

Automatically Extracting Numerical Results from Randomized Controlled Trials with Large Language Models

Hye Sun Yun

*Northeastern University
Boston, MA, USA*

YUN.HY@NORTHEASTERN.EDU

David Pogrebitskiy

*Northeastern University
Boston, MA, USA*

POGREBITSKIY.D@NORTHEASTERN.EDU

Iain J. Marshall

*King's College London
London, UK*

IAIN.MARSHALL@KCL.AC.UK

Byron C. Wallace

*Northeastern University
Boston, MA, USA*

B.WALLACE@NORTHEASTERN.EDU

Abstract

Meta-analyses statistically aggregate the findings of different randomized controlled trials (RCTs) to assess treatment effectiveness. Because this yields robust estimates of treatment effectiveness, results from meta-analyses are considered the strongest form of evidence. However, rigorous evidence syntheses are time-consuming and labor-intensive, requiring manual extraction of data from individual trials to be synthesized. Ideally, language technologies would permit fully automatic meta-analysis, on demand. This requires accurately extracting numerical results from individual trials, which has been beyond the capabilities of natural language processing (NLP) models to date. In this work, we evaluate whether modern large language models (LLMs) can reliably perform this task. We annotate (and release) a modest but granular evaluation dataset of clinical trial reports with numerical findings attached to interventions, comparators, and outcomes. Using this dataset, we evaluate the performance of seven LLMs applied zero-shot for the task of conditionally extracting numerical findings from trial reports. We find that massive LLMs that can accommodate lengthy inputs are tantalizingly close to realizing fully automatic meta-analysis, especially for dichotomous (binary) outcomes (e.g., mortality). However, LLMs—including ones trained on biomedical texts—perform poorly when the outcome measures are complex and tallying the results requires inference. This work charts a path toward fully automatic meta-analysis of RCTs via LLMs, while also highlighting the limitations of existing models for this aim.

1. Introduction

Quantitative measures of comparative treatment effectiveness are reported primarily in unstructured (natural language) published articles that describe the design, protocol, and results of randomized controlled trials (RCTs). Individual trial results are noisy and often biased, motivating the need for rigorous *statistical meta-analysis* of all trials of a particular

treatment to produce a robust estimate of efficacy (Lau et al., 1995; Borenstein et al., 2021).¹ Estimates from meta-analyses of primary findings are considered one of the highest forms of evidence in medicine (Murad et al., 2016).

However, conducting a meta-analysis requires manually extracting from individual articles the data elements necessary for synthesis, i.e., numbers required to calculate metrics of interest—for example, odds ratios or mean differences for dichotomous and continuous outcomes, respectively—and associated variances. This time-consuming but critical step still must be manually performed, which precludes a fully automated “on-demand” meta-analysis of all evidence relevant to a given clinical question. Prior work (Yun et al., 2023) has reported that domain experts who conduct systematic reviews and meta-analyses view data extraction as a key part of the process which may be amenable to automation via LLMs.

In this work, we ask: **Are modern LLMs sufficiently capable of data extraction to permit accurate, fully automated meta-analysis?** To answer this question empirically, we annotate modest but granular validation and test sets of clinical trial reports with numerical findings attached to interventions, comparators, and outcomes (ICO triplets). We use these annotations to evaluate a wide range of LLMs—including accessible, smaller models and closed, massive models—for their ability to infer structured numerical findings for specific comparisons of interest. That is, we assess whether LLMs can reliably extract from trial reports the data necessary for statistical meta-analysis.

We find that LLMs can permit accurate automated meta-analyses with some success. Unsurprisingly, massive LLMs with large input context windows such as GPT-4 outperform smaller, open-source models at extracting binary (dichotomous) outcomes. For extracting continuous outcomes, LLMs perform comparatively poorly, especially in settings where there are multiple similar outcome measures reported in a trial report. The takeaway from this work is that modern LLMs offer a promising path toward fully automatic meta-analysis, but further improvements are needed before this will be reliable. We hope the data we release with this effort will support future work in this direction.

Generalizable Insights about Machine Learning in the Context of Healthcare

LLMs have realized remarkable performance across a diverse range of NLP tasks in recent years, motivating several efforts to investigate the use of LLMs for healthcare specifically (Singhal et al., 2023; Lehman et al., 2023; Naik et al., 2023; Wadhwa et al., 2023b). Here we provide empirical insights concerning the use of such models to fully automate meta-analysis of clinical trial results; the results also have implications for related numerical data extraction tasks in healthcare, and how well LLMs are likely to be able to perform them. More concretely, the main generalizable contributions this work offers are:

1. We release an evaluation dataset containing extensive annotations for the task of extracting numerical clinical findings for a given intervention, comparator, and outcome (ICO triplets) necessary for conducting meta-analyses.
2. We report the quantitative and qualitative results evaluating modern LLMs on extracting numerical findings from RCTs using the annotated dataset. Furthermore,

1. These are essentially weighted averages of comparative effect estimates reported in trials, where weights are inverse to reported variances.

from research abstracts and parsed the numeric outcomes for statistical analysis. However, this work used an encoder-only model (BERT) which is considered small by modern standards and required explicit supervision for the task. Other recent work (Shamsabadi et al., 2024) has shown that LLMs can be fine-tuned to perform information extraction tasks from scientific and biomedical literature in general, though these efforts were not focused on supporting meta-analysis.

More recently, several works have explored extracting experimental findings of clinical trials using LLMs in zero-shot (a setup in which a model observes tasks or data that it has not explicitly seen during training) (Khraisha et al., 2023; Kartchner et al., 2023). However, these works often evaluated a couple of massive, closed LLMs and did not focus on numerical findings. Naik et al. (2023) investigated numerical information extraction using several LLMs. However, theirs was a general extraction scheme intended to support a range of downstream functionalities; by contrast, we focus more narrowly on how well LLMs can extract the numerical data necessary for meta-analysis, conditioned on a specific intervention, comparator, and outcome of interest.

The few focused efforts that have evaluated LLMs for data extraction from full-text articles for synthesis (Gartlehner et al., 2023; Sun et al., 2024; Reason et al., 2024) have used very small convenience samples of articles and evaluated only closed models, providing a somewhat limited view of current model capabilities generally. In our work, we construct an evaluation dataset of several hundred samples with both abstracts and results sections from full-texts and evaluate numerical data extraction for meta-analyses across a diverse range of LLMs in zero-shot settings. In addition to massive, closed LLMs we evaluate smaller, accessible models to provide a more complete view of the capabilities of modern LLMs for numerical data extraction of RCTs.

3. Extracting Numerical Results from Clinical Trial Reports

We are interested in the task of inferring structured findings reported in clinical trial reports that correspond to a specific intervention, comparator, and outcome (ICO). Past work (Lehman et al., 2019; Wadhwa et al., 2023b) has treated this as a *classification* task, categorizing articles as reporting that the given intervention (e.g., aspirin) induced a *significant increase*, *significant decrease*, or *no significant difference* with respect to the outcome of interest (e.g., duration of headache), relative to the specified comparator (e.g., placebo).

In this work, we consider a more challenging variant of this task in which the aim is to infer *numerical* data associated with each ICO. Specifically, we investigate the ability of LLMs to extract the numbers required to derive point estimates (which capture the sign and magnitude of relative treatment effects) and corresponding variances. If we could reliably infer such results from individual articles describing trials, this would permit *automatic statistical meta-analysis on demand*. Briefly, a standard fixed-effects meta-analysis—which assumes studies measure a common underlying effect (Hedges and Vevea, 1998)—aggregates point estimates of interest (e.g., odds ratios) $\hat{\theta}_i$ across k studies (indexed by i) by taking a weighted sum, where weights w_i are inverse to the variance associated with $\hat{\theta}_i$:

$$\hat{\theta}_{\text{FE}} = \frac{\sum_{i=1}^k w_i \hat{\theta}_i}{\sum_{i=1}^k w_i} \quad (1)$$

Meta-analysis of RCT evidence is critical because individual trials are inherently noisy and may reflect statistical biases. Aggregating evidence from independent trials in this way to estimate an overall treatment effect permits a robust effectiveness estimate (Egger et al., 1997). For this reason, meta-analytic results provide reliable clinical evidence, and meta-analyses often inform healthcare policy and patient care guidelines.

A downside of meta-analyses is that they are laborious to produce and keep up to date with new findings, especially given the rate at which new evidence accumulates (Bastian et al., 2010). This has motivated work on automating or semi-automating aspects of evidence synthesis, but reliably extracting the numerical values associated with ICO triplets has been a task too difficult for prior language technologies to reliably execute. Given the rapid progress in NLP—and specifically the capabilities of LLMs—we investigate, in this work, the degree to which modern LLMs can accurately extract numerical results from clinical trials sufficient for meta-analysis, fully automatically.

4. Annotation

To evaluate the feasibility of using LLMs to extract numerical data from RCTs for meta-analysis, we annotated a dataset of 699 abstracts and result sections from randomized controlled trials indexed in PubMed Central. Our dataset contains extensive annotations for the task of extracting numerical clinical findings for a given intervention, comparator, and outcome (ICO triplets) necessary for conducting meta-analyses (Equation 1).

Our dataset is derived from the *Evidence Inference* corpus (Lehman et al., 2019; DeYoung et al., 2020). This comprises articles describing RCTs.² For all of these RCT reports, medical experts have assessed the directionality of the evidence reported concerning given ICO triplets. Specifically, these annotations take the form of ternary classifications indicating whether the Intervention *significantly increased*, *significantly decreased*, or had *no significant effect* on the Outcome, as compared to the Comparator. We enrich this dataset by attaching *numerical* results to ICO triplets for a subset of the articles; these are at the level necessary to permit meta-analysis and so (much) more granular than the existing categorical annotations.

Annotation Details We randomly sampled 120 RCTs from the *Evidence Inference* dataset (DeYoung et al., 2020). Full-text papers are quite lengthy. Therefore, as a practical matter, we focus on only abstracts and the results sections of the full-text papers, where numerical results are most often reported. We compress the XML-formatted sections of these papers to reduce input length.³ We further pre-process the XML by removing all the attributes mostly about style rather than content. Finally, we convert each processed XML to mark-down, further compressing the content while maintaining a vital organizational structure. We release the post-processed version of the data used in this work to ensure reproducibility.

During the annotation process, we made some changes to the Intervention, Comparator, and Outcomes fields for certain cases. For Outcome fields that included multiple parts, we divided the parts into separate individual Outcomes and associated them with the same Intervention and Comparator values. For example, if the original Outcome label was “weight

2. Since it is constructed from articles in the open access subset, this dataset includes full-texts of papers in XML format.

3. We use <https://onlinexmltools.com/minify-xml>.

reduction at 12 months, BMI reduction at 12 months,” then we created two separate Outcome values—one as “weight reduction at 12 months” and the other as “BMI reduction at 12 months”—and associated these two new Outcomes with the same Intervention and Comparator. This results in having more ICO instances in the dataset than the original. In addition, we found cases when one Intervention field would have multiple interventions and the RCT report did not report numerical data for the combined, multiple interventions. In this case, we took a similar approach above and separated the parts into individual Interventions and associated them with the same Outcome and Comparator values.

Finally for cases when the Outcome is associated with multiple numerical measurements in the RCT report, we transformed these Outcomes into multiple Outcomes with the specific measurement included in the value. For example, the Outcome of “motion range” had several measurements including shoulder internal rotation, shoulder external rotation, and shoulder extension. In this particular example, we created three separate records in the annotated dataset each with different Outcomes: “motion range - shoulder internal rotation,” “motion range - shoulder external rotation,” and “motion range - shoulder extension.” These derived Outcomes had the same Intervention and Comparator as the original Outcome.

We annotated each ICO from the dataset by following the annotation schema. The annotation schema included the following:

1. *Type of outcome*: binary or continuous
2. *Data extraction for binary outcomes*: intervention event number, intervention group size, comparator event number, and comparator group size (2x2 contingency table)
3. *Data extraction for continuous outcomes*: intervention mean, intervention standard deviation, intervention group size, comparator mean, comparator standard deviation, and comparator group size

If a specific part of the numerical data is not found in the abstract or results section, the annotators left the field blank. We recorded metadata concerning whether the numerical information could be located within a figure or table and documented whether all of the desired numerical data was successfully retrieved.

To keep the task fairly straightforward, we decided to leave instances of RCTs that reported medians instead of means blank. If there are multiple time points for a given Outcome and the time point is not explicitly stated, we use the primary time point based on the RCT. If this is not provided, we use the last time point. Additionally, we only include the standard deviations if they were explicitly reported, or if the 95% confidence intervals were provided, by back-calculating the standard deviations using the formula outlined in Higgins et al. (2018). We also report the token counts for each RCT abstract and results text (under OpenAI’s tokenizer) to provide a sense of the input document length.

We set aside 10 random RCTs as a development set for prompt engineering and used the remaining 120 RCTs as a test or evaluation set. We release this dataset via a GitHub repository.⁴

Annotator backgrounds Two co-authors of the paper with backgrounds in computer science but familiarity with standard meta-analysis metrics annotated each record independently. The annotators met regularly to agree on the final annotations. We also solicited

4. <https://github.com/hyesunyun/llm-meta-analysis>

Table 1: Statistics for the annotated dataset.

Metric	Dev	Test	Total
# PMC Articles	10	110	120
# Prompts (ICOs)	43	656	699
# Binary Outcomes	11	172	183
# Continuous Outcomes	32	484	516
% With Enough Data for Point Estimates	62.79	58.84	59.08
Mean Articles Tokens	3331	3603	3581

feedback from a clinical researcher with expertise in evidence synthesis to discuss edge cases and ensure annotation quality.

Dataset statistics Table 1 gives an overview of statistics for our final dataset. Our dataset includes 120 RCTs (abstracts and results sections) and 699 total records. The development set has 10 RCTs with 43 records and the test set has 110 RCTs with 656 records. A total of 183 outcomes are categorized as binary while 516 as continuous. We found that a total of 413 records (59.08%) have sufficient data to calculate the point estimates of the clinical trials. The numerical data extracted during annotations were mostly found in the tables of the RCTs as 471 instances fell under this case. The average token number for each RCT was 3,581.

5. Can LLMs Accurately Extract Numerical Results from Trial Reports?

We evaluate whether modern LLMs are capable of extracting the numerical data necessary for meta-analysis from RCT reports. We assume a *zero-shot* setting here, i.e., we instruct models to extract the data elements of interest, without additional supervision. Few-shot learning may realize better performance for this task. However, RCT reports are often quite lengthy, even when considering only abstracts and results sections. This makes it practically difficult (and expensive) to include full examples in context. Therefore, we evaluate only the zero-shot case in this work.

Ideally, one might ask directly for the *point estimate* of interest, for example, the odds ratio with respect to a (dichotomous) outcome of interest between the treatment and control groups. However, deriving this often requires intermediate steps to transform raw data (e.g., a two-by-two table tallying the number of participants in each group who did and did not experience a specific outcome) into point estimates (e.g., odds ratios) and standard errors.

Given that LLMs are middling at math (Hong et al., 2024; Satpute et al., 2024; Urrutia and Araya, 2024)—and the fact that statistical meta-analysis packages like *metafor* (Viechtbauer, 2010) will readily compute such quantities from raw data—we instead adopt a stepwise approach intended to extract raw data. Specifically, we first prompt the model to infer whether a given outcome is binary or continuous based on a (natural language) description of the outcome, e.g. *elevation of glucose after 1 hour*; these descriptions are part of the annotations provided in the evidence inference dataset (DeYoung et al., 2020). Note that this assessment is made independently of the input article, and in some cases, the model may respond that it is unable to make this inference. The output from this step informs whether we prompt the model to extract raw data as would be reported in

Table 2: Full list of instruction-tuned LLMs used for the evaluation experiments.

Model	Type	Parameters	Sequence Length	Training Data
GPT-4 0125	General	1.7T	128000	Unknown
GPT-3.5 Turbo 0125	General	175B	16385	Unknown
Alpaca	General	13B	4096	instruction-following demonstrations generated from OpenAI’s <code>text-davinci-003</code>
Mistral Instruct v2	General	7B	32768	Unknown
Gemma Instruct	General	7B	8192	Web Documents, Code, Mathematical text, Instructions
OLMo Instruct	Science	7B	2048	Tulu 2 SFT Mix, Ultrafeedback Cleaned
PMC LLaMA	Biomedicine	13B	2048	Semantic Scholar Open Research Corpus, PMC LLaMA instructions
BioMistral	Biomedicine	7B	2048	PubMed Central Open Access Subset

the dichotomous case (i.e., a two-by-two table tallying numbers of events in the treatment and control groups, respectively) and continuous instance (measurements and variances associated with the outcome of interest in both groups).

Due to the limited sequence lengths of many smaller, open-source models, it was necessary to divide texts into segments sufficiently small to conform to a maximum length. This *chunking* process involves breaking down texts into smaller sections. Our approach includes pre-processing the text, converting number words to digits, and removing sentences without numerical values. We then iteratively concatenate segments until we have a set of chunks that are together at the token limit.

We pass the raw data extracted from this step through specialized statistical software (`statsmodels`; Seabold and Perktold 2010 or `metafor`; Viechtbauer 2010) to derive point estimates and standard errors, i.e., the inputs necessary for meta-analysis. This sequential approach provides a form of transparency: Users can inspect the “raw” outputs from LLMs to verify their correctness. We release the code and prompts used for the evaluation: <https://github.com/hyesunyun/llm-meta-analysis>.

5.1. Large Language Models (LLMs)

Recent work has demonstrated that LLMs are strong entity and relation extractors. Wadhwa et al. (2023a) showed that LLMs can achieve comparable results to fully supervised models for relation extraction, zero-shot. Related efforts have demonstrated the capabilities of zero-shot LLMs in healthcare specifically (Wei et al., 2021; Agrawal et al., 2022; Singhal et al., 2023; Sivarajkumar et al., 2024). In this work, we evaluate the following LLMs with respect to their ability to extract the data from RCT reports necessary for meta-analysis:

- **GPT-4** (Achiam et al., 2023): The latest version of the GPT-family of models from OpenAI. This is a large multimodal (text and image) model. We use `gpt-4-0125` which has a maximum context length of 128k. Context length is the amount of text that an LLM can process and retain in memory at any given time.
- **GPT-3.5⁵**: Also from OpenAI, the model behind ChatGPT. We use `gpt-3.5-turbo-0125` which can handle contexts of up to 16,385 tokens.

5. <https://platform.openai.com/docs/models/gpt-3-5-turbo>

- **Alpaca 13B** (Taori et al., 2023): An instruction fine-tuned version of the LLaMA-13B (Touvron et al., 2023) model with a max context length of 4,096 tokens.
- **Mistral 7B Instruct v2**⁶: An instruction fine-tuned version of the Mistral-7B-v0.1 (Jiang et al., 2023) model. Allows for a context length of 32,768 tokens.
- **Gemma 7B Instruct** (Team et al., 2024): An open-weights LLM from Google DeepMind. Max context length: 8,192 tokens.
- **OLMo 7B Instruct** (Groeneveld et al., 2024): Open Language Model (OLMo) from AI2. Features an input sequence length upper bound of 2,048 tokens.
- **PMC LLaMa** (Wu et al., 2023): A 13B parameter model which was initialized to LLaMA-13B (Touvron et al., 2023) and then further pretrained over medical corpora. The model was then instruction-tuned. It offers a max input and output sequence length of 2,048.
- **BioMistral** (Labrak et al., 2024): A Mistral-based model which was further pre-trained over data from PubMed Central Open Access. It has an input upper bound of 2,048 tokens (even though the base model has a significantly greater sequence length; presumably this was done due to resource constraints).

5.2. Metrics

To measure the performance of LLMs in terms of their ability to extract numerical data from trial reports, we first consider their accuracy with respect to categorizing outcome types (as binary or continuous), and we then evaluate data extraction accuracy both in terms of exact and partial matches with reference (manually extracted) data.

We consider data extracted from a trial report partially correct where a subset of the extracted numerical data matches the reference. For example, for a binary outcome with 4 numerical values (two-by-two table entries), correctly extracting 1, 2, or 3 of these may be viewed as a partial match. We calculate and report partial accuracy for all possible partial matches.

Articles sometimes report results ambiguously, or not at all. In these cases, we wish the model to abstain from providing data. We also report the number of times the model mistakenly outputs “unknown” for a specific label. This measure is useful for understanding how conservative the model is, as it captures instances where humans do not find the results ambiguous but the models do.

In addition to considering whether extracted data matches reference data (one binary indicator per ICO), we consider the magnitude of difference between extracted values and references (in standardized units). This is informative as it indicates whether errors are likely to have an important impact on downstream meta-analysis (large magnitude differences would affect the overall treatment efficacy estimate $\hat{\theta}_{FE}$ more). We calculate mean standardized errors (average of all the differences between the efficacy estimates from numerical data of LLMs and reference) and their variance (standard error and 95% confidence interval).

6. <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

Table 3: LLM performances for inferring outcome type (binary or continuous). **# Unknowns** refers to instances when the model outputs the unknown token “x”; these are effectively incorrect (here the type should be inferable).

	GPT-4	GPT-3.5	Alpaca	Mistral	Gemma	OLMo	PMC LLaMA	BioMistral
Accuracy	0.713	0.607	0.739	0.201	0.665	0.290	0.732	0.133
F1 - Binary	0.735	0.680	0.000	0.576	0.590	0.424	0.124	0.275
F1 - Continuous	0.836	0.690	0.851	0.183	0.716	0.079	0.848	0.135
# Unknowns	155	152	1	489	0	5	15	409

6. Results

We evaluated the performance of LLMs on three different tasks that are aimed to assist in automating meta-analyses in zero-shot. The three tasks are inferring outcome type, extracting numerical results for binary outcomes, and extracting numerical results for continuous outcomes. These tasks were performed independently of each other. LLMs were tasked to output the answers in a structured format (categorical answer for inferring outcome type and YAML format for data extraction). Lastly, we show the promise of using LLMs end-to-end for automating meta-analyses with a case study where we reproduce a meta-analysis with fully automated data.

6.1. Extracting Data from Trial Reports

Inferring outcome types Results for inferring outcome types are reported in [Table 3](#). Alpaca-13B realizes the highest exact match accuracy of 0.739, followed by PMC LLaMA with 0.732. Although Alpaca-13B and PMC LLaMA have high accuracy, their F1 scores for the binary label are poor because the model mostly predicts the outcome type to be “continuous”, which is the majority class (74.78% of the test set). GPT-4 and GPT-3.5 achieve high accuracy and F1 scores, besting the open-source models. The open, smaller models we evaluated appear less stable, i.e., the F1 score varies considerably. Among these models, Gemma offers the best performance for exact match accuracy, F1 scores, and the number of instances where the model designates the outcome type as “unknown”.

Binary outcomes Performance metrics for binary outcome data extraction are available in [Table 4](#). Based on the accuracy metrics and number of complete data, GPT-4 significantly outperforms all the other models. GPT-4 has the highest overall exact match accuracy of 0.655, followed by GPT-3.5 with 0.298. Similar to the task of inferring outcome types, the open, smaller models vary substantially in terms of their match accuracies. Especially, the models trained on biomedical text have one of the worst performances with both PMC LLaMA and BioMistral achieving a mere 0.035 in total exact accuracy and producing the most number of “unknowns” leading to the lowest percentage of complete data.

The mean standardized errors of the log odds ratio calculated from the extracted data are lowest for the best performing models and highest for the worst performing models. GPT-

Table 4: Performance in terms of extracting numerical values for binary outcomes. **IE**: intervention events; **IGS**: intervention group size; **CE**: comparator events, and; **CGS** comparator group size. Partial match numbers reflect leniency in the number of matches required for an instance to count as (partially) “correct”. The MSE is the mean standardized error of the log odds ratios calculated with data extracted from the model. The number of unknowns refers to the number of times the model produced the unknown token “x” when the reference was not unknown; in this case, the reference data contained unknowns 72 times. The percentage of complete data shows how many model outputs had enough data to calculate the point estimate (and variance), compared to the reference data.

		GPT-4	GPT-3.5	Alpaca	Mistral	Gemma	OLMo	PMC LLaMA	BioMistral
Exact Match	<i>Total</i>	0.655	0.298	0.035	0.164	0.135	0.012	0.035	0.035
	<i>IE</i>	0.749	0.462	0.129	0.345	0.275	0.076	0.146	0.158
	<i>IGS</i>	0.842	0.655	0.094	0.515	0.509	0.170	0.088	0.053
	<i>CE</i>	0.737	0.392	0.129	0.333	0.275	0.123	0.158	0.158
	<i>CGS</i>	0.830	0.649	0.094	0.567	0.556	0.140	0.058	0.053
Partial Match	<i>3</i>	0.690	0.415	0.035	0.251	0.216	0.035	0.035	0.035
	<i>2</i>	0.901	0.696	0.181	0.637	0.585	0.187	0.164	0.175
	<i>1</i>	0.912	0.749	0.193	0.708	0.678	0.275	0.216	0.175
MSE		0.101	0.441	0.485	0.657	0.913	1.253	1.523	-
# Unknowns		41	145	490	28	90	319	524	612
% Complete		87.94	61.70	9.22	87.23	58.87	24.11	7.09	0.00

4’s MSE is 0.101 with a standard error of 0.043 (95% CI 0.017 to 0.186) while BioMistral did not extract any computable data comparable with our reference data.

Continuous outcomes Table 5 presents the performance metrics for continuous outcome data extraction. Due to this task involving more values to extract, LLM performance was quite low with even GPT-4 achieving exact match accuracy of 0.487. The second best model is GPT-3.5 with a total exact match accuracy dropping to 0.280. Similarly to binary outcome data extraction, the smaller, open-source models performed poorly.

The mean standardized errors show the comparison between the effect estimates calculated from model outputs and the reference data. GPT-4 has the lowest value of 0.290 (SE: 0.112; 95% CI: 0.071 to 0.510), followed by GPT-3.5 with 0.951 (SE: 0.420, 95% CI: 0.127 to 1.775). LLMs further trained on biomedical texts consistently performed the worst with BioMistral producing no complete data to calculate the standardized mean difference.

Error Analysis To characterize the types of errors that LLMs make when extracting numerical values from trial reports, we qualitatively evaluated mistakes observed on validation data. For this analysis, we focused on outputs from GPT-4 and Mistral as representative models; these were the best-performing closed and open LLMs evaluated, respectively.

For outcome type inference, GPT-4 produced an “unknown” a total of 155 times, inferred an outcome to be continuous when it was in fact binary 25 times, and predicted a continuous

Table 5: Performance of extracting numerical values for continuous outcomes. **IM**: intervention mean, **ISD**: intervention standard deviation, **IGS**: intervention group size, **CM**: comparator mean, **CSD**: comparator standard deviation, and **CGS**: comparator group size. Partial match numbers refer to how many parts need to match the reference for the instance to count as correct. The MSE is the mean standardized error of the standardized mean differences calculated with data extracted from the model. The number of unknowns refers to the number of times the model produced the unknown token “x” when the reference was not unknown; in this case, the reference data contained unknowns 925 times. The percentage of complete data shows how many of the model outputs had enough data to calculate the point estimate (and variance), compared to the reference data.

		GPT-4	GPT-3.5	Alpaca	Mistral	Gemma	OLMo	PMC LLaMA	BioMistral
Exact Match	<i>Total</i>	0.487	0.280	0.039	0.095	0.087	0.035	0.039	0.041
	<i>IM</i>	0.720	0.538	0.309	0.348	0.328	0.221	0.369	0.390
	<i>ISD</i>	0.751	0.606	0.334	0.375	0.412	0.311	0.447	0.470
	<i>IGS</i>	0.734	0.641	0.216	0.507	0.534	0.190	0.107	0.087
	<i>CM</i>	0.720	0.526	0.330	0.361	0.324	0.227	0.390	0.402
	<i>CSD</i>	0.738	0.584	0.338	0.390	0.404	0.282	0.456	0.472
	<i>CGS</i>	0.691	0.608	0.181	0.427	0.447	0.184	0.109	0.087
Partial Match	<i>5</i>	0.542	0.336	0.045	0.115	0.103	0.060	0.058	0.054
	<i>4</i>	0.724	0.555	0.293	0.293	0.342	0.173	0.375	0.402
	<i>3</i>	0.765	0.645	0.311	0.421	0.408	0.231	0.392	0.408
	<i>2</i>	0.913	0.814	0.470	0.691	0.699	0.408	0.497	0.501
	<i>1</i>	0.922	0.872	0.551	0.794	0.810	0.507	0.518	0.501
MSE		0.290	0.951	6.257	1.138	3.466	1.738	-	-
# Unknowns		422	437	1169	483	775	1213	1778	1985
% Complete		63.64	62.40	31.82	62.81	40.08	11.98	4.96	0.00

outcome as binary 8 times. Mistral 7B Instruct had 480, 34, and 1 errors, respectively. In addition, Mistral produced 9 badly formatted YAML outputs.

For binary outcome extraction, GPT-4 errors broke down as follows: Producing incorrect numbers (34); Producing “unknown” for inferable values (20); and Outputting actual numbers when the reference is “unknown” (12). Mistral 7B Instruct yielded invalid output formats 3 times. In addition, Mistral produced incorrect numerical values (123); “unknown” values when they should not have (7); and confabulated numerical values for “unknown” reference values (25).

For continuous outcome extraction, the majority of errors involved producing the wrong numerical value (GPT-4: 113; Mistral: 226); producing “unknown” values inappropriately (GPT-4: 142; Mistral: 171); and generating numerical values for “unknown” reference values (GPT-4: 38; Mistral: 171). Mistral also produced 5 poorly formatted outputs in this case. [Appendix A](#) visualizes the types and number of errors from GPT-4 and Mistral.

We sampled 136 errors from GPT-4 and 148 errors from Mistral. We sampled up to 20 errors from each general error category mentioned above. Then, we conducted a qualitative error analysis to better characterize the mistakes made by these LLMs. We inductively

annotated possible sources and reasons for each sampled error and aggregated commonly found reasons. Selected samples of errors are available in [Appendix B](#).

In addition to LLMs producing the wrong labels for clearly defined outcomes, our analysis revealed that errors in inferring outcome type are often attributable to ambiguous phrasing of the outcome values. Since the additional context of the trial reports is not provided as input, some outcome values can be either reported as a binary outcome or continuous. Examples of this include words such as “rate” and “level” even in cases where they are treated as dichotomous, and “failure” and “discontinuation” even when they are continuous. This suggests that providing additional context from the article when making this classification may improve performance.

For outcome extraction tasks, we found that smaller models such as Mistral sometimes produced formats that do not conform to the expected outputs whereas GPT-4 did not have this problem. These bad formats were often the contributing factor for having more “unknowns” since badly formatted outputs were treated as “unknown” numerical data.

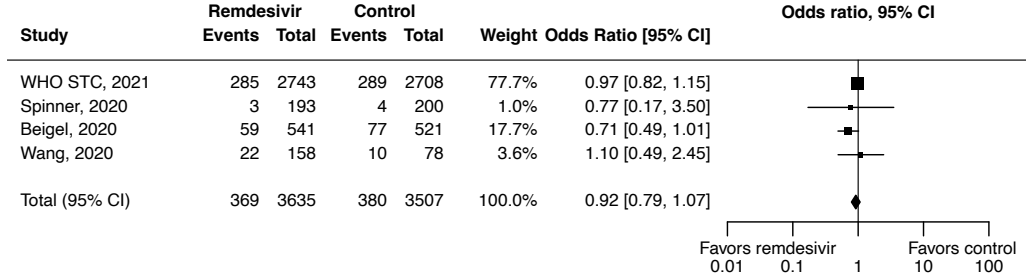
We observed that incorrect outputs were often due to “hallucinations” (outputting numerical data not found anywhere in the trial report) or extracting values from the wrong intervention, comparator, or outcome measures from the wrong timepoint. Also, models often produced “unknown” as an answer if the target numerical value was a negative value. One possible explanation can be that these negative numbers are often reported in RCTs with two hyphens (“--”) rather than a single one preceding a number which can lead to the model ignoring them as proper numerical values. When RCTs report pre- and post-treatment measures, LLMs sometimes extract the values only from the pre-treatment and ignore the post-treatment numbers (whereas we need both for meta-analysis; typically one is interested in the comparative differences).

One common apparent source of error occurs when the LLMs need to perform simple math such as division or subtraction to get the correct total group sizes. For example, an abstract might report that 40 participants were randomly assigned to two study groups. In this case, most humans would assume (absent additional information) that each group was assigned 20 participants. However, LLMs sometimes fail to make such inferences and output totally irrelevant numbers or “unknown.” Another common group size mistake appears to be due to the trials containing several different group sizes which correspond to different ICO triplets. Some outcome measures are reported based on the sample that successfully finished the study, for example. Finally, while apparently rare, LLMs sometimes confuse medians and means, or standard errors and standard deviations.

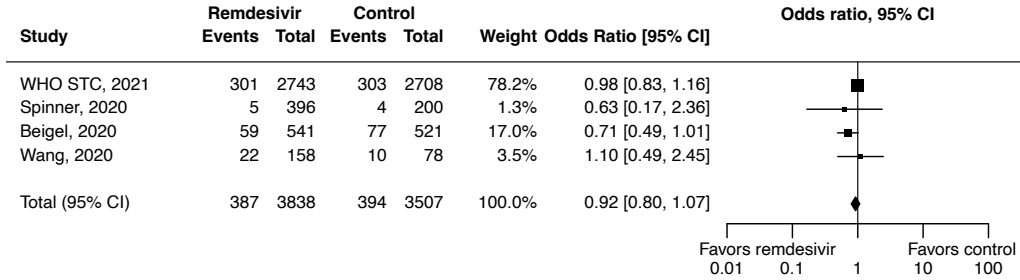
6.2. Towards Fully Automatic Meta-analysis: An Example

To demonstrate the feasibility of conducting a standard (fixed effects) meta-analysis with LLMs, we provide a case study based on an existing meta-analysis published in the Cochrane Library on remdesivir for treatment of COVID-19 ([Grundeis et al., 2023](#)). For this exercise, we focus on one outcome: “All-cause mortality at up to day 28”. This outcome is a dichotomous (binary) outcome, and the comparison is between remdesivir as the intervention and a comparator of standard care. The original meta-analysis includes four trial reports and reports all important numerical values extracted from each trial.

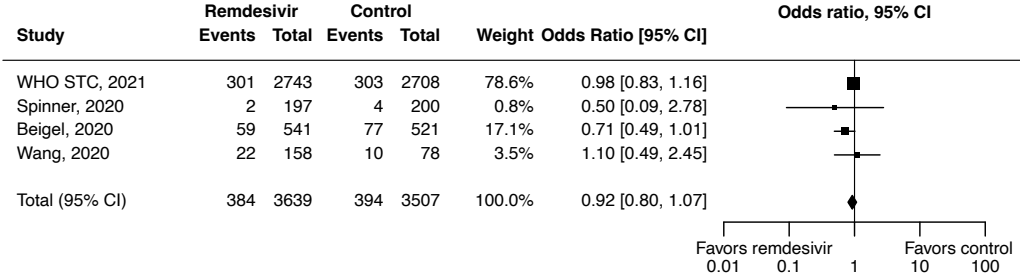
We followed the same pre-processing approach for preparing the trial reports after downloading their full text in XML via `Bio.Entrez`⁷ Python package (Cock et al., 2009). We prompted both GPT-4 and Mistral Instruct 7B to extract the numerical values. We start by inferring the outcome type, and then we use the outputs from this step to prompt the model to output the numerical values relevant to the given ICO (here: *remdesivir, standard care, all-cause mortality at up to day 28*).



(A) Cochrane meta-analysis (reference)



(B) meta-analysis from GPT-4 outputs



(C) meta-analysis from Mistral Instruct 7B outputs

Figure 2: Forest plots showing fixed-effect meta-analyses from (A) Cochrane review (reference; performed with manually extracted data), (B) GPT-4 extracted data, (C) Mistral Instruct 7B extracted data. For this particular case, both LLMs permit highly accurate meta-analytic estimates, fully automatically.

7. <https://biopython.org/docs/1.75/api/Bio.Entrez.html>

Both GPT-4 and Mistral correctly categorize the outcome type as binary. GPT-4 had an exact match accuracy of 0.500 and was able to match at least 2 values of the 2x2 outcome table correctly for all trials. Mistral performed comparably to GPT-4 in this case. The exact extracted data from the Cochrane reference and the two models are shown in [Figure 2](#) as forest plots for comparison.⁸ Both models made mistakes in extracting data from two studies ([Consortium, 2021](#); [Spinner et al., 2020](#)). The errors for [Consortium \(2021\)](#) were related to the correct numbers not being readily available in the trial text but requiring extra calculations. For [Spinner et al. \(2020\)](#), GPT-4 hallucinated the events and the total numbers for the intervention, while Mistral extracted the wrong number due to choosing from the wrong timepoint for the outcome measure. These errors were (very) minor when put into a fixed-effect meta-analysis; using outputs from both models resulted in the correct total log odds ratio and the 95% CI was off by only 0.1 ([Figure 2](#)).

7. Discussion

In this work, we aim to assess whether modern LLMs are sufficiently capable of extracting data to produce accurate, fully automated meta-analyses. To this end, we comprehensively annotated a modest evaluation set of clinical trial reports with numerical findings attached to interventions, comparators, and outcomes. We then used this dataset to evaluate a wide range of LLMs in terms of their ability to infer structured numerical findings.

Our results indicate that modern LLMs can perform this task with some accuracy, but difficulties remain. Massive general models like GPT-4 and GPT-3.5 perform fairly well, especially for tasks inferring outcome types and extracting binary outcomes. Smaller, open-source models also showed some capability but fared comparatively worse. For some of the smaller models with minimal context windows, this may be because we had to pass chunks of the input trial reports through the network, increasing noise. Mistral 7B Instruct performed the best of the open-source models considered. Interestingly, despite being further trained on domain-specific text, PMC LLaMA and BioMistral performed considerably worse than other similar-sized general-purpose LLMs by producing the most number of “unknowns” and the lowest percentage of complete data. For extracting continuous outcomes, all the LLMs we evaluated performed below 50% exact match accuracy.

We further demonstrated the use of LLMs in a case study in which we conducted a meta-analysis with automatically extracted results. This showed good results in almost replicating the exact total log odds ratio and confidence interval, suggesting a future in which meta-analysis is fully automatable. However, LLMs performed fairly well in this example probably because the constituent RCT reports have a clear structure in the outcome measure reports with limited ambiguity (such as different time points). Also, the example only contains 4 RCTs, which helps in reducing the accumulation of errors that can often happen with the inclusion of more data.

Our work shows that modern LLMs such as GPT-4 and Mistral are promising for extracting numerical findings from trial reports to produce automated meta-analyses. However, they fail to extract findings that require more specialized inference when multiple similar outcome measures are reported or the outcome measures are ambiguous. For now,

8. Forest plots show point estimates and associated confidence intervals corresponding to individual studies, as well as the aggregated (overall) estimate and interval.

these should remain an assistive technology, although it seems plausible that near-term improvements in LLMs may permit robust automatic meta-analysis.

Limitations There are several limitations to this work. Due to the time required to perform detailed annotations, the evaluation dataset we have introduced is a small sample. In addition, while the annotations were performed carefully by individuals familiar with meta-analysis, they did not have clinical background (that said, this work was done in consultation with a clinician to ensure accuracy).

We only investigated zero-shot applications of LLMs, mostly because of the context window that would be required to include examples of data extraction from full-texts. We did not attempt few-shot or fine-tuning. Furthermore, we did only minimal and informal “prompt engineering”; it is certainly possible that alternative instructions may yield better (or worse) performance. A related limitation is that we tasked LLMs with inferring outcome types from only descriptions; providing additional context from articles may improve the performance on this first subtask.

Acknowledgments

This research was partially supported by National Science Foundation (NSF) grants RI-2211954 and IIS-1750978, and by the National Institutes of Health (NIH) under the National Library of Medicine (NLM) grant 2R01LM012086.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.130. URL <https://aclanthology.org/2022.emnlp-main.130>.
- Hilda Bastian, Paul Glasziou, and Iain Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9):e1000326, 2010.
- Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John Wiley & Sons, 2021.
- Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.
- WHO Solidarity Trial Consortium. Repurposed antiviral drugs for covid-19—interim who solidarity trial results. *New England journal of medicine*, 384(6):497–511, 2021.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Jay DeYoung, Eric Lehman, Ben Nye, Iain J. Marshall, and Byron C. Wallace. Evidence inference 2.0: More data, better models, 2020.
- Matthias Egger, George Davey Smith, and Andrew N Phillips. Meta-analysis: principles and procedures. *Bmj*, 315(7121):1533–1537, 1997.
- Gerald Gartlehner, Leila Kahwati, Rainer Hilscher, Ian Thomas, Shannon Kugley, Karen Crotty, Meera Viswanathan, Barbara Nussbaumer-Streit, Graham Booth, Nathaniel Erskine, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *medRxiv*, pages 2023–10, 2023.
- Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.
- Felicitas Grundeis, Kelly Ansems, Karolina Dahms, Volker Thieme, Maria-Inti Metzendorf, Nicole Skoetz, Carina Benstoem, Agata Mikolajewska, Mirko Griesel, Falk Fichtner, et al. Remdesivir for the treatment of covid-19. *Cochrane Database Syst Rev*, pages CD014962–CD014962, 2023.
- Larry V Hedges and Jack L Vevea. Fixed-and random-effects models in meta-analysis. *Psychological methods*, 3(4):486, 1998.
- J Higgins, Toby Lasserson, Jackie Chandler, David Tovey, and Rachel Churchill. Standards for the conduct and reporting of new cochrane intervention reviews, reporting of protocols and the planning, conduct and reporting of updates. *Methodological Expectations of Cochrane Intervention Reviews (MECIR)*, 2018.
- Pengfei Hong, Deepanway Ghosal, Navonil Majumder, Somak Aditya, Rada Mihalcea, and Soujanya Poria. Caught in the quicksand of reasoning, far from agi summit: Evaluating llms’ mathematical and coding competency through ontology-guided interventions, 2024.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- David Kartchner, Selvi Ramalingam, Irfan Al-Hussaini, Olivia Kronick, and Cassie Mitchell. Zero-shot information extraction for clinical meta-analysis using large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 396–405, 2023.
- Qusai Khraisha, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. Can large language models replace humans in the systematic review process? evaluating gpt-4’s efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *arXiv preprint arXiv:2310.17526*, 2023.

- Svetlana Kiritchenko, Berry De Bruijn, Simona Carini, Joel Martin, and Ida Sim. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10:1–17, 2010.
- Wojciech Kusa, Oscar E Mendoza, Matthias Samwald, Petr Knuth, and Allan Hanbury. Csmed: Bridging the dataset gap in automated citation screening for systematic literature reviews. *Advances in Neural Information Processing Systems*, 36, 2024.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.
- Joseph Lau, Christopher H Schmid, and Thomas C Chalmers. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *Journal of clinical epidemiology*, 48(1):45–57, 1995.
- Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3705–3717, 2019.
- Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models? In Bobak J. Mortazavi, Tasmie Sarker, Andrew Beam, and Joyce C. Ho, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 578–597. PMLR, 22 Jun–24 Jun 2023. URL <https://proceedings.mlr.press/v209/eric23a.html>.
- M Hassan Murad, Noor Asi, Mouaz Alsawas, and Fares Alahdab. New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127, 2016.
- Faith Wavinya Mutinda, Kongmeng Liew, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. Automatic data extraction to support meta-analysis statistical analysis: a case study on breast cancer. *BMC Medical Informatics and Decision Making*, 22(1):158, 2022.
- Aakanksha Naik, Bailey Kuehl, Erin Bransom, Doug Downey, and Tom Hope. Care: Extracting experimental findings from clinical literature, 2023.
- Tim Reason, Emma Benbow, Julia Langham, Andy Gimblett, Sven L Klijn, and Bill Malcolm. Artificial intelligence to automate network meta-analyses: Four case studies to evaluate the potential application of large language models. *PharmacoEconomics-Open*, pages 1–16, 2024.
- Ankit Satpute, Noah Giessing, Andre Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. Can llms master math? investigating large language models on math stack exchange, 2024.
- Skipper Seabold and Josef Perktold. Statsmodels: econometric and statistical modeling with python. *SciPy*, 7:1, 2010.

- Mahsa Shamsabadi, Jennifer D’Souza, and Sören Auer. Large language models for scientific information extraction: An empirical study for virology. *arXiv preprint arXiv:2401.10040*, 2024.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318, 2024.
- Christoph D Spinner, Robert L Gottlieb, Gerard J Criner, José Ramón Arribas López, Anna Maria Cattelan, Alex Soriano Viladomiu, Onyema Ogbuagu, Prashant Malhotra, Kathleen M Mullane, Antonella Castagna, et al. Effect of remdesivir vs standard care on clinical status at 11 days in patients with moderate covid-19: a randomized clinical trial. *Jama*, 324(11):1048–1057, 2020.
- Rodney L Summerscales, Shlomo Argamon, Shangda Bai, Jordan Hupert, and Alan Schwartz. Automatic summarization of results from clinical trials. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 372–377. IEEE, 2011.
- Zhuanlan Sun, Ruilin Zhang, Suhail A Doi, Luis Furuya-Kanamori, Tianqi Yu, Lifeng Lin, and Chang Xu. How good are large language models for automated data extraction from randomized trials? *medRxiv*, pages 2024–02, 2024.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivi re, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth e Lacroix, Baptiste Rozi re, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Felipe Urrutia and Roberto Araya. Who’s the best detective? large language models vs. traditional machine learning in detecting incoherent fourth grade math answers. *Journal of Educational Computing Research*, 61(8):187–218, 2024.
- Wolfgang Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010. doi: 10.18637/jss.v036.i03.

- Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation extraction in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.868. URL <https://aclanthology.org/2023.acl-long.868>.
- Somin Wadhwa, Jay DeYoung, Benjamin Nye, Silvio Amir, and Byron C Wallace. Jointly extracting interventions, outcomes, and findings from rct reports with llms. In *Machine Learning for Healthcare Conference*, pages 754–771. PMLR, 2023b.
- Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 173–182, 2010.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.
- Hye Yun, Iain Marshall, Thomas Trikalinos, and Byron Wallace. Appraising the potential uses and harms of LLMs for medical systematic reviews. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10122–10139, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.626. URL <https://aclanthology.org/2023.emnlp-main.626>.

Appendix A.

We present bar plots of the breakdown of the general error types we have identified from GPT-4 and Mistral Instruct 7B outputs. Our error analysis reveals that Mistral produces significantly more errors across all three tasks than GPT-4.

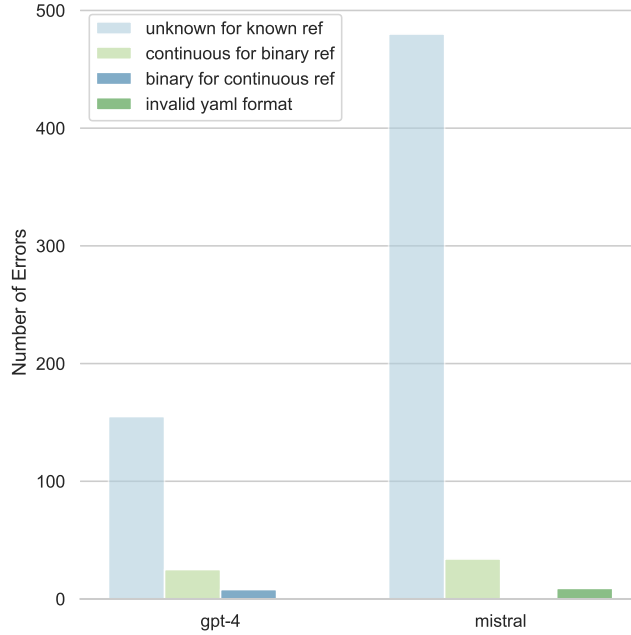


Figure 3: Bar plot showing the types of errors from GPT-4 and Mistral Instruct 7B for *outcome type inference* task.

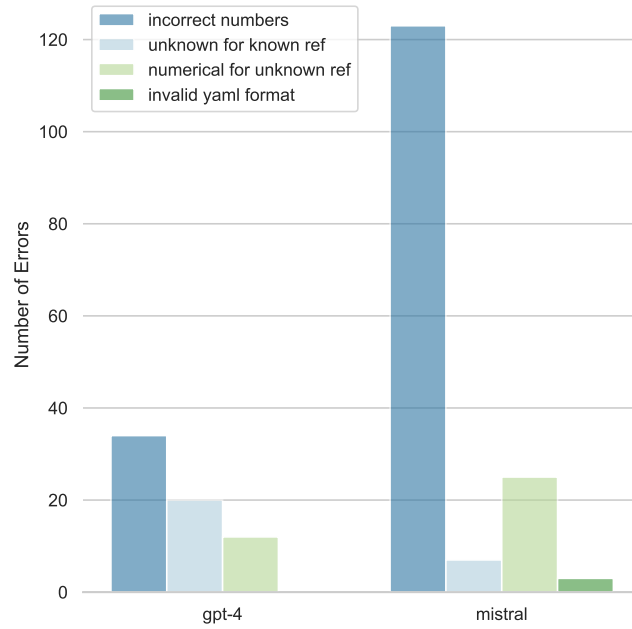


Figure 4: Bar plot showing the types of errors from GPT-4 and Mistral Instruct 7B for *binary outcomes extraction* task.

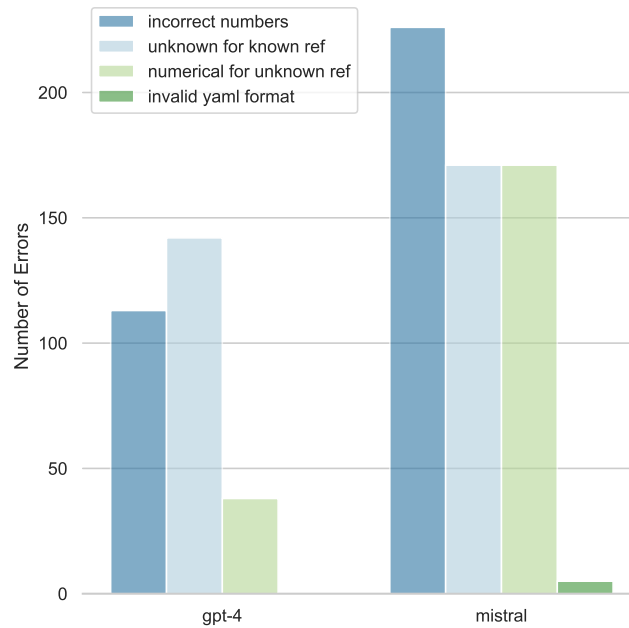


Figure 5: Bar plot showing the types of errors from GPT-4 and Mistral Instruct 7B for *continuous outcomes extraction* task.

Appendix B.

Some samples of errors from GPT-4 and Mistral Instruct 7B are presented below. We sampled these as part of the qualitative error analysis from [Section 6](#).

Table 6: Samples of errors from inferring outcome type with GPT-4 and Mistral Instruct 7B.

Model	Outcome (Input)	Output	Reference
GPT-4	communication ability	unknown	continuous
GPT-4	survival	binary	continuous
GPT-4	differences in disc herniation	continuous	binary
Mistral	glycemic level	unknown	continuous
Mistral	completed treatment	binary	continuous
Mistral	level of postoperative pain	continuous	binary

GPT-4		Mistral	
I: Amifostine C: Control O: Grade III/IV stomatitis		I: 0.2 mg oral clonidine C: 2 mg oral midazolam O: the percentages of patients in the agitated sedation scale	
Trial Report: ... 40 patients with solid tumours were randomized to receive HD-VIC chemotherapy with or without amifostine (910 mg/m ² at day 1--3)... Stomatitis grade III/IV occurred in 25% without versus 0% of patients with amifostine (*P* = 0.01)		Trial Report: ... In a randomized clinical, trial 80 patients enrolled in the study, 40 in the clonidine group and 40 in the midazolam group... Further, the percentages of patients in the agitated and very agitated group were significantly higher in the midazolam group (20 and 25%, respectively), compared to the clonidine group (2.5 and 7.5%, respectively)...	
OUTPUT: intervention: events: 0 group_size: x comparator: events: 25 group_size: x	REFERENCE: intervention: events: 0 group_size: 20 comparator: events: 5 group_size: 20	OUTPUT: intervention: events: 7.5 group_size: 40 comparator: events: 25 group_size: 40	REFERENCE: intervention: events: 14 group_size: 40 comparator: events: 7 group_size: 40

Figure 6: Samples of errors from extracting binary outcomes with GPT-4 and Mistral Instruct 7B. The GPT-4 error is from failing to infer that the group sizes should be half the total number of participants. For Mistral, the model received a binary outcome with the word “percentages.” Mistral was faithful to the given outcome by extracting the percent numbers found in the text but failed to output participant numbers which are more often used for binary outcomes.

GPT-4		Mistral	
I: Allogeneic mesenchymal precursor cells C: Control O: Activities of daily living		I: Multi-component physical activity C: Usual PE and sport programmes O: BMI reduction at 12 months	
Trial Report: ... Baseline characteristics of study participants ----- MPC + HA (n = 11) HA alone (n = 6) ----- Activities of daily living 82.1 (11.0) 92.8 (6.4) ...		Trial Report: ... Changes in adiposity outcomes from baseline to 12-month follow-up ----- Intervention (n=592) mean (95% CI) Control (n=459) mean (95% CI) ----- BMI 21.07 (20.56, 21.59) 21.27 (20.73, 21.81) ...	
OUTPUT: intervention: mean: 82.1 sd: 11.0 group_size: 11 comparator: mean: 92.8 sd: 6.4 group_size: 6	REFERENCE: intervention: mean: x sd: x group_size: 11 comparator: mean: x sd: x group_size: 6	OUTPUT: intervention: mean: 21.86 sd: 0.84 group_size: 560 comparator: mean: 21.27 sd: 0.81 group_size: 425	REFERENCE: intervention: mean: 21.07 sd: 6.39 group_size: 592 comparator: mean: 21.27 sd: 5.9 group_size: 459

Figure 7: Samples of errors from extracting continuous outcomes with GPT-4 and Mistral Instruct 7B. The error from GPT-4 is due to the model extracting values from the baseline and not the post-intervention which is the outcome measure of interest. The error from Mistral is related to the model extracting the means and standard deviations from the wrong locations of the text.