Beyond Abstracts: A New Dataset, Prompt Design Strategy and Method for Biomedical Synthesis Generation

James O'Doherty[1]*, **Cian Nolan**[1]*, **Yufang Hou**[2], **Anya Belz**[1]
[1]Dublin City University, [2] IBM Research Europe - Ireland
james.odoherty3@mail.dcu.ie, cian.nolan95@mail.dcu.ie,
yhou@ie.ibm.com, anya.belz@dcu.ie

Abstract

The biomedical field relies on cost and time intensive systematic reviews of papers to enable practitioners to keep up to date with research. Impressive recent advances in large language models (LLMs) have made the task of automating at least part of the systematic review process feasible, but progress is slow. This paper identifies some factors that may have been holding research back, and proposes a new, enhanced dataset and prompting-based method for automatic synthesis generation, the most challenging step for automation. We test different models and types of information from and about biomedical studies for their usefulness in obtaining high-quality results. We find that, surprisingly, inclusion of paper abstracts can worsens results. Instead, study summary information, and system instructions informed by domain knowledge, are key to producing high-quality syntheses.

1 Introduction

Medical practitioners need to keep up to date with the latest medical research, but the ever increasing volume of studies makes it difficult to separate signal from noise. The goal of systematic reviews is to synthesise all relevant evidence for a clinical query (Higgins et al., 2023) and provide clear, up-to-date answers based on high-quality research. Systematic reviews are considered the most reliable form of evidence in the biomedical field. Consequently, they have a huge influence on the medical decisions made by doctors, health authorities and individuals.

Producing systematic reviews is a slow and costly process. A study in 2019 estimated that the average cost of producing a biomedical systematic review was $141,194.80 (Michelson and Reuter, 2019). The high cost is due to reviewers having to sift through hundreds or thousands of potentially relevant studies to find the high quality studies that are included in their final analysis which must then undergo rigorous statistical analysis before a final conclusion is reached. Unsurprisingly, there is a lot of interest in automating different steps in the process, and recent advancements in LLMs offer a promising avenue to do just this.

Prior work in this area has tended to take an end-to-end approach to the task and to use limited information about reviews and included studies in the input which does not reflect a deep enough understanding of the systematic review process. Below we start by setting out this process and the information collected in repositories like the Cochrane Library (Section 2), followed by an overview of previous work where we identify important details not included in prior work that may be useful in solving the task (Section 3). We use these insights to create a new, richer dataset for the biomedical synthesis task (Section 4), and a new prompting-based approach to generating biomedical scientific syntheses (Section 5). We show the promise of this new approach via evaluation with diverse metrics and discuss key observations (Section 8). We make our dataset and code available on GitHub.[1]

2 Background

The Cochrane Library is one of the most highly respected institutions for creating systematic medical reviews, which it collects as the Cochrane Reviews, a public repository of systematic reviews. The following are identified[2] as the key steps in creating a Cochrane Review: (1) identification of relevant studies; (2) selection of studies for inclusion / evaluation of their strengths and limitations; (3) systematic collection of data; and (4) appropriate synthesis of data.

Our focus in the work presented here is on the

*Equal contribution

[1] https://github.com/JOD-code/Beyond-Abstracts-Biomedical-Synthesis-Generation

[2] https://www.cochranelibrary.com/about/about-cochrane-reviews

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 358–377

fourth step where experts review the data to form the final conclusion of the systematic review. Conclusions are typically provided in both quantitative and qualitative forms. Our focus is on automating the qualitative analysis of systematic reviews. We leave the automation of PICO (Population, Intervention, Comparator and Outcome) extraction and quantitative analysis (see 4.6) to future work. Figure 1 provides an overview of the entire process, with the portion enclosed within the dashed box indicating the part that our approach focuses on automating.

2.1 Papers and studies

An important distinction in the context of systematic reviews is between *papers* and *studies*. Medical studies can produce a large amount of data with results often reported in multiple papers. A single *study* may itself have been reported in a single *paper,* or across several *papers*. When selecting relevant studies, many studies will be reviewed but ultimately excluded from the final synthesis of evidence for various reasons. The remaining studies that are included in the final synthesis are referred to as the *included studies*. Systematic review authors may include references to other papers that are not part of the basis of the final synthesis but are referred to in the analysis for other reasons (e.g. as background).

2.2 PICO information

The key elements of biomedical intervention studies are their Population, Interventions, Comparators and Outcomes, together known as PICO elements (Higgins et al., 2023). The *Population* element contains information about study participants, including their number, demographics and risk factors. *Interventions* describes the treatments under investigation. *Comparison* refers to the treatment alternative tested (e.g. placebo, other drugs). *Outcomes* summarises the impact of interventions on the population as compared to the comparison group.

The Cochrane Library distinguishes three types of PICO: (1) *Included Study PICO* which characterises an individual included study; (2) *Systematic Review PICO* which is a combined PICO for all studies included in the systematic review; and (3) *Comparison PICO* which is created as part of the quantitative analysis during scientific synthesis.[3]

[3] https://www.cochranelibrary.com/about-pico

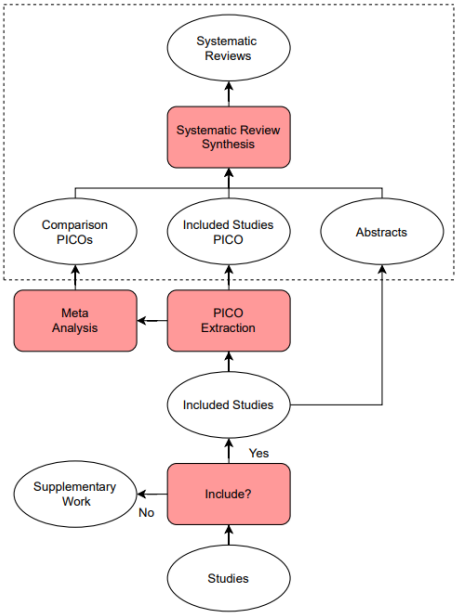

Figure 1: Steps in manual or automated systematic review synthesis.

One systematic review may contain multiple Comparison PICOs, each grouping different subsets of the data from Included Study PICOs to answer specific questions. Comparison PICOs are displayed as forest plots (see Figure 8). More details about the different types of PICO is provided in Appendix A.

3 Related Work

Synthesis from abstracts. Wallace et al. (2021) provide abstracts of individual studies to a BART model tasked with generating conclusions which are evaluated against the *authors' conclusions* (Section 4.2) from corresponding systematic reviews. Otmakhova et al. (2022) also address scientific synthesis from the abstracts of included studies, but augment their input data with manually annotated additional information. Shaib et al. (2023) assess the ability of GPT-3 to summarise and synthesise biomedical evidence, finding that it can provide high-quality summaries of a single paper, but not high-quality synthesis of multiple studies.

The above works all miss important information from their datasets. Wallace et al. (2021) and Shaib et al. (2023) appear to have downloaded included studies listed in systematic reviews from PubMed. When there is no PubMed link for an included study, that study is simply left out of the dataset entirely. This matters because the average systematic review in the Cochrane Library (on which both

datasets are based) is based on 5.5 included studies (Useem et al., 2015), whittled down from a much larger set of possible studies which are excluded if they are not of high enough quality. Because the remaining studies are all significant and highly relevant, omission of even a single included study is likely to have a serious impact on the synthesis stage. This limitation is not explicitly discussed, but potentially affects the findings, in these papers.

Shaib et al. (2023) collect abstracts of papers rather than studies. Because, as discussed above, studies are often reported in multiple papers, this results in some studies being represented in the input to synthesis multiple times, creating the illusion that multiple different studies have reached the same conclusions. This would certainly confuse a person trying to weigh the importance of the evidence and likely has the same effect on an LLM. This issue is not mentioned in the paper, but may in part explain the reported low performance of GPT-3 on this task.

Synthesis from PICOs. Lehman et al. (2019) created the Evidence Inference dataset to support synthesis generation from PICOs, but use the Systematic Review PICO itself rather than the Included Studies PICOs. DeYoung et al. (2020) and Labrak et al. (2023) add to this dataset, but include not only Included Reviews but also other referenced papers that did not feed into a systematic review, an over-inclusion previously criticised by Otmakhova et al. (2022). Wallace et al. (2021) automatically extract PICO information in papers. They only evaluate the downstream task of synthesis generation, and do not use gold standard PICOs to evaluate the quality of their PICO extraction. Otmakhova et al. (2022) use sentence-level PICOs rather than document-level which is not how human systematic reviewers understand PICOs.

4 A New Enriched Dataset for Systematic Review Synthesis

In this section we describe our new dataset and how it differs from datasets used in previous work. Table 1 sets out key statistics of our dataset that are discussed in more detail below.

The dataset consists of 45 systematic reviews each represented by the following fields: (1) systematic review title; (2) target text; (3) included study data structure; and (4) Comparison PICO data structure.

Each included study data structure is composed of the following fields: (1) included study reference; (2) included study title; (3) included study abstracts; (4) included study PICOs; and (5) included study risk of bias. In the following, we outline how we selected the 45 reviews, and describe the above fields in more detail.

	Inc.	Tot.	Cov.
Target summaries	45	45	100%
Inc. study ref	394	394	100%
Inc. study title	394	394	100%
Inc. study abstract	320	394	81%
Inc. study PICO	394	394	100%
Comparator PICO	829	829	100%

Table 1: Summary of our dataset. 'Inc.' column lists how many of each row are actually included in our dataset. The 'Tot.' column lists the total that would have been available to the human systematic reviewers when carrying out the synthesis. The 'Cov.' column lists the percentage of the total that is included in our dataset. 'Inc. study abstract' refers to the number of included studies that have at least one relevant paper abstract included.

4.1 Selection of Systematic Reviews

Initially, we selected the same 50 systematic reviews from the Cochrane Library used by Shaib et al. (2023) to enable direct comparison with their results. Note that we did not use Shaib et al. (2023)'s dataset itself, as it includes only LLM-generated summaries of the original abstracts, which may not faithfully capture the key information contained in them.

On closer examination, we found that three systematic reviews in Shaib et al.'s (2023) dataset focused on prognosis or diagnosis (systematic review types that do not have PICO data). We excluded these, as our research focuses on intervention systematic reviews. We also removed two duplicate reviews leaving us with a final dataset of 45 systematic reviews, encompassing 394 included studies.

We do not include Systematic Review PICOs, because they are not relevant to our task.

4.2 Target text

Each systematic review contains an abstract summarising its findings. Within these abstracts, there is an Authors' Conclusions section that encapsulates the ultimate conclusions derived from the systematic review process. Following Wallace et al. (2021), Shaib et al. (2023) and Otmakhova et al. (2022), we include the Authors' Conclusions in our

dataset as the target output text. Texts generated by the models we test are evaluated by comparing them with these target texts. In this paper we compare the generated texts to the target texts with the metrics set out in Section 6. An example of a target text is given in Appendix C.

4.3 Included Study Reference

Included study reference is a unique identifier given to each included study by Cochrane Library. An example of a study reference is "*Dorris 2017*". The same study reference is used in the Comparison PICO. The inclusion of the study reference should allow an LLM performing the synthesis to draw connections between the references to studies in the Comparison PICO and the included study information.

4.4 Included Study Abstracts

Each systematic review contains a list of included studies, and for each of these, a list of papers based on it. Previous approaches included information at the paper, rather than the study, level, resulting in the inclusion in datasets of multiple papers based on the same individual study. This may bias models by giving too much weight to a single piece of research merely because the authors published multiple papers based on it. We reviewed a sample of different paper abstracts related to the same study and found that they were very repetitive. For this reason we choose only one abstract / title pair from the papers to represent the included study, from the Cochrane Library itself where available, otherwise from PubMed and then the linked journal. If multiple abstracts were available, we chose the first.

As noted in Section 3, datasets from prior work contain significant gaps in included studies. We went to significant effort to improve over this, but full coverage was not possible. For certain papers, no data (other than the citation details of the title, authors etc.), not even abstracts, were available online. Usually they were not available because they were behind paywalls. Nevertheless, we substantially increase the coverage of underlying studies. Where Shaib et al. (2023) contained 239 summarised abstracts related to our 45 systematic reviews, our dataset includes 320 full abstracts. In addition, we properly distinguish papers and studies, including one abstract per *study*. Where Shaib et al. (2023) includes 239 relevant abstracts, they only cover 200 of the 394 relevant studies. Ultimately, our abstracts cover 81% of the 394 underlying studies whereas Shaib et al.'s (2023) dataset covers 50% of the underlying studies.

4.5 Included Study PICOs / Risk of Bias

Each systematic review contains one PICO for each included study. An example PICO is included in Figure 9. In addition to the main elements of the Included Study PICO, each Included Study also contains a Risk of Bias element which we also include in our dataset. An example Risk of Bias element is included in Appendix D. Our dataset has 100% coverage of Included Study PICOs ensuring that information about all 394 studies included in the systematic reviews in our dataset is represented.

4.6 Comparison PICOs

Each systematic review typically includes a series of forest plots, which are invaluable tools for synthesising data. These plots provide a concise and visual summary of the results, enabling readers to quickly assess the consistency of findings across studies, the overall effect size, and the precision of the estimates (see Figure 8 in Appendix I for typical layout and features of a forest plot).

The information in forest plots is referred to as *Comparison PICOs* (Section 2). Much of it is stored in Scalable Vector Graphics (SVG) format. While it appeared to us that the SVG format is quite easily readable by LLMs, if we had included the SVG data in its totality in our prompt it would have increased our input token count substantially. Instead, we use Claude 3 Haiku (Section 5.2) to extract the key information from the SVGs as a preprocessing step and include both its output and the original SVG data in our dataset.[4] Forest plots also include a risk of bias section specific to each included study. We extract this information and include it in the comparison PICO section. See Appendix I for more details on preprocessing forest plot SVG files. See Appendix E for an example of a reconstructed comparison PICO.

The number of Comparison PICOs provided for each systematic review can vary substantially. Three systematic reviews in our dataset did not contain any Comparison PICO, because there were no direct comparisons between the relevant included studies. One study in contrast contained 80 Comparison PICOs. The average number of Compar-

[4]Note that Claude 3 Haiku, Claude 3 Sonnet and Claude 3.5 Sonnet were accessed through the Anthropic API at: https://api.anthropic.com/v1/messages

ison PICOs in a systematic review in our dataset was 18.4 and the median was 14.

5 Biomedical Synthesis Generation via LLM Prompting

The complete biomedical synthesis generation task takes as input (a) a research question, and (b) a repository of papers, and produces as output a text representing the answer to the question based on the scientific consensus as evidenced by the papers in the repository.

In the work presented here, we address part of the complete task. Rather than starting from the raw papers, we avail of the meta-information available in the Cochrane Library, to test what performance can be achieved when such information is available in high-quality form (here human produced).

Our basic approach is to put (a) the research question, and (b) information representing each included study we wish to use as evidence to an LLM in a prompt, and interpret the LLM response as the answer to the question. More specifically, for each systematic review in our dataset, we use its title (e.g. *Care delivery and self-management strategies for children with epilepsy*) as the question, and the key information from all Included Studies as the evidence set (as illustrated in Figure 2 and described in Section 5.1). To evaluate the quality of the answer (synthesis) generated by the model, we compare it to the human-authored synthesis (the Author Conclusion section) from the systematic review.

Our aim is to improve over previous approaches by including more complete, more detailed and higher quality information about each included study (as set out in Section 4), along with detailed instructions based on textbook guidelines about how to conduct a systematic review, in the prompt to the LLM. Below we describe prompt composition (Section 5.1), and the LLMs we test (Section 5.2).

5.1 Prompt construction

Figure 2 is a flow diagram illustrating how we construct our prompts to the LLM. The boxes shaded in blue indicate prompt components that we tested in our selective ablation study for their impact (Section 7). A complete prompt has the following structure:

> Base prompt (Part 1) + Included Study Information + Comparison PICO Information + Base Prompt (Part 2) + Guidelines + Examples

We outline each of the above prompt components below.

Base prompt (Part 1): Part 1 of our Base Prompt is as follows: *"You are a systematic reviewer tasked with synthesizing information from multiple clinical studies. Below is the data you need to review. The title of the systematic review is: {Systematic Review Title}".*

Included Study Information: The Included Study Information is made up of four components which are concatenated together: Included Study Reference, Included Study Title, Included Study Abstract and Included Study PICO (which includes Risk of Bias as described in Section 4.5).

Comparison PICO Information: As described in Section 4.6, Comparison PICOs are a key tool for synthesizing research findings. We therefore include them as a separate component after the included study information is provided. An example Comparison PICO in included in Figure 8.

Base Prompt (Part 2): Part 2 of our Base Prompt reads as follows: *"What does the above evidence conclude about {Systematic Review Title}?"*

Guidelines: The guidelines component consists of summary excerpts from Cochrane Library's systematic review guidelines (Higgins et al., 2023) and instructs the model on the structure and depth of analysis expected. See Appendix F for the full prompt.

Examples: We then incorporate three gold output examples into the prompts to provide clear indication of the desired summary style and content. Note that these examples were selected from systematic reviews that are not included in our dataset, but like the systematic reviews in our dataset involved the study of interventions (as opposed to diagnosis or prognosis). Candidate examples were split into three categories based on their outcomes: (1) there was no definitive benefit to the intervention; (2) there is not enough evidence to reach a conclusion; (3) there is a benefit to the intervention. One example from each category was then randomly picked in an attempt to not bias the LLM to favor one type of conclusion over another. These three examples were used throughout all of our experiments. As we do not provide the corresponding inputs, we consider our approach to be *zero-shot* rather than *few-shot*.

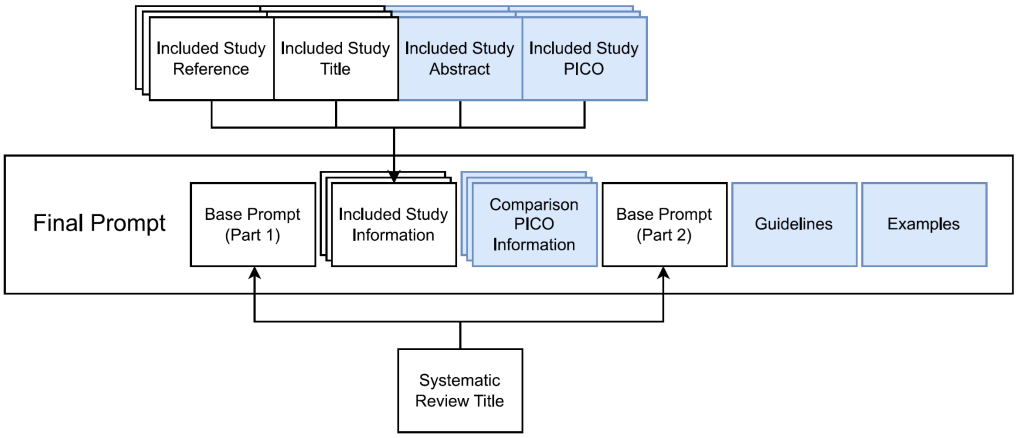

Figure 2: Flow diagram illustrating the composition of our final input prompt for a single biomedical scientific synthesis generation. The coloured boxes represent components that are removed in some of our experiments.

5.2 Models

We use Claude 3 Haiku, Claude 3 Sonnet and Claude 3.5 Sonnet from Anthropic[5] (which we refer to as Haiku, Sonnet and Sonnet 3.5 respectively) to conduct experiments. These models have a context window size of 200k tokens and are notable for their strong recall over long context lengths. Claude 3 models are trained on a mix of public internet data[6], non-public third-party data, labeled data, and internal data generated by Anthropic. They employ training techniques such as pretraining on large diverse data for next word prediction, as well as reinforcement learning with human feedback that encourage helpful, harmless, and honest responses. Additionally, they use Constitutional AI (Bai et al., 2022) to align Claude with human values during reinforcement learning, explicitly specifying rules and principles based on sources like the UN Declaration of Human Rights. We access these models through the Anthropic API. For a cost breakdown see Appendix K. On a variety of benchmarks Sonnet 3.5 is the strongest model, Sonnet is the second strongest and Haiku is the least strong.

6 Evaluation Methods

As outlined in Section 4.2, we use the Authors' Conclusions from the abstract of the target systematic review as our reference text to assess the performance of our system. The aim is to quantify the agreement between conclusions drawn in our generated biomedical syntheses with those in this reference text. Below we describe metrics we used to evaluate the agreement between the two texts.

LLM Judge. We use LLM-as-Judge as our primary approach to evaluation, as metrics based on it have been shown to have the highest correlation with human judgements in multiple studies (Wang et al., 2023; Sottana et al., 2023; Zheng et al., 2024). We use the most recent version of GPT (GPT-4o)[7] as our LLM-as-Judge.

More specifically, the judge LLM is provided with the reference text and the generated text, and is instructed to determine whether the generated text agrees or disagrees with the conclusions in the reference text. It is instructed to set out its reasoning first, and then give a score as a number between 1 and 4, with the following meaning: 1 (Strongly Disagree), 2 (Disagree), 3 (Agree), 4 (Strongly Agree). This scoring is similar to the scale used in Shaib et al. (2023). We report both the average LLM Judge Score and *Agreement Percentage* which is the percentage of generated syntheses that are scored *Agree* and *Strongly Agree* (Table 2).

We tuned the prompt for our LLM-as-Judge metric to make sure it would give the appropriate response for a set of four synthetic examples that we designed to match the four different scores above. We iterated on the prompt design until the scores assigned by the LLM matched what we expected to see for each of our synthetic inputs. The temperature for the LLM-as-Judge calls was set to 0, intended to ensure reproducibility. However, in our experiments we noted that rerunning with the same

[5] https://www.anthropic.com/news/claude-3-family

[6] Training data contained information up to August 2023 for Haiku and Sonnet. Training data contained information up to April 2024 for Sonnet 3.5.

[7] We accessed GPT-4o through the OpenAI API at: https://api.openai.com/v1/models

prompt and temperature 0 does not always produce the same response. Therefore, we run each call three times for each generated summary and assign the majority score. If the model assigned three different scores, we instead assign a zero score indicating lack of agreement. This happened four times in 45 systematic reviews over the 15 experiments listed in Table 2. When this occurred, we reran the entire experiment. A single rerun was enough in each case to eliminate zero scores.

The final prompt was as follows: "*You are to judge the quality of the output of an automatically generated 'Author's Conclusion' section for a biomedical systematic review. The user will provide the gold standard reference text and the generated text. You will use the submit_analysis tool to provide your analysis. Reference Summary: {reference} Generated Summary: {generated}*"

The model is required to submit its response in a structured JSON format using the function calling feature of the OpenAI API.[8] The model must submit the reasoning for its answer first and then a score between 1-4 as described above. The description of the function given to the model is: "*Accepts analysis of generated text against reference text.*" The description of the reasoning parameter is "*The reasoning of the reviewer about whether the generated text agrees or disagrees with the conclusions in the reference text.*" The description of the score parameter is "*Give the result as a number 1-4 meaning: 1: Strongly disagree, 2: Disagree, 3: Agree, 4: Strongly Agree.*" We require the model to fill out the reasoning parameter first to avail of the benefits of chain-of-thought reasoning (Wei et al., 2022) and also to aid in analysis of the model's ultimate decisions.

Other Automatic Metrics. Following established practice, we also employ a range of string-similarity metrics to assess the quality of our generated texts, specifically: BLEU (Papineni et al., 2002), ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), and ChrF score (Popović, 2015).

7 Experimental Results

Overall Results. Table 2 sets out the main results of our experiments, in terms of information included in the prompt (first five columns), the model used (sixth column) and the evaluation scores achieved (last 7 columns). Rows are ordered in descending order of Agreement Percentage.

Our 'kitchen sink' experiment which includes all of the components of our input described in Section 5.1 (abstracts, Included Study PICO, Comparison PICO, base prompt, guidelines and examples) achieved a 51% LLM Agreement Percentage with the Haiku model (row 3), 47% with Sonnet (row 5), and 44% with Sonnet 3.5 (row 7). However, in all cases, better results in terms of this metric are achieved when leaving out abstracts, as illustrated further in Figure 3. With Haiku and Sonnet 3.5, leaving out the abstract produced the overall best result of 53% Agreement Percentage (row 1, 2); for Sonnet it was 49% (row 4).

Initially, we hypothesised that this could be due to the fact that our dataset only has 81% coverage of abstracts. We wanted to test whether the same effect would happen if there was 100% coverage of abstracts. We proceeded to conduct two more experiments to further research this surprising result. Note that these experiments are not listed in Table 2 as the dataset used is different and therefore the results are not directly comparable. Of our 45 systematic reviews, 21 contain 100% coverage of abstracts. We again employed our 'kitchen sink' approach to this new filtered dataset. Again, we find that including abstracts still had a small negative effect on performance. Using Haiku as our model, we achieved an Agreement Percentage of 42.85% with an average LLM Judge Score of 2.523. This compares to an increased agreement score of 47.61% and an LLM Judge Score of 2.571 when abstracts were excluded.

Regarding the strength of the model, in all cases, there is only a slight difference in performance between models. This is despite the fact the Claude 3.5 Sonnet is generally considered to be a far stronger model (Section8).

Impact of PICO Components. We further assessed the impact on Agreement Percentage of including different types of PICO elements. We tested four configurations: (1) both Included Study PICO and Comparison PICO (row 3 in Table 2); (2) Included Study PICO only (row 7); (3) Comparison PICO only (row 9); and (4) neither PICO (row 10). In all of these experiments, Claude 3 Haiku was the model, abstracts were included, and our otherwise full prompt was used (base prompt + guidelines + examples).

The results show that the combination of both In-

[8] https://platform.openai.com/docs/guides/function-calling

	Info included in prompt					Model	N-gram Metrics					LLM Metrics	
	Abs	PICO		Prompt			BLEU	R-1	R-2	R-L	chrF	LLM Sc.	Agr. Per.
		Inc.	Comp.	Guide.	Exam.								
1		✓	✓	✓	✓	Sonnet 3.5	0.358	0.288	0.078	0.247	0.478	**2.689**	**53.33**
2		✓	✓	✓	✓	Haiku	0.358	**0.291**	0.080	0.253	0.481	2.644	**53.33**
3	✓	✓	✓	✓	✓	Haiku	0.353	0.288	**0.083**	**0.255**	**0.486**	2.555	51.11
4		✓	✓	✓	✓	Sonnet	0.349	0.272	0.062	0.239	0.467	2.622	48.89
5	✓	✓	✓	✓	✓	Sonnet	0.344	0.265	0.060	0.231	0.466	2.533	46.67
6	✓	✓	✓	✓	✓	Sonnet 3.5	0.347	0.282	0.076	0.249	0.478	2.511	44.44
7	✓	✓		✓	✓	Haiku	0.344	0.267	0.064	0.232	0.466	2.444	44.44
8	✓	✓	✓	✓		Haiku	**0.383**	0.285	0.078	0.250	0.475	2.422	42.22
9	✓		✓	✓	✓	Haiku	0.341	0.281	0.071	0.246	0.474	2.288	35.56
10	✓			✓	✓	Haiku	0.343	0.265	0.063	0.229	0.463	2.022	31.11
11	✓	✓	✓		✓	Haiku	0.259	0.254	0.067	0.236	0.445	2.356	28.89
12	✓	✓	✓			Haiku	0.360	0.270	0.063	0.233	0.460	2.067	28.89
13	✓			✓		Haiku	0.243	0.253	0.063	0.230	0.430	2.356	28.89
14	✓				✓	Haiku	0.278	0.240	0.055	0.219	0.444	1.756	11.11
15	✓					Haiku	0.270	0.232	0.056	0.210	0.437	1.778	8.89

Table 2: Comparison of experiments using different models (Claude 3 Haiku or Claude 3 Sonnet) and different combinations of inputs. Abs refers to full length abstracts of Included Studies described in Section 4.4. The PICO column indicates whether PICO information was provided to the model. It is broken down into two sub-columns: Included Study PICO (Inc) and Comparison PICO (Comp). All experiments include the base prompt described in Section 5.1. The prompt column indicates which additional elements of our prompt were included and is broken down into sub-columns: 'guidelines' (Guide) and 'examples' (Exam). Additional metrics: BLEU, ROUGE-1, ROUGE-2, ROUGE-L, chrF, LLM Judge Score, and Agreement Percentage are also included. See Section J for a more detailed description of some of these automatic metrics.

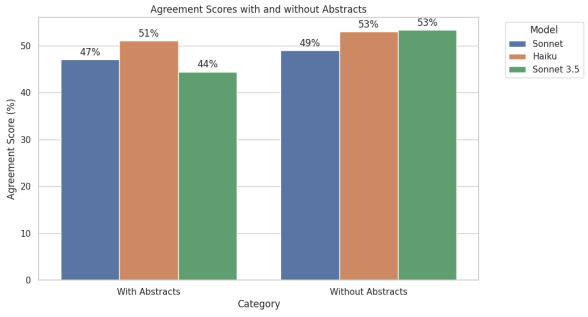

Figure 3: Comparison of Agreement Percentages with and without Abstracts for Haiku and Sonnet Models. The removal of abstracts resulted in higher Agreement Percentage for both models.

cluded Study PICO and Comparison PICO yielded the highest Agreement Percentage of 51% (row 3). When only the Included Study PICO was used, the Agreement Percentage dropped to 44% (row 7). The use of only the Comparison PICO resulted in an Agreement Percentage of 36% (row 9), and the absence of any PICO elements led to the lowest score of 31% (row 10). In combination, these results indicate that the Included Study PICO may be a more crucial part of the puzzle than the Comparison PICO, but it could also be due to the high variability in the number of Comparison PICOs for different systematic reviews.

In terms of inclusion of guidelines and examples, Table 2 shows results for (1) both (row 10), (2) just guidelines (row 13), (3) just examples (row 14), and (4) neither (row 15), in all cases with abstract and without PICO information. The corresponding Agreement Percentages are illustrated in Figure 4. We also tested this when PICO information was included (row 3, 8, 11, 12) where the same ordering of results was found: inclusion of both guidelines and examples performs the best; guidelines only is the next highest performer; examples only is the second lowest performer and; and neither is the lowest performer.

Comparison with Shaib et al. (2023). We observed notable improvements when using our high-quality dataset compared to the dataset used by

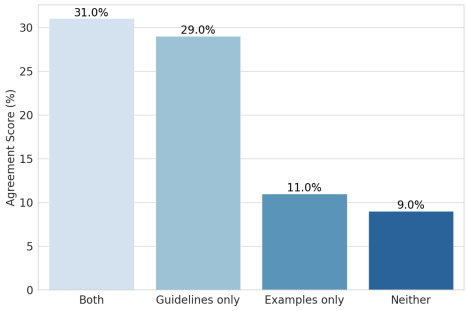

Figure 4: Impact of inclusion of guidelines and/or examples in prompt on Agreement Percentage. Note that no Included Study PICO or Comparison PICO information was included in these experiments.

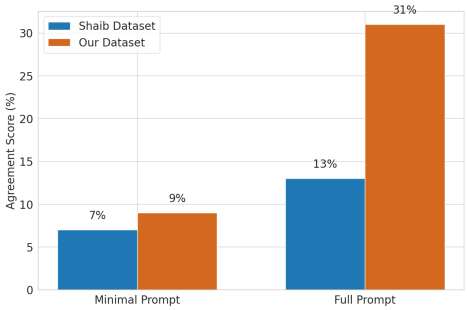

Figure 5: Impact of using Shaib et al. (2023)'s dataset vs. ours on Agreement Percentages. Bars are coloured according to the dataset used: Shaib et al. (2023) dataset vs. our dataset. The minimal prompt and full prompt configurations are compared.

Shaib et al. (2023), as illustrated in Figure 5.

The dataset in Shaib et al. (2023) contains summaries of abstracts from the same 45 Systematic Reviews as contained in our datasets. These summaries were automatically generated using an LLM. As noted above, the Shaib et al. (2023) dataset has much lower coverage of Included Study abstracts than our own dataset.

Using our base prompt only, along with the abstract summaries from Shaib et al.'s (2023) dataset, the system reaches a 7% agreement percentage. When we substitute the full abstracts from our dataset this rises to 9%. However, when we applied our full prompt (base prompt + guidelines + examples), the improvement was much more substantial, with the Agreement Percentage jumping from 13% to 31%. This demonstrates the combined effect of a high-quality dataset and a comprehensive prompt in enhancing the synthesis process.

Correlations between metrics. We found that there was little correlation between the Agreement percentage scores produced by the LLM, and the other automatic metrics that we used. For the full correlation matrix, see Appendix J.

8 Discussion

Model Strength. Interestingly, Haiku outperformed Sonnet despite the Sonnet model being superior across multiple benchmarks. Sonnet 3.5, generally the strongest of the three models, had a more mixed result; it matched the Agreement Percentage of the other highest scoring model (Haiku) when abstracts were included but performed worse than both other models when abstracts were not included. This result suggests that model strength may not be the primary bottleneck for this task. The difference in performance between these models was not statistically significant (as measured by a chi-squared test on the binary Agreement Percentages, see Appendix L) but they do indicate that, with the right approach, strong performance can be achieved with more cost efficient models.

Dataset Quality. Our results show that our improved prompting strategy has a small impact when applied to prior datasets with much lower coverage of abstracts from Included Studies. However, when applied to our more comprehensive dataset, the same prompts are more effective. With our best prompt and the abstracts from our dataset, the Agreement Percentage is 31%, compared to 13% with the Shaib et al. (2023) dataset.

9 Conclusion

In the study reported in this paper, we leveraged LLMs to generate biomedical scientific syntheses by incorporating diverse types of crucial information from included studies as input. We evaluated our approach using a carefully constructed dataset that addresses limitations of existing datasets. Our results show that we can improve over previous approaches and guide models to produce higher-quality output by providing them with included study PICO information, as well as crafting structured prompts incorporating instructions informed by domain knowledge gleaned from textbooks. It seems likely that further performance improvement can be achieved by further developing the prompt design. However, an important focus for future research will need to be the confident *automatic* extraction of relevant information from studies and papers for incorporation into such prompts.

10 Limitations

Prompting Strategy. The largest improvement in our experiments came from using a better prompting strategy (see rows 10, 13, 14 and 15 of Table 2). This improvement was achieved without conducting any rigorous evaluation of prompting strategies. Our intuition was that providing a summary of certain parts of the Cochrane Library Handbook would increase performance. This did lead to a statistically significant improvement (Section L). This suggests that there is likely more low-hanging fruit in this area. A more systematic approach may lead to even greater increases in performance. Examples of advanced prompting strategies include chain of thought (Wei et al., 2022), tree of thought (Yao et al., 2024), graph of thought (Besta et al., 2023), prompt evolution (Fernando et al., 2023) and automated prompt optimisation (Yang et al., 2023). These avenues are left for future investigation.

Automatic Metrics. Our findings indicate that basic n-gram-based metrics are inadequate for assessing LLM-generated summaries (Wallace et al., 2021). They fail to capture the intended message and content of the summaries. In this study, we leverage the LLM-as-Judge approach to approximate human judgments. A detailed analysis of the basic automatic evaluation metrics and their correlation with the LLM-as-Judge model can be found in Appendix J. Future research directions include validating the reliability of our LLM-as-Judge model through expert evaluations from domain specialists.

LLM-as-Judge. Further work needs to be done on standardising the approach to using LLM-as-Judge for evaluating automatically generated biomedical synthesis text. Our LLM-as-Judge was designed to be highly stringent. For example, when we put Shaib et al.'s (2023) outputs through our LLM-as-Judge evaluator, the results showed a striking 0% agreement with the reference conclusions. These are the outputs that Shaib et al. (2023) generated using their dataset of summarised abstracts and using GPT-3 as the LLM for synthesis.

This stands in stark contrast to the nearly 50% agreement given by human annotators reported by Shaib et al. (2023). These human annotators had medical training and were recruited on Upwork. A review of a sample of the differences indicates that our LLM-as-Judge evaluator is applying a much higher standard than the human evaluators. See Appendix H for examples of the score given to generated summaries in comparison to human annotators from Shaib et al. (2023) study. For the reasons set out in this paper (Section 6) we believe that we have calibrated the LLM-as-Judge to the appropriate level of strictness given the importance of accuracy in this task. However, future work should look to reach a consensus on how exactly the strictness of these systems should be calibrated to ensure that results are comparable across studies.

Abstracts vs. PICO. Including abstracts in the input data, to our surprise, decreased the scoring of our synthesised outputs when all of our other inputs were included (see rows 1-6 of Table 2). We hypothesise two reasons why this could be the case. First, abstracts tend to be more verbose and less focused than PICO elements, which cut straight to the essential information. Second, including abstracts increases the context length, which is known to degrade the performance of LLMs (Beltagy et al., 2020, Tay et al., 2022, Brown et al., 2020). Due to the only minor decrease in performance, we suggest future work should focus on obtaining a dataset with 100% coverage of these abstracts and retesting this theory to prove it with statistical significance (Section L).

Gold PICO information. In this study, we concentrate on generating systematic syntheses based on gold-standard PICO information extracted by human experts from the Cochrane Library. While this approach provides high-quality input, a more pragmatic setup would involve using automated systems to extract PICO information. We consider this avenue a promising direction for future research.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski,

Piotr Nyczyk, et al. 2023. Graph of Thought: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence Inference 2.0: More Data, Better Models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*.

Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian Welch. 2023. *Cochrane Handbook for Systematic Reviews of Interventions*, 6.4 edition. Cochrane. Accessed: 09-02-2024.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. 2023. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. *arXiv preprint arXiv:2307.12114*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Matthew Michelson and Katja Reuter. 2019. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary clinical trials communications*, 16:100443.

Julia Otmakhova, Karin Verspoor, Timothy Baldwin, Antonio Jimeno Yepes, and Jey Han Lau. 2022. M3: Multi-level dataset for Multi-document summarisation of Medical studies. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3887–3901.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

S Rosumeck, A Nast, and C Dressler. 2018. Ivermectin and permethrin for treating scabies. *Cochrane Database of Systematic Reviews*, (4).

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, Simplifying, and Synthesizing Medical Evidence using GPT-3 (with Varying Success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1387–1407, Toronto, Canada. Association for Computational Linguistics.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: reliably evaluating large language models on sequence to sequence tasks. *arXiv preprint arXiv:2310.13800*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28.

Johanna Useem, Alana Brennan, Michael LaValley, Michelle Vickery, Omid Ameli, Nichole Reinen, and Christopher J Gill. 2015. Systematic differences between Cochrane and non-Cochrane meta-analyses on the same topic: a matched pair analysis. *PloS one*, 10(12):e0144980.

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.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? A preliminary study. *arXiv preprint arXiv:2303.04048*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Types of PICO

As an example, consider a systematic review evaluating the effectiveness of ivermectin and permethrin for treating scabies. This example is based on Rosumeck et al. (2018), focusing on only the first two included studies mentioned therein. One included study examines four different interventions: ivermectin, permethrin, benzyl benzoate, and sulfur ointment. The second included study evaluates ivermectin as a treatment.

The Included Study PICO for these two studies will differ. The first study lists four interventions, while the second study lists only one.

The Systematic Review PICO is defined at the beginning of the systematic review process and determines its scope. In this case, the systematic review only considers ivermectin and permethrin as interventions, so only these two interventions are included as *Interventions* in the Systematic Review PICO. The systematic review ignores the results related to benzyl benzoate and sulfur ointment from the first study because they fall outside of its scope.

The systematic review may contain a Comparison PICO comparing the results of ivermectin as an intervention. The Comparison PICO would include the results related to ivermectin from both studies. Thus, the *Interventions* component of the Comparison PICO is only comparing one intervention: ivermectin.

This example illustrates how the different types of PICO relate to each other, focusing on the *Interventions* element. The same principles apply to the other elements of PICO as well. Systematic Review PICOs set the scope for the review. Included studies may contain information outside this scope or information that is only a subset of the Systematic Review PICO. This difference is reflected between the Included Study PICO and the Systematic Review PICO. Comparison PICOs focus on specific sub-questions and will include only the subset of PICO information from included studies relevant to the question.

B Agreement Scores of Strongest Performing Setups

Figure 6 and Figure 7 show the level of agreement between the different setups.

C Target Text Example

The following is an example of one of the target texts in our dataset:

> *Group CBTp appears to be no better or worse than standard care or other psychosocial interventions for people with schizophrenia in terms of leaving the study early, service use and general quality of life. Group CBTp seems to be more effective than standard care or other psychosocial interventions on overall mental state and global functioning scores. These results may not be widely applicable as each study had a low sample size. Therefore, no firm conclusions concerning the efficacy of group CBTp for people with schizophrenia can currently be made. More high-quality research, reporting useable and relevant data is needed.*

D Risk of Bias example

- Random sequence generation (selection bias): Low risk
- Allocation concealment (selection bias): Low risk
- Blinding of participants and personnel (performance bias): All outcomes High risk
- Blinding of outcome assessment (detection bias): All outcomes Low risk
- Incomplete outcome data (attrition bias): All outcomes Low risk
- Selective reporting (reporting bias): Low risk
- Other bias: Low risk

E Reconstructed Comparison PICO

Comparison 1: Seizure frequency and severity, Outcome 1: Number of seizures at 12 months

- Meta-analysis:
- Study or Subgroup

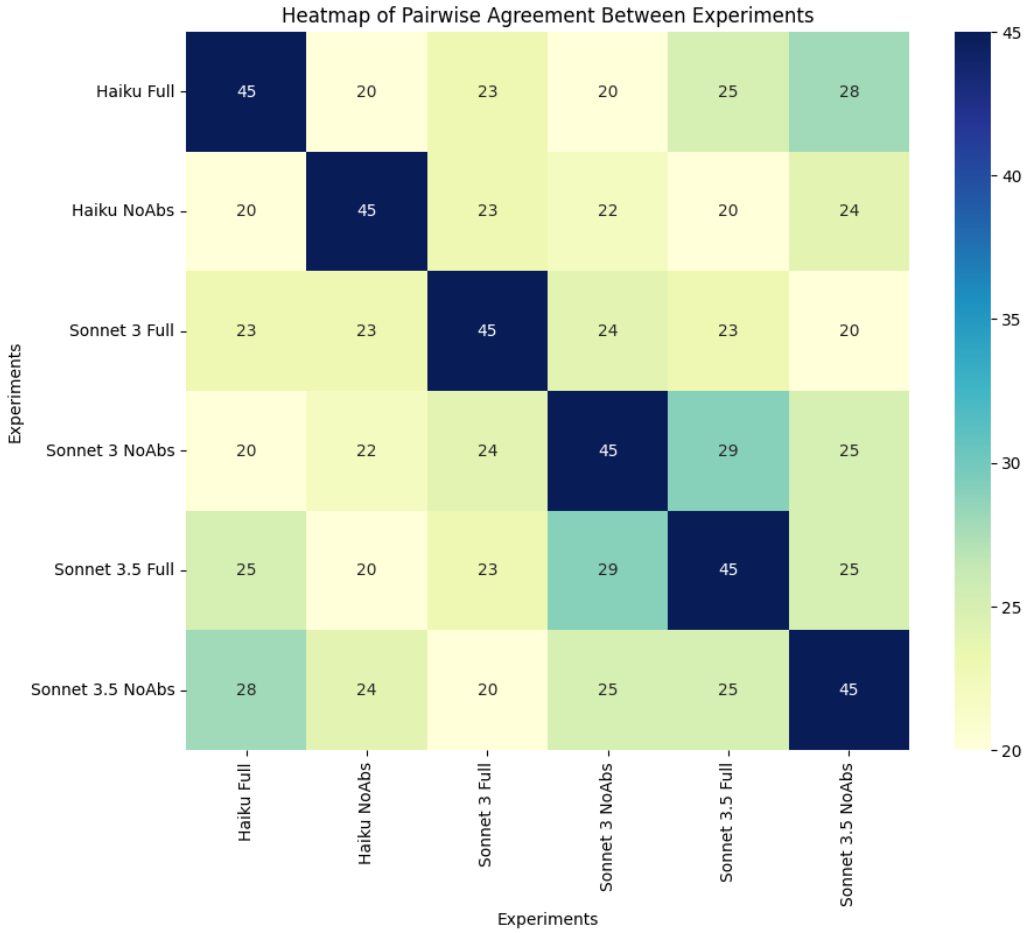

Figure 6: Heatmap showing pairwise agreement between experiments based on LLM Judge Scores which rank between 1 and 4. Darker colors indicate higher agreement between experiments.

- Tieffenberg 2000
- Experimental: Mean: 0.34, SD: 0.98, Total: 103
- Control: Mean: 1.11, SD: 2.77, Total: 64
- Mean Difference: IV, Fixed, 95

Risk of Bias:

- A: ?
- B: ?
- C: -
- D: ?
- E: -
- F: +
- G: ?

F Guidelines prompt

The guidelines prompt is as follows: "*The following is a summary of the instructions given to Cochrane Reviewers for drafting the Authors' Conclusions section of a systematic review: Implications for Practice: Cochrane Reviews provide valuable information for practice but do not make direct recommendations due to the need for additional evidence and judgments. Authors should discuss the certainty of evidence, benefits versus harms, and patient values/preferences without making specific recommendations. If authors discuss possible actions, they should consider all factors influencing decisions, including patient-important outcomes, costs, and resource availability. Implications for Research: This section highlights the need for further research and specifies desirable research characteristics. Authors should use the PICO framework (Population, Intervention, Comparison, Outcomes) to detail areas needing more investigation. The GRADE framework helps in understanding*

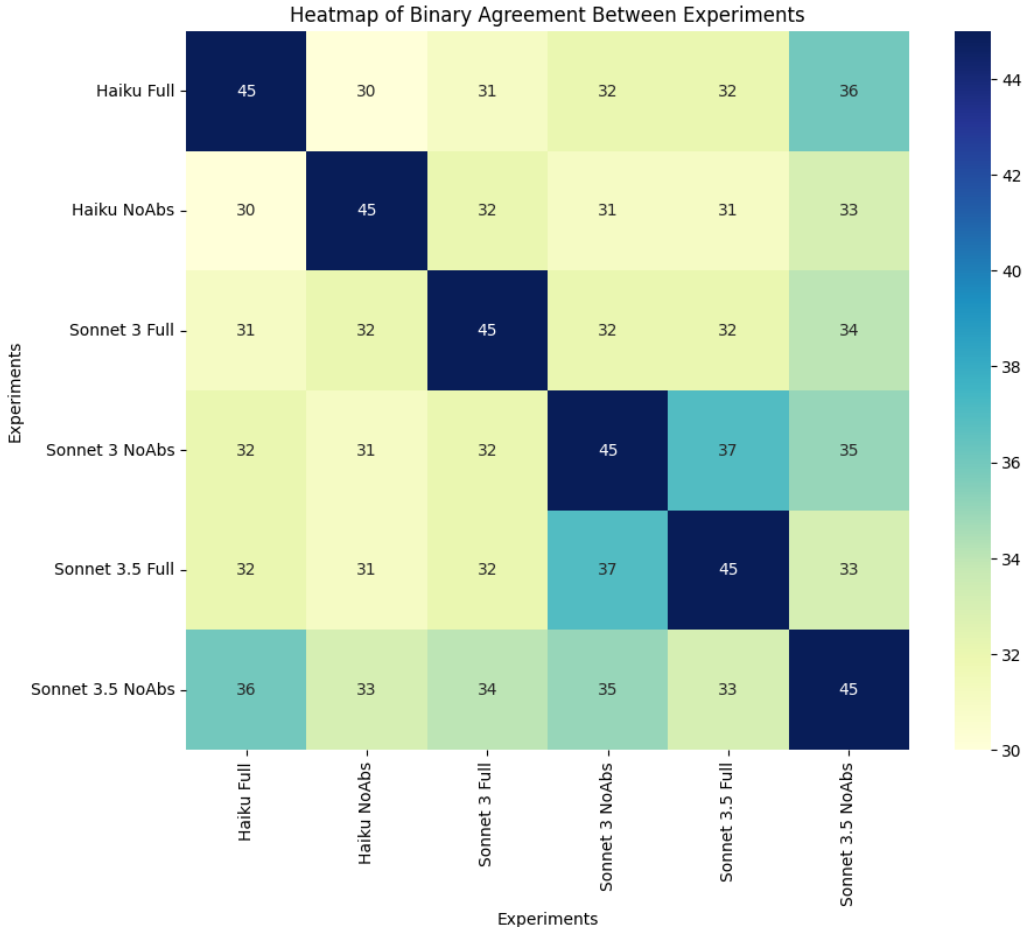

Figure 7: Binary heatmap representing the agreement between experiments. Each cell shows whether pairs of experiments agreed or disagreed based on binary classification (agree: scores 3 or 4, disagree: scores 1 or 2). Darker colors indicate higher agreement between experiments.

how further research could improve evidence certainty. Implications by GRADE Domains:

Risk of Bias: Call for better-designed studies. Inconsistency: Need for studies in relevant subgroups to understand differences. Indirectness: Studies that better fit the PICO question. Imprecision: More studies with larger participant numbers. Publication Bias: Investigate unpublished data and conduct large studies. Large Effects and Dose Effects: No direct research implications but large effects likely reflect the true impact. Opposing Bias and Confounding: Studies controlling for residual biases and confounders.

You are to draft the summary of the Author's Conclusions that is to go in the abstract. It should be no more than 200 words."

G Examples prompt

The following is our Examples prompt: "*Below are some example outputs to give you an indication of style, length and what information is to be included: Title: Antiviral medications for preventing cytomegalovirus disease in solid organ transplant recipients Author's conclusions: Prophylaxis with antiviral medications reduces CMV disease and CMV-associated death, compared with placebo or no treatment, in solid organ transplant recipients. These data support the continued routine use of antiviral prophylaxis in CMV-positive recipients and CMV-negative recipients of CMV-positive organ transplants.*

Title: Magnesium sulphate for women at risk of preterm birth for neuroprotection of the fetus Author's conclusions: The currently available evidence indicates that magnesium sulphate for women at risk of preterm birth for neuroprotection of the fetus, compared with placebo, reduces cerebral palsy, and death or cerebral palsy, in children up to two years' corrected age, and probably reduces severe intraventricular haemorrhage for infants. Magnesium sulphate may result in little to no difference in outcomes in children at school age.

While magnesium sulphate may result in little to no difference in severe maternal outcomes (death, cardiac arrest, respiratory arrest), it probably increases maternal adverse effects severe enough to stop treatment. Further research is needed on the longer-term benefits and harms for children, into adolescence and adulthood. Additional studies to determine variation in effects by characteristics of women treated and magnesium sulphate regimens used, along with the generalisability of findings to low- and middle-income countries, should be considered.

Title: Early pharmacological interventions for prevention of post-traumatic stress disorder (PTSD) in individuals experiencing acute traumatic stress symptoms Author's conclusions: This review provides uncertain evidence regarding the use of escitalopram, hydrocortisone, intranasal oxytocin and temazepam for people with acute stress symptoms. It is therefore unclear whether these pharmacological interventions exert a positive or negative effect in this population. It is important to note that acute traumatic stress symptoms are often limited in time, and that the lack of data prevents the careful assessment of expected benefits against side effects that is therefore required. To yield stronger conclusions regarding both positive and negative outcomes, larger sample sizes are required. A common operational framework of criteria for inclusion and baseline assessment might help in better understanding who, if anyone, benefits from an intervention. As symptom severity alone does not provide the full picture of the impact of exposure to trauma, assessment of quality of life and functional impairment would provide a more comprehensive picture of the effects of the interventions. The assessment and reporting of side effects may facilitate a more comprehensive understanding of tolerability."

H Example of GPT Annotator vs Shaib Annotator

LLM Judge Score: Strongly disagree
Shaib Annotator Score: Agree
LLM Judge Reasoning: The generated summary is overly positive and does not reflect the variability and methodological issues highlighted in the reference summary. The reference summary emphasizes the lack of sufficient evidence and the need for further research, while the generated summary suggests a more definitive positive impact of the interventions without acknowledging the limitations.
Generated Summary:
The evidence suggests that tailored interventions and educational programs can have a positive impact on the care delivery and self-management of children with epilepsy. These interventions can improve knowledge and understanding of the condition, reduce anxiety, improve decision-making skills, and improve adherence to treatments. Additionally, child-centered, family-focused interventions can reduce crises and visits to physicians, decrease fears of death, reduce disruption of family life and patient-physician relationships, and reduce school absenteeism.
Target Summary:
Whilst each of the programmes evaluated in this review showed some benefit to children with epilepsy, their impact was extremely variable. No programme showed benefits across the full range of outcomes, and all studies had methodological problems. There is currently insufficient evidence in favour of any single programme. Further evidence from randomised controlled trials using validated measures and considering clinical meaningfulness as well as statistical significance of results is required.

LLM Judge Score: Disagree
Shaib Annotator Score: Agree
LLM Judge Reasoning:
The generated summary partially captures the essence of the reference summary but introduces several inaccuracies and omissions. The reference summary emphasizes the reduction in antibiotic prescriptions due to C-reactive protein point-of-care tests without affecting recovery rates, and it highlights the need for further research in specific populations and on other biomarkers. The generated summary, however, introduces elements not mentioned in the reference, such as enhanced communication skills and Internet-based training, and it incorrectly states that rapid tests for C-reactive protein did not reduce antibiotic prescriptions. These discrepancies lead to a lower score.
Generated Summary:
The evidence suggests that using C-reactive protein point-of-care testing and training in enhanced communication skills can reduce antibiotic prescribing for lower respiratory tract infections in general practices without compromising patient recovery and satisfaction with care. Additionally, Internet-

based training in C-reactive-protein point-of-care testing, communication skills, or a combination of the two can substantially reduce antibiotic prescribing for lower RTIs. However, point-of-care procalcitonin and lung ultrasonography were not found to further reduce antibiotic prescription, and the use of a rapid test for C-reactive protein did not reduce prescription of antibiotics.

Target Summary:

The use of C-reactive protein point-of-care tests as an adjunct to standard care likely reduces the number of participants given an antibiotic prescription in primary care patients who present with symptoms of acute respiratory infection. The use of C-reactive protein point-of-care tests likely does not affect recovery rates. It is unlikely that further research will substantially change our conclusion regarding the reduction in number of participants given an antibiotic prescription, although the size of the estimated effect may change. The use of C-reactive protein point-of-care tests may not increase mortality within 28 days follow-up, but there were very few events. Studies that recorded deaths and hospital admissions were performed in children from low- and middle-income countries and older adults with comorbidities. Future studies should focus on children, immunocompromised individuals, and people aged 80 years and above with comorbidities. More studies evaluating procalcitonin and potential new biomarkers as point-of-care tests used in primary care to guide antibiotic prescription are needed. Furthermore, studies are needed to validate C-reactive protein decision algorithms, with a specific focus on potential age group differences.

On average the LLM judge scored the outputs one score lower than the Human annotators from (Shaib et al., 2023).

I SVG Forest Plot Reconstuction

Obtaining the comparison PICO proved to be quite a challenge. In Cochrane Library, for our dataset, there could be up to 88 comparison PICOs for one systematic review. These were in the form of forest plots. These forest plot are saved as images on a surface level, but when accessing the html code we find that the images can be accessed to obtain their svg data. Note that for a few of the forest plots they are actually saved as images but it is a very small subset. The svg data means that we were able to scrape the forest plots quite easily but due to the wide variety of formats and layouts, reconstructing these using code alone would have been extremely challenging. We found that when giving this svg data to an LLM, it was able to read and reconstruct it with ease. This means that if we wanted to we could have incorporated the entire SVG data into the prompt of our synthesiser, however due to their massive length, this would have increased cost, computation time and would have exceeded the token limits of almost every model available. We employed the novel approach of using LLMs to reconstruct the SVG data for us. This way we could accommodate the wide variety of formats while being able to reduce token length to streamline the integration of the comparison PICOs in our main experiments. We used Claude Haiku for the reconstruction after testing a variety of different LLMs on individual SVGs. Haiku provided the lowest cost for the largest token limit of the models tested. Temperature was set to 0.

Prompt used

> Extract the key information from the forest plot above. List the information in each row of the forest plot separately (this may require repeating the headings row(s)). Note that "Weight" is an independent heading (where included). Include the risk of bias information for each row (if included). Include the risk of bias legend (if included). Only include the risk of bias legend once. Do not provide a summary or analysis just provide the key information. Begin with "Meta analysis:"

This prompt was improved upon iteratively by reconstructing one forest plot at a time and addressing any issues with formatting or content that arose. Here is an example of the forest plot and the SVG reconstruction and the forest plot it refers to in Figure 8.

> Meta analysis:
> Study or Subgroup,Group CBTp Events,Group CBTp Total,Control Events,Control Total,Weight,Risk Ratio M-H, Random, 95% CI
> Barrowclough 2006,2,57,1,56,1.2%,1.96 [0.18 , 21.06]
> Bechdolf 2004,9,40,8,48,8.4%,1.35 [0.57 , 3.17]
> Chadwick 2016,6,54,9,54,6.8%,0.67 [0.25 , 1.74]

Deng 2014,18,59,18,59,18.2%,1.00 [0.58 , 1.72]
Granholm 2007,5,37,6,39,5.3%,0.88 [0.29 , 2.63]
Granholm 2013,14,41,6,38,8.6%,2.16 [0.93 , 5.05]
Granholm 2014,37,73,30,76,32.9%,1.28 [0.90 , 1.84]
Li 2013a,12,60,5,60,6.6%,2.40 [0.90 , 6.39]
Mortan Sevi 2020,2,12,7,14,3.5%,0.33 [0.08 , 1.31]
Penn 2009,5,32,1,33,1.5%,5.16 [0.64 , 41.74]
Shi 2015,4,60,2,60,2.4%,2.00 [0.38 , 10.51]
Tao 2015,1,60,3,60,1.3%,0.33 [0.04 , 3.11]
Wykes 2005,4,43,3,42,3.2%,1.30 [0.31 , 5.47]
Total (95% CI),119,628,99,639,100.0%,1.22 [0.94 , 1.59]
Total events,119,99
Risk of Bias
A,?,+,+,+,+,+,+
B,+,?,+,+,+,+,+
C,-,-,-,-,-,-,+
D,+,+,+,?,+,+,+
E,+,+,+,+,+,+,+
F,+,+,+,+,+,+,+
G,+,+,+,+,+,+,+

Risk of bias legend
(A) Random sequence generation (selection bias)
(B) Allocation concealment (selection bias)
(C) Blinding of participants and personnel (performance bias)
(D) Blinding of outcome assessment (detection bias)
(E) Incomplete outcome data (attrition bias)
(F) Selective reporting (reporting bias)
(G) Other bias

See Figure 8 for how this Comparison PICO looks in SVG format.

J Automatic Evaluation Metrics

BLEU We used the sentence-level BLEU score, which is a metric for evaluating a generated sentence to a reference sentence. BLEU measures the precision of n-grams in the generated text compared to the reference text, accounting for brevity and exact matches. Specifically, we use the sentence_bleu function from the nltk.translate.bleu_score module, which calculates BLEU scores at the sentence level. The sentence_bleu function from the nltk.translate.bleu_score module, when used as sentence_bleu([target], summary), by default calculates the cumulative BLEU-4 score. This means it considers n-grams from 1 to 4, giving you the combined score for 1-gram, 2-gram, 3-gram, and 4-gram matches between the target and the summary.

ROUGE For ROUGE, we used the py-rouge library, which provides various ROUGE metrics, including ROUGE-1, ROUGE-2, and ROUGE-L. These metrics are defined as follows:

ROUGE-1: Measures the overlap of unigrams (single words) between the generated summary and the reference summary. ROUGE-2: Measures the overlap of bigrams (two consecutive words) between the generated summary and the reference summary. ROUGE-L: Measures the longest common subsequence (LCS) between the generated summary and the reference summary. This metric captures the sequence similarity, taking into account the order of words.

chrF The character F-score (chrF) is another evaluation metric we used, which is calculated using the sentence_chrf function from the nltk.translate.chrf_score module. ChrF measures the precision and recall of character n-grams (typically 6-grams) rather than word n-grams, making it particularly effective for capturing both lexical and grammatical correctness in the generated summaries. It is useful for assessing the readability and coherence of the text at the character level, providing a complementary perspective to word-level metrics like BLEU and ROUGE.

Table 3 shows the correlation between all metrics used.

K Cost Breakdown

We spent $11 on the Anthropic API accessing the Claude 3 Haiku and Claude 3 Sonnet models. This was used for converting SVG to human readable

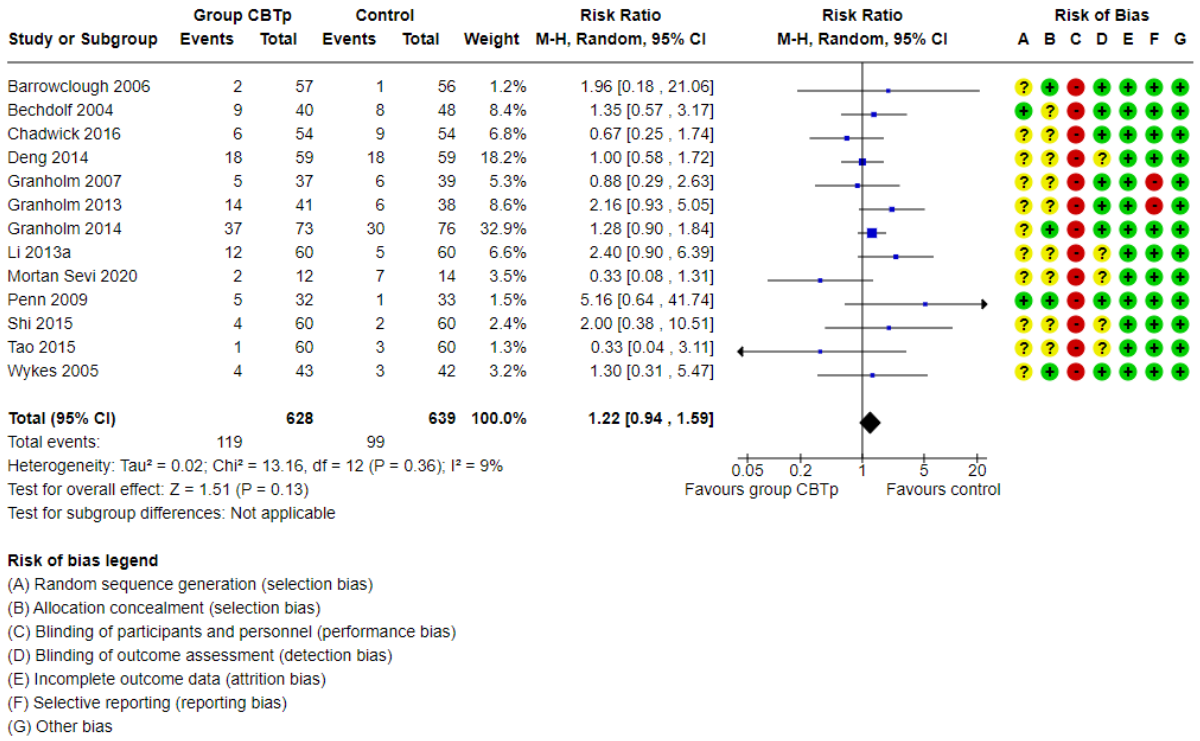

Figure 8: Example of Comparison PICO Forest Plot.

	BLEU	chrF	R-1	R-2	R-L	LLM Score
BLEU	1.000	0.368	0.486	0.256	0.413	-0.138
chrF	0.368	1.000	0.490	0.561	0.492	0.212
R-1	0.486	0.490	1.000	0.706	0.927	0.001
R-2	0.256	0.561	0.706	1.000	0.723	0.096
R-L	0.413	0.492	0.927	0.723	1.000	0.040
LLM Score	-0.138	0.212	0.001	0.096	0.040	1.000

Table 3: Correlation Matrix of Metrics. Formed by concatenating results from all experiments

text and performing the biomedical synthesis generation in our experiments.

Using GPT-4o as a judge cost $5.60.

L Statistical Tests

We implemented chi-squared tests to determine the significance level of changes in agreement percentages. The improvements we made to methodology by integrating PICO elements, enhancing prompting with guidelines and examples, and including our improved dataset all have high levels of statistical significance. However, the observed differences in performance when changing models and including or excluding abstracts cannot be said with any degree of certainty that an improvement was indeed made.

Methodology Comparison Comparing Abs Only, Full Prompt, Our Dataset, Haiku with NoAbs, PICO, ComPICO, Full Prompt, Haiku:

- **Chi2:** 3.75
- **p-value:** 0.0528
- **Confidence Interval:** 94.72%

Prompt Comparison Comparing Abs, PICO, ComPICO, Full Prompt, Haiku with Abs, PICO, ComPICO, Minimal Prompt, Haiku:

- **Chi2:** 2.99
- **p-value:** 0.0837
- **Confidence Interval:** 91.63%

Li 2013a

Study characteristics	
Methods	Allocation: randomised Blinding: not reported Duration: 9 months Design: parallel Setting: inpatient Country: China
Participants	Diagnosis: diagnosis of schizophrenia according to CCMD-3 diagnostic criteria n = 120 Gender: 67 males, 36 females (data only available for treatment completers) Age: mean: group CBTp: 37.6 years; control: 38.3 years; range not reported History: duration of illness mean: group CBTp: 5.3 years; control: 5.4 years. All participants taking concomitant medication (risperidone) Exclusion criteria: not reported
Interventions	1. Group CBTp, n = 48 The group focused on psychoeducation on schizophrenia, implementing CBT techniques, such as the application of self-monitoring and response strategies, using a voice diary, allowing the participant to recognise any sound that may have appeared and existed, encouraging the analysis of experiences and suggestions of avoidance methods; later, the participants began to use these coping strategies when they heard the sound, and to develop awareness of the disease. Sessions ran once a week for the first 3 months, once every 2 weeks after 3 months, and once a month after 6 months; each session lasted 50–60 minutes 2. Standard care, n = 55 No further information
Outcomes	Leaving study early for any reason Overall mental state (PANSS total score) Hallucinations (AHRS score)

Figure 9: Example of Included Study PICO.

Dataset Comparison Comparing Abs Only, Full Prompt, Our Dataset, Haiku with Abs Only, Full Prompt, Shaib Dataset, Haiku:

- **Chi2:** 3.40
- **p-value:** 0.0651
- **Confidence Interval:** 93.49%

Model Comparison Comparing NoAbs, PICO, ComPICO, Full Prompt, Sonnet with NoAbs, PICO, ComPICO, Full Prompt, Haiku:

- **Chi2:** 0.0
- **p-value:** 1.0
- **Confidence Interval:** 0.0%

Abstract Inclusion Comparison Comparing Abs, PICO, ComPICO, Full Prompt, Haiku with NoAbs, PICO, ComPICO, Full Prompt, Haiku:

- **Chi2:** 0.0
- **p-value:** 1.0
- **Confidence Interval:** 0.0%

Total Comparison Comparing Abs Only, Minimal Prompt, Shaib Dataset, Haiku with NoAbs, PICO, ComPICO, Full Prompt, Haiku:

- **Chi2:** 19.53

- **p-value:** 9.93e-06
- **Confidence Interval:** 99.999%

Comparison	Chi2	p-value	Confidence Interval (%)
Methodology Comparison	3.75	0.0528	94.72
Prompt Comparison	2.99	0.0837	91.63
Dataset Comparison	3.40	0.0651	93.49
Model Comparison	0.0	1.0	0.0
Abstract Inclusion Comparison	0.0	1.0	0.0
Total Comparison	19.53	9.93e-06	99.999

Table 4: Statistical Significance Results of Various Comparisons