatGPT in Remedy were incorrect numbers of patients in treatment/control groups and the hallucination of a control group when none was present. This most commonly occurs when the group sizes are not explicitly given in the paper, in which case ChatGPT assumed that the population was split evenly between treatment and control groups. When the disease subtype was absent in the abstract, ChatGPT had the propensity to guess the study subtype instead of indicating its absence. Hallucinations from GPT-JT showed a higher propensity to seek to answer the question with fabricated numbers when the actual answer was missing from the study.

6 Conclusion

In conclusion, this study demonstrates the potential of large language models in enhancing the efficiency of clinical meta-analyses of randomized clinical trials. By comparing its performance with traditional manual annotation methods, we provide valuable insights into the advantages and challenges of implementing AI-based solutions in evidence-based medicine. The results of our research indicate that LLMs can contribute to more streamlined, transparent, and reproducible results in clinical research. It also reveals that they still make significant errors and should be used cautiously with additional quality checks when used as a tool to extract data from clinical research.

Ethics Statement

Using large language models in a clinical domain has inherent risks. As demonstrated in this paper, LLMs sometimes hallucinate and fabricate false answers to questions posed about research articles. If done at scale, these extraction errors could propagate to downstream analyses, potentially leading to false conclusions. While LLMs may be able to significantly speed the process of human data curation and even help in detecting errors, they still require manual verification of results to ensure high data quality.

References

Irfan Al-Hussaini, Davi Nakajima An, Albert J. Lee, Sarah Bi, and Cassie S. Mitchell. 2022. Ccs explorer: Relevance prediction, extractive summarization, and named entity recognition from clinical cohort studies. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5173–5181.

Dalia Al-Karawi and Luqman Jubair. 2016. Bright light therapy for nonseasonal depression: meta-analysis of clinical trials. *Journal of affective disorders*, 198:64–71.

Sultan Alrowili and Vijay Shanker. 2021. BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227, Online. Association for Computational Linguistics.

Roopal Bhatnagar, Sakshi Sardar, Maedeh Beheshti, and Jagdeep T Podichetty. 2022. How can natural language processing help model informed drug development?: a review. *JAMIA open*, 5(2):ooac043.

Alison Booth, Mike Clarke, Gordon Dooley, Davina Ghersi, David Moher, Mark Petticrew, and Lesley Stewart. 2012. The nuts and bolts of prospero: an international prospective register of systematic reviews. *Systematic reviews*, 1(1):1–9.

Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7(2):e012545.

Normand G Boulé, Elizabeth Haddad, Glen P Kenny, George A Wells, and Ronald J Sigal. 2001. Effects of exercise on glycemic control and body mass in type 2 diabetes mellitus: a meta-analysis of controlled clinical trials. *Jama*, 286(10):1218–1227.

Anjani Dhrangadhariya and Henning Müller. 2023. Not so weak pico: leveraging weak supervision for participants, interventions, and outcomes recognition for systematic review automation. *JAMIA open*, 6(1):ooac107.

Emory. 2023. Remedy (repurposed medicines)-cancer database.

S Gopalakrishnan and P Ganeshkumar. 2013. Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. *Journal of family medicine and primary care*, 2(1):9.

Márcia Riromi Henna, Rozemeire GM Del Nero, Cristina Zugaiar S. Sampaio, Álvaro Nagib Atallah, Sérgio Tomaz Schettini, Aldemar Araújo Castro, and Bernardo Garcia de Oliveira Soares. 2004. Hormonal cryptorchidism therapy: systematic review with metanalysis of randomized clinical trials. *Pediatric surgery international*, 20:357–359.

Steven D Heys, Leslie G Walker, Ian Smith, and Oleg Eremin. 1999. Enteral nutritional supplementation with key nutrients in patients with critical illness and cancer: a meta-analysis of randomized controlled clinical trials. *Annals of surgery*, 229(4):467.

Elias Jabbour and Hagop Kantarjian. 2018. Chronic myeloid leukemia: 2018 update on diagnosis, therapy and monitoring. *American journal of hematology*, 93(3):442–459.

Tian Kang, Yingcheng Sun, Jae Hyun Kim, Casey Ta, Adler Perotte, Kayla Schiffer, Mutong Wu, Yang Zhao, Nour Moustafa-Fahmy, Yifan Peng, and Chunhua Weng. 2023. EvidenceMap: a three-level knowledge representation for medical evidence computation and comprehension. *Journal of the American Medical Informatics Association*. Ocad036.

Olivia Kronick, Xinyu Chen, Nidhi Mehra, Armon Varmeziar, Rachel Fisher, Vamsi Kota, and Cassie Mitchell. 2023. Frequency of hematological adverse events in chronic myeloid leukemia patients on tyrosine kinase inhibitor therapy: A systematic review with meta-analysis.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Valentina R Minciacchi, Rahul Kumar, and Daniela S Krause. 2021. Chronic myeloid leukemia: a model disease of the past, present and future. *Cells*, 10(1):117.

Prahathishree Mohanavelu, Mira Mutnick, Nidhi Mehra, Brandon White, Sparsh Kudrimoti, Kaci Hernandez Kluesner, Xinyu Chen, Tim Nguyen, Elaina Horlander, Helena Thenot, et al. 2021. Meta-analysis of gastrointestinal adverse events from tyrosine kinase inhibitors for chronic myeloid leukemia. *Cancers*, 13(7):1643.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Laurel Orr. 2022. Manifest. `https://github.com/HazyResearch/manifest`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Philip Sedgwick. 2013. Meta-analyses: heterogeneity and subgroup analysis. *Bmj*, 346.

Fujian Song, Trevor A Sheldon, Alex J Sutton, Keith R Abrams, and David R Jones. 2001. Methods for exploring heterogeneity in meta-analysis. *Evaluation & the health professions*, 24(2):126–151.

Dalene Stangl and Donald A Berry. 2000. *Meta-analysis in medicine and health policy*. CRC Press.

Shivashankar Subramanian, Ioana Baldini, Sushma Ravichandran, Dmitriy A Katz-Rogozhnikov, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Kush R Varshney, Annmarie Wang, Pradeep Mangalath, and Laura B Kleiman. 2020. A natural language processing system for extracting evidence of drug repurposing from scientific publications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13369–13381.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Fine-tuning large neural language models for biomedical natural language processing. *arXiv preprint arXiv:2112.07869*.

Together. 2022. Releasing gpt-jt powered by open-source ai.

Together. 2023. Gpt-jt-6b.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605.

Ben Wang and Aran Komatsuzaki. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 175–180, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of clinical foundation models: A survey of large language models and foundation models for emrs. *arXiv preprint arXiv:2303.12961*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.

Markus Zlabinger, Marta Sabou, Sebastian Hofstätter, and Allan Hanbury. 2020. Effective crowd-annotation of participants, interventions, and outcomes in the text of clinical trial reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3064–3074, Online. Association for Computational Linguistics.