

Evaluating Professional Acceptability of LLM-Generated Systematic Review Summaries in Healthcare: Psychiatrists' Perspectives

Paul Thompson¹, Artemis Boulogeorgou², Fotini Kaponi², Efstathia Soufleri²
and Sophia Ananiadou^{1,2}

¹National Centre for Text Mining, School of Computer Science, University of Manchester

²Archimedes/Athena Research Center, Greece

{paul.thompson, sophia.ananiadou}@manchester.ac.uk

bo.artemis@gmail.com, fotinikaponi@hotmail.gr, e.soufleri@athenarc.gr

Abstract

Cochrane systematic reviews evaluate the effectiveness and safety of medical interventions. Patients can benefit from clinicians' integration of outcomes of these reviews into their daily practices. However, systematic reviews are usually long documents; even their abstracts can extend to 1000 words, making rapid appraisal challenging for busy health professionals. Large language models (LLMs) offer potential to further distil these abstracts. Nevertheless, generating high-quality, clinician-oriented summaries in this context is non-trivial. They must comprehensively cover the original abstract, while remaining accurate and *professionally acceptable*, i.e., retaining all clinically important details. To address this challenge, we have developed a novel dataset, PsycSumEval, comprising summaries generated by four different LLMs for 115 Cochrane abstracts concerning mental health. Psychiatrists evaluated each summary across nine content dimensions, assigning scores and providing free-text justifications that highlight inaccuracies and missing details. The corpus provides fine-grained insight into how psychiatrists assess professional acceptability of compressed medical evidence. Rather than treating agreement as a merely statistical endpoint, we capture structured expert judgments alongside their rationales, enabling transparent analysis of where professional norms are stable and where interpretive latitude persists. We contribute both a rigorous evaluation dataset and an explicit model of expert acceptability criteria for medical evidence summarisation.

Keywords: Systematic Reviews, Summarisation, Evaluation Framework, Large Language Models

1. Introduction

Systematic reviews are the cornerstone of evidence-based medicine (Higgins et al., 2024; Antes et al., 2022; Dickersin et al., 1994; Chalmers and Altman, 1995; Needleman, 2002). By identifying, appraising and synthesising data from multiple trials and observational studies, such reviews provide reliable overviews of the advantages and disadvantages of medical interventions. These overviews can benefit patients, e.g., by allowing clinicians to integrate the latest research findings into their daily practices, or to support the development of clinical practice guidelines aimed at optimising patient care (Morton et al., 2011).

Since the complex nature of the review process results in long documents, busy health professionals may rely on the accompanying abstracts to acquire an understanding of the most important aspects of the review (Tsou and Treadwell, 2016). Past concerns about poor quality abstracts (Beller et al., 2011) were part of the motivation behind the development of the now widely-endorsed PRISMA guidelines (Liberati et al., 2009; Page et al., 2021a,b), which aim to ensure quality and consistency in systematic reviews of healthcare interventions. A dedicated checklist for abstracts (PRISMA-A) (Beller et al., 2013) aims to ensure

that the abstracts of such reviews can stand alone in presenting a clear and truthful account of the research. Particular journals or organisations may extend PRISMA-A with their own requirements, e.g., Cochrane systematic reviews must supplement results with assessments of evidence certainty (Antes et al., 2022).

The introduction of PRISMA-A has resulted in a steady increase in the median length of Cochrane abstracts, rising from 353 words in 2000 to 838 words in 2022 (Helbach et al., 2023b). However, evidence that guideline compliance can be achieved using far fewer words (Helbach et al., 2023a; Maticic et al., 2019) suggests that Cochrane abstracts could be further compressed, without sacrificing vital details. This would be especially beneficial for busy health professionals seeking rapid appraisal of multiple reviews.

Given the large amount of research into their use for biomedical text summarisation (Xie et al., 2023; Huang et al., 2025), large language models (LLMs) constitute an attractive option to produce summaries of Cochrane abstracts automatically. Indeed, previous work has investigated the LLM-driven generation of plain language summaries of these abstracts for non-expert readers (Devane et al., 2026; Arias-Russi et al., 2025a,b). Here, we are concerned with a different task, i.e., evaluat-

ing the suitability of LLMs to produce compressed summaries of Cochrane abstracts that remain suitable for healthcare professionals. To support critical decision making, such summaries must meet quality and acceptability criteria:

1. A high degree of **accuracy/factual correctness**. Hallucination remains one of the most significant issues in LLM-generated text (Zhang et al., 2025), and has the potential to cause significant harm in clinical decision making scenarios (Tang et al., 2023)
2. **Compliance with key aspects of PRISMA-A and Cochrane guidelines**, to ensure adequate coverage of relevant aspects of the systematic review.
3. **Professional acceptability**, meaning that compressed summaries must preserve clinically salient details required by experts.

Existing automated evaluation metrics (e.g., Lin (2004); Papineni et al. (2002); Lavie and Agarwal (2007); Zhang et al. (2019)) are insufficient to assess compliance with the above criteria, since they can only judge similarity between LLM-generated and gold-standard summaries at a general level. This motivates the development of more fine-grained and clinically-informed evaluation frameworks (Huang et al., 2025). While there is promising evidence in the general domain that LLMs could themselves replace humans as effective evaluators (Zheng et al., 2023; Dubois et al., 2024), this success does not yet extend to assessing medical evidence summaries automatically (Vasilev et al., 2026; Zhang et al., 2024; Joseph et al., 2024). Accordingly, human involvement in applying these frameworks remains irreplaceable.

We have thus developed a novel framework tailored to the evaluation of automatically-generated, expert-oriented summaries of Cochrane systematic review abstracts. The framework has subsequently been used to assess summaries generated by four different LLMs for 115 Cochrane abstracts in the mental health domain (i.e., a total 460 LLM-generated summaries). Evaluation was performed by two psychiatrists to ensure acceptability from a professional perspective. The assessment of each summary involves recording up to four pieces of information for each of nine different content dimensions, which are based on key requirements of the PRISMA-A and Cochrane handbook guidelines (Higgins et al., 2024).

2. Related Work

2.1. LLM-Generated Summaries in Biomedical and Clinical Domains

LLMs have demonstrated strong performance across a range of natural language processing tasks (Qin et al., 2024), including summarisation in the biomedical (Xie et al., 2023; Huang et al., 2025) and clinical (Bednarczyk et al., 2025) domains. Indeed, there is evidence that LLMs can sometimes outperform humans in accurately summarising medical evidence (Van Veen et al., 2024; Tang et al., 2023). Nevertheless, several concerns remain regarding the reliability of these summaries. Hallucination, i.e., the fabrication of details not present in the source text, remains a significant issue (Zhang et al., 2025). Biomedical and clinical summarisation pose additional challenges: many LLMs lack deep familiarity with specialised terminology and clinical concepts used in these domains (Huang et al., 2025), and even fine-tuned models can struggle to produce output that is sufficiently faithful and accurate (Zhang et al., 2024).

The risks become particularly acute in clinical decision-making contexts, where even seemingly small errors in the generated output could mislead readers, with potentially serious consequences (Kirmani et al., 2024; Jose et al., 2024). Tang et al. (2023) defined *harmfulness* as the potential for a generated summary to cause physical or psychological harm, or inappropriate changes in therapy, due to misinterpretation. Harmfulness in medical evidence summaries was primarily attributed to insufficient comprehensiveness (e.g., omission of Population, Intervention, Comparison, Outcome (PICO) elements (Richardson et al., 1995)) or to misinterpretation errors, including *certainty illusions*. These correspond to errors in the degree of certainty expressed towards particular findings, which could cause readers to under- or overestimate their significance.

2.2. Automated Summary Evaluation

The high-stakes nature of LLM-generated medical summaries demands robust evaluation measures. Although general automated metrics (e.g., Lin (2004); Papineni et al. (2002); Lavie and Agarwal (2007); Zhang et al. (2019)) may suffice in non-critical scenarios, e.g., plain language summaries for non-experts (Arias-Russi et al., 2025a,b), they are inadequate for assessing compliance with clinically detailed criteria. Prior work demonstrates a failure of these metrics to reliably detect factual errors, hallucinations and harmfulness in medical summaries (Tang et al., 2023; Vasilev et al., 2025).

Such limitations motivate the development of domain-specific evaluation methods that explicitly

address terminological accuracy and clinical risk (Huang et al., 2025). While such frameworks are usually applied manually, there is growing interest in automation. The *LLM-as-judge* paradigm (Zheng et al., 2023) uses LLMs to evaluate model outputs across flexible criteria at lower cost and more efficiently than human application (Yu, 2025). Although promising in general domains (Zheng et al., 2023; Dubois et al., 2024), the effectiveness of this approach in detecting factual errors and hallucinations in medical summaries remains limited, particularly at the full-summary level (Vasilev et al., 2026; Zhang et al., 2024). However, LLM judges that evaluate finer-grained domain-specific aspects (e.g., PICO elements, side effects or medications) show stronger correlation with human judgments (Chu et al., 2025; Gero et al., 2024).

2.3. Human-Annotated Datasets

Although LLM-based evaluation of medical summaries remains imperfect, fine-tuning judge models on high-quality human-annotated data is a promising direction (Kim et al., 2023; Huang et al., 2025). Recent efforts to produce suitable datasets include Luo et al. (2024), Joseph et al. (2024) and Chu et al. (2025), which evaluate LLM-generated medical evidence summaries along multiple dimensions. These include a mixture of structural sections (e.g., *Objectives* and *Findings*) and specific content elements, such as PICO components, treatments and their effects, assessed through graded evaluation of aspects such as factuality, completeness and conciseness.

Joseph et al. (2024) further required evaluators to justify their scores through free-text rationales. These can support the development of *explainable* LLM judges, capable of articulating the basis for their evaluations. As base LLMs were found to generate weak explanations, human-authored rationales provide valuable supervisory data to improve evaluative reasoning. These higher-quality critiques may also support iterative refinement of generated summaries (Saunders et al., 2022).

3. Methods

3.1. Evaluation Framework

Existing evaluation frameworks for medical summarisation provide useful foundations, but do not fully address the requirements of assessing professionally-oriented systematic review summaries. To bridge this gap, we have developed a structured evaluation framework aligned with aspects of the PRISMA-A and Cochrane guidelines. The design of the framework targets the assessment not only of factual accuracy, but also of pro-

fessional acceptability in compressed summaries intended for healthcare decision makers.

Each summary is evaluated across nine different dimensions reflecting key structural and evidential components of systematic review abstracts. For each dimension, evaluators assign a single ordinal score capturing factual adequacy and professional suitability. Certain dimensions with identified deficiencies additionally require the specification of the type of corrective action that would be needed to achieve content acceptability (i.e., *add*, *remove* or *modify*). This information enables a systematic analysis of common error types.

To ensure interpretability and traceability of errors (Kambhamettu et al., 2025; Xie et al., 2024), non-maximal scores require erroneous content in generated summaries to be linked to corresponding evidence in the source abstract. Specifically, problematic sentences in the summary are identified at the dimension level, while accompanying dimension-specific rationales explain the issues.

The nine evaluation dimensions are outlined below, with scoring details provided in Table 1. Appendix A provides an example of how the evaluation scheme is applied to a specific summary.

Dimensions 1–4: PICO elements. PRISMA-A guidelines require systematic review abstracts to specify study objectives in terms of PICO elements (see Table 2). Since Tang et al. (2023) identified missing PICO details as a source of harmfulness in LLM-generated medical summaries, we evaluate *Population*, *Intervention*, *Comparator* and *Outcome* elements as separate dimensions.

Dimensions 5–7: Methodology, Main Findings and Conclusions. These reflect PRISMA-A requirements that abstracts should report study methods, key findings and a concluding interpretation, including possible limitations.

Dimension 8: Quality of Evidence. The Cochrane handbook specifies that evidence syntheses should report a GRADE assessment of evidence quality/certainty (Balshem et al., 2011), i.e., *high*, *moderate*, *low* or *very low*. This dimension assesses whether the generated summary accurately reflects the quality of evidence reported in the original abstract.

Dimension 9: Currency. PRISMA-A requires abstracts to report when information sources were last searched. This dimension evaluates whether the generated summary correctly represents the search date(s) specified in the abstract.

3.2. Data Source

Our evaluation corpus, which we name *Psyc-SumEval*, was constructed using the abstracts of 115 randomly selected, publicly available Cochrane systematic reviews in the mental health domain. Reviews were sampled from the

Dimension	Scoring scheme and criteria
PICO elements (Population, Intervention, Comparator, Outcome)	0 = missing; 1 = partially correct or inaccurate; 2 = accurate and complete. Rationale required for 0/1.
Quality of Evidence	0 = inconsistent with abstract or only partially reported; 1 = consistent with abstract (both present or both absent). Rationale required for 0.
Main Findings	0 = missing; 1 = partial/inaccurate; 2 = accurate. Experts also specify corrective action: add, remove, modify or none. Rationale required for 0/1
Methodology	Same 0–2 scale and corrective action scheme as Main Findings.
Conclusions	Same 0–2 scale and corrective action scheme as Main Findings.
Currency	0 = date inaccurate or omitted; 1 = date accurate; “No date reported” = absent in both abstract and summary.

Table 1: Rubric used to evaluate LLM-generated summaries

PICO Element	Description	Examples
Population	Characteristics of trial participants (e.g, demographics, medical condition(s))	<i>Adult (19–44 years), Schizophrenia</i>
Intervention	Active treatment assessed for efficacy	<i>Family Therapy, Relapse Prevention</i>
Comparator	Alternative treatment against which the intervention is compared	<i>Usual care</i>
Outcome	Measures used to determine trial results	<i>Hospital admission, Quality of life</i>

Table 2: Descriptions and examples of PICO elements

Cochrane Library¹, to ensure diversity of interventions and populations. No patient-level or individual trial data were included.

3.3. Summary Generation

For each of the 115 abstracts, a prompt designed to encourage the generation of a compressed summary compliant with key PRISMA-A and Cochrane guidelines (see Figure 1) was issued to four different LLMs, i.e., *GPT-4o* (OpenAI, 2024a), *GPT-4o-mini* (OpenAI, 2024b), *Claude Opus* (Anthropic, 2024) and *LLaMA-3.1-70B-Instruct* (Meta, 2024). No content beyond the abstract itself was provided to the models. This yielded 460 summaries, each of which evaluated by one of the two psychiatrists using the framework described above.

3.4. Evaluation of Professional Judgement Stability

The primary evaluation phase was followed by a cross-review phase. The objective was not to establish inter-annotator reliability in the traditional sense, but rather to examine the stability of professional judgment under a shared, structured rubric. In high-stakes biomedical summarisation, variation between experts may reflect legitimate interpretive discretion rather than labelling error. The

Summarize the following research abstract using the PICO framework. Identify the Population (P) being studied, the Intervention (I), the Comparison (C), and Outcomes (O) measured. Highlight the study's methodology, main findings, and conclusions. Emphasize any key clinical outcomes, the quality of the evidence, and how up-to-date the evidence is. Ensure the summary captures the overall significance of the research in the context of patient care and treatment efficacy, as well as any potential limitations or recommendations for future research.

Figure 1: Prompt issued to LLMs

cross-review phase was therefore designed to permit a cross-dimensional analysis of convergence and divergence in professionally grounded perspectives on acceptable abstraction of evidence, rather than to produce a statistically optimised annotation dataset. As mentioned above, every non-maximal score required an explanatory rationale, linking evaluative outcomes to explicitly documented reasoning. Accordingly, disagreement was analysed at the level of justification, rather than being restricted to numerical divergence.

Although Cohen's κ is commonly used to quantify inter-annotator agreement beyond chance, it was not adopted here. κ assumes independent annotation over fully overlapping datasets with

¹<https://www.cochranelibrary.com/>

fixed categorical labels and treats disagreement as noise relative to a latent ground truth. However, we aimed to examine professional judgment stability rather than categorical reliability. Moreover, several dimensions involve graded or interpretive assessments (e.g., adequacy of methodological detail or appropriateness of conclusions), where variation may reflect professional discretion. Under these conditions, a κ coefficient could conflate meaningful interpretive differences with unreliability. We therefore report dimension-level disagreement patterns alongside analysis of qualitative rationales to characterise expert judgment stability.

The cross-review phase was undertaken using only a subset of the complete corpus (i.e., 30 randomly selected abstracts), in order to balance analytical depth with realistic constraints on expert time, prioritising dimension-level interpretability over large-scale redundancy. Each expert reviewed the four generated summaries for 15 abstracts originally annotated by the other expert (60 summaries total) and indicated agreement or disagreement for each dimension.

4. Results

4.1. Cross-Review Phase

Table 3 reports the number of disagreements between the psychiatrists identified in the cross-review phase. Patterns of disagreement provide insight into which aspects of summarisation are primarily governed by explicit guideline adherence and which aspects are more open to interpretive professional discretion. For most models and dimensions, disagreements occurred in a relatively small proportion of cases (generally $\leq 15\%$ of the 30 summaries reviewed). This provides evidence of substantial convergence between experts regarding professional acceptability across dimensions. A notable exception is the *Conclusions* dimension for *LLaMA-3.1-70B*, whose closing content was often repetitive and poorly structured, making consistent evaluation difficult. We provide a dimension-level analysis below.

PICO Elements — The *Objectives* sections of original abstracts often specify PICO elements at a broad, review design level, while the *Main Results* may provide more detailed PICO characteristics that are specific to the studies finally included in the review (e.g., age ranges of study populations or names of specific interventional drugs). The two experts sometimes disagreed over the necessity of including such study-specific details in compressed summaries. In the original abstracts, these additional details are most frequently provided for the *Population* and *Intervention* elements, but were most commonly omitted in the summaries

of *LLaMA-3.1-70B-Instruct* and *GPT-4o-mini*, thus explaining the concentration of disagreements in these two dimensions and models.

Methodology — Experts showed high agreement regarding core methodological information, particularly the inclusion of total numbers of studies and participants. Disagreements questioned the need to include more granular information, such as specific registers searched (e.g., *Cochrane Schizophrenia study-based register of trials*), countries covered, or the level of detail required to describe data collection and analysis, and risk-of-bias assessment.

Main Findings — Most disagreements concerned models' handling of low-certainty evidence. In particular, there was some difference of opinion on whether results specified as having *very low* certainty in the original abstract could be safely omitted in the compressed summary. Variation also emerged in terms of tolerance towards subtle shifts in certainty statements. For example, the original abstract statement *None of the trials reported clinically important change in quality of life* was rendered by one model as *Modafinil has no effect on quality of life*, which only one expert accepted. In contrast, more hedged formulations (e.g., *may have little or no effect*) were consistently judged to be acceptable.

Conclusions — Some model-generated summaries separated *Conclusions*, *Limitations* and *Recommendations for Future Research*. One expert assessed *only* the explicitly labelled *Conclusions* sections, whereas the other expert also considered content in *Limitations* and *Recommendations*, given that such information may occur within original abstract conclusions. Disagreements typically arose when the latter expert found potentially misleading or unsupported statements in these additional sections. If the original abstract contained little or no discussion of limitations or future research directions, then all models were susceptible to introducing new speculative content without explicit attribution. Notably, however, no disagreements concerned the content of explicitly labelled *Conclusions* sections of the generated summaries, suggesting that the experts had shared expectations regarding the core function of these sections.

Quality of Evidence & Currency — Disagreements for *Quality of Evidence* were confined to a single abstract involving an atypical case, in which no eligible studies were identified. The sole disagreement for *Currency* appears attributable to evaluator oversight, as the search date was found to be explicitly stated in the generated summary.

4.2. Per-Model Performance

Using the psychiatrist-assigned assessment scores, we compared model performance across

Category	GPT-4o	Claude Opus	LLaMA-3.1-70B	GPT-4o-mini
Population	2	3	4	4
Intervention	1	2	5	2
Comparator	0	0	1	1
Outcome	0	0	0	0
Methodology	2	2	2	3
Main Findings	4	3	5	4
Conclusions	5	5	13	6
Quality of Evidence	1	1	1	1
Currency	0	1	0	0

Table 3: Number of disagreements in the cross-checked subset ($n = 30$ reviews per model).

Dimension	GPT-4o	Claude Opus	LLaMA-3.1-70B	GPT-4o-mini
Population (0–2)	1.91	1.77	1.71	1.70
Intervention (0–2)	1.87	1.83	1.79	1.80
Comparator (0–2)	1.90	1.92	1.81	1.90
Outcome (0–2)	1.84	1.86	1.77	1.76
Methodology (0–2)	1.65	1.43	1.28	1.42
Main Findings (0–2)	1.74	1.77	1.37	1.70
Conclusions (0–2)	1.72	1.72	1.30	1.57
Quality of Evidence (1)	0.93	0.96	0.98	0.97
Currency (1)	0.41	0.39	0.97	0.77

Table 4: Per-model performance statistics. Means (0–2 scale) are provided for ordinal dimensions and proportion of “1” labels is shown for Quality of Evidence and Currency.

dimensions. For dimensions evaluated on a 0–2 scale, mean scores were calculated to provide an indication of overall model adequacy. For the dimensions with binary values (i.e., *Quality of Evidence* and *Currency*), we report the proportion of summaries assigned the value “1”. Although *Currency* allowed three labels (“0”, “1” and “No date reported”), only the first two labels were ever assigned. Results are presented in Table 4.

Table 5 reports the distribution of *corrective actions* needed to render the content of the *Methodology*, *Main Findings* and *Conclusions* sections acceptable. These actions are almost exclusively *Add* or *Remove*, showing that most content errors in generated summaries correspond to omissions or hallucinated/unsupported content, as discussed further below. The tiny proportion of *Modify* actions suggests that misinterpretations/distortions of original abstract content are rare.

4.2.1. PICO Dimensions

Performance on extracting PICO elements is generally high. Although expert rationales occasionally note completely missing items in generated summaries, the most common issue concerns mismatches between the required and provided levels of detail. Although a small number of disagreements regarding the required level of detail in PICO element summaries were identified in Section 4.1, the two experts generally agreed

that such summaries should include both *broader scope characteristics* defined in the original review objectives and more *specific characteristics* of the studies actually included in the review. Errors in PICO elements often correspond to cases in which one or other of these types of characteristics is missing from the generated summary.

When only a single study is included in a review, for example, models often extract specific PICO characteristics from that study alone, and omit broader scope review objectives. However, for reviews including multiple studies, it is more common for summaries to include these broader characteristics, but to omit additional study-specific details. This pattern of behaviour is most prevalent for *Population*, although the notably higher performance of *GPT-4o* for this dimension reflects its greater tendency to generate *Population* summaries that include not only general information from the *Objectives* section of the original abstract (e.g., conditions suffered), but also more specific details from the *Selection Criteria* and *Main Results* sections, such as age, ethnicity and condition severity.

More stable cross-model performance is observable for *Intervention*, with all models demonstrating an ability to extract relevant information from the *Objectives*, *Selection Criteria* and *Main Results* sections of the original abstract. This allows generated summaries to successfully supplement, e.g., classes of interventional drugs introduced in the

Dimension	Corrective Action	GPT-4o	Claude Opus	LLaMA-3.1-70B	GPT-4o-mini
Methodology	None	65.2%	42.6%	27.8%	41.7%
	Add	33.0%	55.7%	71.3%	58.3%
	Remove	0.0%	0.9%	0.0%	0.0%
	Modify	1.7%	0.9%	0.9%	0.0%
Main Findings	None	74.8%	76.5%	37.4%	70.4%
	Add	25.2%	21.7%	60.9%	29.6%
	Remove	0.0%	0.0%	0.0%	0.0%
	Modify	0.0%	1.7%	1.7%	0.0%
Conclusions	None	71.3%	73.0%	30.4%	57.4%
	Add	0.9%	0.9%	0.9%	0.9%
	Remove	27.8%	26.1%	68.7%	41.7%
	Modify	0.0%	0.0%	0.0%	0.0%

Table 5: Distribution of corrective actions for *Methodology*, *Main Findings* and *Conclusions*

objectives with the names of more specific drugs evaluated in the included studies. The slightly higher performance of *GPT-4o* for this dimension appears to be due to its stronger ability to extract specific treatments when their mentions are scattered throughout the *Main Results* section of the original abstract.

Performance for *Comparator* is high for all models. However, when multiple comparators are mentioned in the original abstract, *LLaMA-3.1-70B* appears to be slightly more prone than other models to omitting one or more of these comparators from the generated summary.

In certain original abstracts, some or all of the *Outcomes* are mentioned only in the *Main Results* sections. Because these sections are often information-dense, comprehensive outcome extraction can be more challenging than other elements. The larger *GPT-4o* and *Claude Opus* handle this complexity more consistently, which likely explains their slightly stronger performance for the *Outcome* dimension.

4.2.2. Methodology Dimension

This dimension exhibits the lowest performance scores across all models. Achieving acceptability is reliant on integrating selected information scattered across multiple sections of original abstracts (typically *Search Methods*, *Selection Criteria*, *Data Collection and Analysis* and *Main Results*). This makes appropriate content selection challenging.

Expert rationales indicate strong agreement on a core set of expected elements, i.e., whether the review updates a previous one; key resources searched; number of included studies; total participant numbers; and brief descriptions of data analysis, risk-of-bias assessment and evidence grading. Table 5 shows that *Methodology* errors almost exclusively correspond to omission of certain information that is considered necessary from a professional perspective. Usually, however, the deficiency stems from a single missing core item,

rather than more widespread inaccuracies.

Inspection of generated summaries reveals that all models are capable of extracting information relevant to multiple core elements, but they do so inconsistently. This may reflect a tendency to generate relatively fixed-length *Methodology* sections, in which the inclusion of less critical details displaces more essential ones. Consequently, fundamental information, such as the number of included studies, may be omitted. Compared to the stronger *GPT-4o*, this pattern is more evident for *Claude Opus*, *LLaMA-3.1-70B-Instruct* and *GPT-4o-mini*, whose methodological summaries frequently restate PICO information narratively and/or redundantly include the search date, even when it appears elsewhere in the summary.

4.2.3. Main Findings Dimension

Generating acceptable summaries of *Main Findings* may appear simpler than for *Methodology*, partly because this task primarily involves compressing a single section of the original abstract, i.e., *Main Results*, and also since the performance for all models is higher than for *Methodology*. Nevertheless, the task remains complex: original *Main Results* sections are often dense, given their integration of results, interpretations and certainty levels across multiple PICO permutations. Effective summarisation thus requires both accurate extraction and disciplined compression.

This complexity helps to explain why the smaller *LLaMA-3.1-70B-Instruct* struggles with this dimension. Table 5 shows that, similarly to *Methodology*, most identified issues concern omissions of necessary information. Contrary to concerns raised in the cross-review analysis, overt *certainty illusions* are rare. More frequent is the failure to explicitly associate reported results with their originally stated level of evidence quality/certainty.

By far the most common problem, however, is the omission of one or more outcomes mentioned in the original abstract. While one expert was

sometimes slightly more lenient, both experts were largely unanimous that *all* outcomes identified in the *Outcomes* dimension should be mentioned in the *Main Findings* of the compressed summary, regardless of evidence quality level and whether sufficient data was available to evaluate the outcome.

Table 5 indicates that high-quality information extraction and compression appears to be feasible for all models, with the exception of *LLaMA-3.1-70B-Instruct*. Different models use varying techniques to achieve the goal of generating *Main Findings* summaries that are shorter than the corresponding section in the original abstract. One method that appears to be used to varying degrees by all models is the implicit application of thresholding criteria to filter out some of the originally reported outcomes. Examples include secondary outcomes; outcomes qualified by phrases like *very low certainty*, *no clear evidence* or *interpreted with caution*; outcomes with no usable data; and cases in which the intervention(s) evaluated made little or no difference to specified outcome(s). However, given the experts' opinion that all outcomes from the original abstract should also be mentioned in the generated summaries, such filtered results lead to lower acceptability scores.

As shown in Table 5, *Claude Opus* and *GPT-4o* generate the highest proportions of *Main Findings* summaries that are considered to be acceptable without modification (76.5% and 74.8%, respectively). This likely reflects their stronger abstractive capabilities, enabling compact integration of multiple results with minimal loss of information. The slightly lower performance of the smaller *GPT-4o-mini* (70.4% of *Main Results* sections deemed acceptable) may reflect its tendency to adhere more closely to the more verbose source phrasing of results, meaning that a greater degree of outcome filtering may sometimes be required to produce *Main Findings* summaries of a suitable length. The far lower proportion of fully acceptable *Main Findings* generated by *LLaMA-3.1-70B-Instruct* appears to be due to its frequent generation of shorter (and hence, more extremely filtered) findings sections than other models, possibly due to the disproportionate length allocated to its concluding content.

4.2.4. *Conclusions* Dimension

Performance on the *Conclusions* dimension is lower than for *Main Findings* across all models. This may seem surprising, since *Conclusions* sections from the original abstracts should be more straightforward to compress than the more detailed *Main Results* sections; conclusions are uniformly shorter, and aim to clearly summarise the most significant outcomes of the review.

Indeed, the expert who evaluated only explicitly labelled *Conclusions* sections judged the vast ma-

majority of these generated sections to be acceptable without modification. The main exception was for *LLaMA-3.1-70B-Instruct*, whose conclusions were frequently marked down for repetition of information.

The lower scores for this dimension mainly stem from problems in the generated content of sections labelled *Limitations*, *Future Research* or similar, which only the second expert considered. The issues were mostly the same as those identified in the cross-review analysis, i.e., fabricated content within these additional sections, which was particularly likely to occur if the original abstract provided little or no information about limitations or future work. In such cases, models often generated statements that appeared plausible, but which were unsupported by the original abstract content. This pattern was especially pronounced in the outputs of *LLaMA-3.1-70B-Instruct* (and to a lesser extent, *GPT-4o-mini*), which sometimes framed invented content as directly attributable to the study (e.g., *The study highlights ...*). Such observations help to explain why the *Corrective Action* specified for the vast majority of erroneous *Conclusions* sections is to *remove* information (see Table 5).

However, given that the second expert identified few issues within explicitly labelled *Conclusions* sections, all models appear generally proficient at generating accurate summaries of primary study outcomes. Similarly to *Main Findings*, preserving nuanced expressions of certainty within the *Conclusions* sections remains important to avoid misinterpretation, and expert rationales indicate that fidelity to certainty levels is generally maintained.

4.2.5. *Quality of Evidence* Dimension

Results show that the vast majority of generated summaries preserve sufficient information from original abstracts regarding evidence quality. Adherence to Cochrane Handbook guidelines regarding GRADE evidence certainty levels usually results in uniform specification of these levels across original abstracts, thus making this information straightforward to extract. Examination of generated summaries reveals variations in exactly how and where information about quality of evidence is specified. This may, for example, be at the level of individual results in the *Main Findings* section; within a separate *Quality of Evidence* section that specifies the range of certainty levels across all results; or as a statement in the *Limitations* section that low or very low evidence quality limits the validity of the results. All of these reporting methods appear to be professionally acceptable, as long as the complete range of evidence levels observed across all results is specified. Only a small number of summaries fail to meet this criterion.

4.2.6. Currency Dimension

The results for this dimension are unexpected. In contrast to other dimensions, the smaller *LLaMA-3.1-70B-Instruct* and *GPT-4o-mini* models considerably outperform the larger *GPT-4o* and *Claude Opus*. Particularly noteworthy is *LLaMA-3.1-70B-Instruct* which, despite under-performing other models in all other dimensions, is able to correctly extract information about search currency from 97% of abstracts, compared to only around 40% of abstracts for *GPT-4o* and *Claude Opus*. Extraction of this information from original abstracts should theoretically be very straightforward, since search dates are usually clearly stated in the *Search Methods* section. Therefore, the failure of larger models to extract this information consistently is surprising. A feasible explanation could be the wording of the prompt issued to the LLMs, which explicitly elicits information about how up to date the search *evidence* is. It is therefore possible that other models are carrying out more complex processing to ensure that a search date is *only* included in the generated summary if it is clear that all available evidence up until the specified search date was considered. As an example, some original abstracts state that a search was carried out *without date restrictions*. However, further investigation is needed to determine whether different models' decisions to exclude search dates from summaries is based on prompt wording and/or contextual cues, or is simply due to space constraints.

5. Discussion

We have shown that our proposed multi-dimensional framework can be applied by clinical experts to evaluate LLM-generated summaries from a professional perspective. The cross-review results indicate substantial convergence in expert judgments, suggesting that shared professional norms guide evaluation across dimensions. The evaluations enabled systematic comparison of model performance. No single model outperformed the others across all dimensions; instead, performance was dimension-specific. Analysis of corrective actions revealed consistent patterns: deficiencies in *Methodology* and *Main Findings* largely involved omissions of essential details, whereas problematic *Conclusions* more often contained unsupported or fabricated content. The frequency of such issues varied by model.

Expert-authored rationales provided further insight into implicit professional criteria, including core methodological elements expected in summaries and the requirement that all pre-specified outcomes be acknowledged in *Main Findings*, even when evidence is inconclusive. These expectations are not always explicit in reporting guide-

lines but are central to professional acceptability. Capturing such implicit standards highlights the value of structured expert annotation for improving both evaluation and generation of automatically generated systematic review summaries.

Our findings also reveal that different dimensions pose distinct challenges. For *PICO* elements, the key issue is balancing design-level scope with study-specific detail. *Methodology* requires selective integration of key information that may be scattered across multiple abstract sections. *Main Findings* demand effective abstraction to preserve *all* outcomes within limited space, while *Conclusions* require strict fidelity to the source text to avoid unsupported extrapolation or hallucination. To meet professional expectations for *Quality of Evidence*, the complete range of evidence levels observed across reported results must be present in the generated summary. Finally, the unexpectedly poor results of the *Currency* dimension for most models demand further investigation to ensure that models' understanding is aligned with expert requirements for this dimension; the latter appear to be restricted to the need for a simple specification of the date(s) when resources relevant to the review were searched.

6. Conclusion

This paper has introduced a novel framework for evaluating the professional acceptability of LLM-generated summaries of systematic review abstracts in high-stakes clinical settings. The framework has been applied by psychiatrists to assess summaries generated by four different LLM models for 115 Cochrane abstracts to create the *PsycSumEval* corpus. Beyond benchmarking model performance, *PsycSumEval* makes expert evaluative reasoning explicit through structured scoring and documented rationales. In scenarios where summaries may influence clinical understanding and decision making, interpretability and transparency of professional judgement is essential for responsible evaluation.

7. Future Work

Future work will explore the effects of alternative prompting strategies on model performance. We also plan to extend our work beyond the mental health domain to systematic reviews in other clinical areas, improving coverage and generalisability. Involving additional clinical experts will allow us to examine the stability of professional acceptability criteria across domains. Furthermore, we will use *PsycSumEval* to fine-tune and evaluate LLM judges tailored to professionally oriented assessment of systematic review summaries.

8. Limitations

Several limitations should be noted. The evaluation involved only two psychiatrists, potentially restricting diversity of perspectives. The cross-review was conducted on a subset of 30 reviews (120 summaries), while the remaining summaries were evaluated by a single expert. Broader double evaluation and the involvement of additional experts would strengthen generalisability. The cross-review protocol was not fully blinded, as experts had access to prior labels and rationales. While this supported interpretability, it may have introduced mild anchoring effects. Future work should adopt and compare fully independent evaluations. The study assessed summaries of Cochrane abstracts, rather than full systematic reviews. Although abstracts provide a structured and controlled setting, they represent only part of the underlying evidence. Extending evaluation to full-text reviews may yield different performance patterns.

9. Ethics Statement

The PsycSumEval corpus is derived from publicly available Cochrane abstracts and contains no patient-level or personally identifiable data. However, it involves clinically sensitive content: summaries of medical evidence may influence decisions if misused. The resource is therefore released solely for methodological research and not for clinical application. Risks include reinforcement of model errors in high-stakes settings, particularly in interpretive dimensions such as conclusions. Continued human oversight remains essential. Although abstracts include review dates, medical evidence evolves and findings may become outdated. All evaluation protocols and guidelines are released transparently to support responsible and reproducible research.

10. Data Availability

The PsycSumEval corpus and evaluation guidelines are available at <https://github.com/paul-rich-t/PsycSumEval>.

11. Acknowledgements

This work was partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union under the NextGenerationEU Program.

12. Bibliographical References

Gerd Antes, David Tovey, and Nancy Owens. 2022. Cochrane: Trusted evidence. informed decisions. better health. *Systematic Reviews in Health Research: Meta-Analysis in Context*, pages 413–423.

Felipe Arias-Russi, Carolina Salazar-Lara, and Rubén Manrique. 2025a. [Bridging the gap in health literacy: Harnessing the power of large language models to generate plain language summaries from biomedical texts](#). In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 269–284, Albuquerque, New Mexico. Association for Computational Linguistics.

Felipe Arias-Russi, Carolina Salazar Lara, and Ruben Manrique. 2025b. [A multi-agent framework with diagnostic feedback for iterative plain language summary generation from cochrane medical abstracts](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 87–104, Suzhou, China. Association for Computational Linguistics.

Howard Balshem, Mark Helfand, Holger J. Schünemann, Andrew D. Oxman, Regina Kunz, Jan Brozek, Gunn E. Vist, Yngve Falck-Ytter, Joerg Meerpohl, Susan Norris, and Gordon H. Guyatt. 2011. [GRADE guidelines: 3. Rating the quality of evidence](#). *Journal of Clinical Epidemiology*, 64(4):401–406.

Lydie Bednarczyk, Daniel Reichenpfader, Christophe Gaudet-Blavignac, Amon Kenna Ete, Jamil Zagher, Yuanyuan Zheng, Adel Bensahla, Mina Bjelogrić, and Christian Lovis. 2025. Scientific evidence for clinical text summarization using large language models: scoping review. *Journal of Medical Internet Research*, 27:e68998.

Elaine M Beller, Paul P Glasziou, Douglas G Altman, Sally Hopewell, Hilda Bastian, Iain Chalmers, Peter C Gøtzsche, Toby Lasserson, David Tovey, and PRISMA for Abstracts Group. 2013. PRISMA for abstracts: reporting systematic reviews in journal and conference abstracts. *PLoS medicine*, 10(4):e1001419.

Elaine M Beller, Paul P Glasziou, Sally Hopewell, and Douglas G Altman. 2011. Reporting of effect direction and size in abstracts of systematic reviews. *Jama*, 306(18):1981–1982.

- BE Belsher, E Beech, D Evatt, DJ Smolenski, MT Shea, JL Otto, CS Rosen, and PP Schnurr. 2019. [Present-centered therapy \(pct\) for post-traumatic stress disorder \(ptsd\) in adults](#). *Cochrane Database of Systematic Reviews*, (11).
- Iain Chalmers and Douglas G Altman. 1995. *Systematic reviews*. BMJ Publishing London.
- Bohao Chu, Meijie Li, Sameh Frihat, Chengyu Gu, Georg Lodde, Elisabeth Livingstone, and Norbert Fuhr. 2025. [TracSum: A new benchmark for aspect-based summarization with sentence-level traceability in medical domain](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 844–864, Suzhou, China. Association for Computational Linguistics.
- Declan Devane, Johanna Pope, Paula Byrne, Evan Forde, Isabel O'Byrne, Steven Woloshin, Eileen Culloty, Darren Dahly, Ingeborg Hess Elgersma, Heather Munthe-Kaas, Conor Judge, Martin O'Donnell, Finn Krewer, Sandra Galvin, Nikita N. Burke, Theresa Tierney, KM Saif-Ur-Rahman, Tom Conway, and James Thomas. 2026. [Comparison of AI-assisted and human-generated plain language summaries for Cochrane reviews: a randomised non-inferiority trial \(HIET-1\) \[Registered Report - stage II\]](#). *Journal of Clinical Epidemiology*, 191:112102.
- Kay Dickersin, Roberta Scherer, and Carol Lefebvre. 1994. Identifying relevant studies for systematic reviews. *BMJ*, 309(6964):1286–1291.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. [Length-controlled AlpacaEval: A simple way to debias automatic evaluators](#). *arXiv preprint arXiv:2404.04475*.
- Zelalem Gero, Chandan Singh, Yiqing Xie, Sheng Zhang, Praveen Subramanian, Paul Vozila, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2024. [Attribute structuring improves LLM-based evaluation of clinical text summaries](#). *arXiv preprint arXiv:2403.01002*.
- Jasmin Helbach, Falk Hoffmann, Dawid Pieper, and Katharina Allers. 2023a. Reporting according to the preferred reporting items for systematic reviews and meta-analyses for abstracts (PRISMA-A) depends on abstract length. *Journal of Clinical Epidemiology*, 154:167–177.
- Jasmin Helbach, Kathrin Wandscher, Dawid Pieper, and Falk Hoffmann. 2023b. Cochrane review abstracts are getting longer, but this has no large impact on the reporting quality. *Journal of Evidence-Based Medicine*, 16(3):294–302.
- JPT Higgins, J Thomas, J Chandler, M Cumpston, T Li, MJ Page, and VA Welch. 2024. *Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024)*. Cochrane.
- Zhenyu Huang, Xianlai Chen, Yunbo Wang, Jincan Huang, and Xing Zhao. 2025. [A survey on biomedical automatic text summarization with large language models](#). *Information Processing & Management*, 62(5):104216.
- Sagar Jose, Khanh T.P Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. 2024. [Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models](#). *Expert Systems with Applications*, 255:124603.
- Sebastian Joseph, Lily Chen, Jan Trienes, Hannah Göke, Monika Coers, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. [FactPICO: Factuality evaluation for plain language summarization of medical evidence](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8437–8464, Bangkok, Thailand. Association for Computational Linguistics.
- Hita Kambhamettu, Jamie Flores, and Andrew Head. 2025. Traceable texts and their effects: A study of summary-source links in ai-generated summaries. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. [Prometheus: Inducing fine-grained evaluation capability in language models](#). In *The Twelfth International Conference on Learning Representations*.
- Mahira Kirmani, Gagandeep Kour, Mudasir Mohd, Nasrullah Sheikh, Dawood Ashraf Khan, Zahid Maqbool, Mohsin Altaf Wani, and Abid Hussain Wani. 2024. [Biomedical semantic text summarizer](#). *BMC Bioinformatics*, 25(1):152.
- Alon Lavie and Abhaya Agarwal. 2007. [ME-TEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.
- Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gøtzsche, John PA Ioannidis, Mike Clarke, Philip J Devereaux, Jos Kleijnen, and David Moher. 2009.

- The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*, 339.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2024. [Factual consistency evaluation of summarization in the era of large language models](#). *Expert Systems with Applications*, 254:124456.
- Katarina Maticic, Marina Krnic Martinic, and Livia Puljak. 2019. Assessment of reporting quality of abstracts of systematic reviews with meta-analysis using PRISMA-A and discordance in assessments between raters without prior experience. *BMC Medical Research Methodology*, 19(1):32.
- Sally Morton, Alfred Berg, Laura Levit, and Jill Eden. 2011. *Finding what works in health care: standards for systematic reviews*. National Academies Press.
- Ian G Needleman. 2002. A guide to systematic reviews. *Journal of clinical periodontology*, 29:6–9.
- Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021a. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.
- Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021b. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372.
- 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.
- Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet NLP: A survey. *arXiv preprint arXiv:2405.12819*.
- W Scott Richardson, Mark C Wilson, Jim Nishikawa, and Robert S Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12–13.
- William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.
- Liyan Tang, Zhaoyi Sun, Betina Ilnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *NPJ Digital Medicine*, 6(1):158.
- Amy Y Tsou and Jonathan R Treadwell. 2016. Quality and clarity in systematic review abstracts: an empirical study. *Research synthesis methods*, 7(4):447–458.
- Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.
- Yuriy Vasilev, Irina Raznitsyna, Anastasia Pamova, Tikhon Burtsev, Tatiana Bobrovskaya, Pavel Kosov, Anton Vladzomyrskyy, Olga Omelyanskaya, and Kirill Arzamasov. 2025. Evaluating medical text summaries using automatic evaluation metrics and LLM-as-a-judge approach: A pilot study. *Diagnostics*, 16(1):3.
- Yuriy Vasilev, Irina Raznitsyna, Anastasia Pamova, Tikhon Burtsev, Tatiana Bobrovskaya, Pavel Kosov, Anton Vladzomyrskyy, Olga Omelyanskaya, and Kirill Arzamasov. 2026. [Evaluating medical text summaries using automatic evaluation metrics and llm-as-a-judge approach: A pilot study](#). *Diagnostics*, 16(1).
- Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey for biomedical text summarization: From pre-trained to large language models. *arXiv preprint arXiv:2304.08763*.
- Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. [DocLens: Multi-aspect fine-grained evaluation for medical text generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 649–679, Bangkok, Thailand. Association for Computational Linguistics.

Fangyi Yu. 2025. When AIs judge AIs: The rise of agent-as-a-judge evaluation for LLMs. *arXiv preprint arXiv:2508.02994*.

Gongbo Zhang, Qiao Jin, Yiliang Zhou, Song Wang, Betina Ilnay, Yiming Luo, Elizabeth Park, Jordan G Nestor, Matthew E Spotnitz, Ali Soroush, et al. 2024. Closing the gap between open source and commercial large language models for medical evidence summarization. *NPJ Digital Medicine*, 7(1):239.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, pages 46595–46623, Red Hook, NY, USA. Curran Associates Inc.

shows the summary of this abstract, generated by *GPT-4o*, whose length is around half that of the original abstract (i.e., 422 words).

Table 6 shows the expert-provided scores and rationales (for dimensions whose content is not considered to be fully professionally acceptable) for the *Population*, *Intervention*, *Comparator*, *Outcome*, *Quality of Evidence* and *Currency* dimensions. Meanwhile, Table 7 shows the expert-provided scores, corrective actions and rationales (for dimensions whose content is not considered to be fully professionally acceptable) for the *Methodology*, *Main Findings* and *Conclusions* dimensions. For all dimensions in each table, a short discussion explaining the assigned score is provided, by making comparisons between the content of the original abstract and the generated summary. Note that these discussions were not provided by the experts as part of their evaluation effort, but rather have been added to aid understanding of the evaluation process.

The evaluation results shows that *GPT-4o* was largely successful in generating a significantly compressed summary of the original abstract, which largely conforms to the required standards of professional acceptability. Although the final sentence of the generated summary was found to constitute hallucinated content that is not supported by the original abstract, only a small number of required details were found to be missing from the remainder of the summary.

13. Language Resource References

Anthropic. 2024. *Claude Opus*.

Meta. 2024. *LLaMA 3.1 70B Instruct*.

OpenAI. 2024a. *GPT-4o*.

OpenAI. 2024b. *GPT-4o-mini*.

A. Worked Evaluation Example

This appendix provides a detailed, worked example of how a specific summary, generated by *GPT-4o*, was evaluated by one of the psychiatrist experts against the original Cochrane abstract (Belsher et al., 2019) (ID: *CD012898.PUB2*) according to the nine dimensions of our evaluation scheme.

Figure 2 provides the text of the original abstract, which is 865 words long. Meanwhile, Figure 3

Dimension	Score	Rationale	Discussion
Population	2	-	Relevant population information (i.e., <i>adults with PTSD</i>) is specified in <i>Objectives</i> section of the original abstract, and was correctly extracted. In the <i>Population</i> element of the generated summary, information from the <i>Background</i> section of the original abstract has also been used to expand <i>PTSD</i> to the full form <i>post-traumatic stress disorder</i> .
Intervention	2	-	Relevant information about the single intervention of interest (i.e., <i>PCT</i>) is correctly extracted from the <i>Objectives</i> section of original abstract, and is supplemented in the generated abstract with both the full form, i.e., <i>Present-centered therapy</i> and additional descriptive information, both of which are provided in the <i>Background</i> section of the original abstract.
Comparator	2	-	Basic information about comparators is correctly extracted from the <i>Objectives</i> section of the original abstract, and supplemented with information from a) the <i>Background</i> section, to expand the abbreviation <i>TF-CBT</i> to <i>trauma-focused cognitive-behavioral therapy</i> and b) the <i>Main Results</i> section, to identify the specific control conditions evaluated in the included studies, i.e., <i>wait-list</i> and <i>minimal attention</i> .
Outcome	1	missing <i>self-reported PTSD and depression symptoms</i>	Although the primary outcomes identified in the <i>Selection Criteria</i> section of the original abstract are correctly extracted, two further outcomes are mentioned in the <i>Main Results</i> section of the abstract, i.e., <i>self-reported PTSD</i> and <i>depression symptoms</i> . Although these outcomes are mentioned in the <i>Main Findings</i> section of the generated summary, they are missing from the <i>Outcomes</i> element, resulting in the assignment of the score 1.
Quality of Evidence	1	-	In the <i>Main Findings</i> section of the generated summary, evidence quality levels accompany all results for which such a level was specified in the <i>Main Results</i> section of the original abstract. Hence a score of 1 (i.e., <i>consistent with abstract</i>) was assigned.
Currency	1	-	Based on the information about the search date provided in the <i>Search Methods</i> section of the original abstract, the <i>Methodology</i> section of the generated summary correctly states that relevant studies up to February 2019 were screened. Accordingly, a value of 1 was assigned for this dimension.

Table 6: Evaluation scores, corrective actions, rationales (where appropriate) and associated discussions for the *Population*, *Intervention*, *Comparator*, *Outcome*, *Quality of Evidence* and *Currency* dimensions

Background

Present-centered therapy (PCT) is a non-trauma, manualized psychotherapy for adults with post-traumatic stress disorder (PTSD). PCT was originally designed as a treatment comparator in trials evaluating the effectiveness of trauma-focused cognitive-behavioral therapy (TF-CBT). Recent trials have indicated that PCT may be an effective treatment option for PTSD and that patients may drop out of PCT at lower rates relative to TF-CBT.

Objectives

To assess the effects of PCT for adults with PTSD. Specifically, we sought to determine whether (1) PCT is more effective in alleviating symptoms relative to control conditions, (2) PCT results in similar alleviation of symptoms compared to TF-CBT, based on an a priori minimally important differences on a semi-structured interview of PTSD symptoms, and (3) PCT is associated with lower treatment dropout as compared to TF-CBT.

Search methods

We searched the Cochrane Common Mental Disorders Controlled Trials Register, the Cochrane Library, Ovid MEDLINE, Embase, PsycINFO, PubMed, and PTSDpubs (previously called the Published International Literature on Traumatic Stress (PILOTS) database) (all years to 15 February 2019 search). We also searched the World Health Organization (WHO) trials portal (ICTRP) and ClinicalTrials.gov to identify unpublished and ongoing trials. Reference lists of included studies and relevant systematic reviews were checked. Grey literature searches were also conducted to identify dissertations and theses, clinical guidelines, and regulatory agency reports.

Selection criteria

We selected all randomized clinical trials (RCTs) that recruited adults diagnosed with PTSD to evaluate PCT compared to TF-CBT or a control condition. Both individual and group PCT modalities were included. The primary outcomes of interest included reduced PTSD severity as determined by a clinician-administered measure and treatment dropout rates.

Data collection and analysis

We complied with the Cochrane recommended standards for data screening and collection. Two review authors independently screened articles for inclusion and extracted relevant data from eligible studies, including the assessment of trial quality. Random-effects meta-analyses, subgroup analyses, and sensitivity analyses were conducted using mean differences (MD) and standardized mean differences (SMD) for continuous data or risk ratios (RR) and risk differences (RD) for dichotomous data. To conclude that PCT resulted in similar reductions in PTSD symptoms relative to TF-CBT, we required a MD of less than 10 points (to include the 95% confidence interval) on the Clinician-Administered PTSD Scale (CAPS). Five members of the review team convened to rate the quality of evidence across the primary outcomes. Any disagreements were resolved through discussion. Review authors who were investigators on any of the included trials were not involved in the qualitative or quantitative syntheses.

Main results

We included 12 studies ($n = 1837$), of which, three compared PCT to a wait-list/minimal attention (WL/MA) group and 11 compared PCT to TF-CBT. PCT was more effective than WL/MA in reducing PTSD symptom severity (SMD -0.84, 95% CI -1.10 to -0.59; participants = 290; studies = 3; $I^2 = 0\%$). We assessed the quality of this evidence as moderate. The results of the non-inferiority analysis comparing PCT to TF-CBT did not support PCT non-inferiority, with the 95% confidence interval surpassing the clinically meaningful cut-off (MD 6.83, 95% CI 1.90 to 11.76; 6 studies, $n = 607$; $I^2 = 42\%$). We assessed this quality of evidence as low. CAPS differences between PCT and TF-CBT attenuated at 6-month (MD 1.59, 95% CI -0.46 to 3.63; participants = 906; studies = 6; $I^2 = 0\%$) and 12-month (MD 1.22, 95% CI -2.17 to 4.61; participants = 485; studies = 3; $I^2 = 0\%$) follow-up periods. To confirm the direction of the treatment effect using all eligible trials, we also evaluated PTSD SMD differences. These results were consistent with the primary MD outcomes, with meaningful effect size differences between PCT and TF-CBT at post-treatment (SMD 0.32, 95% CI 0.08 to 0.56; participants = 1129; studies = 9), but smaller effect size differences at six months (SMD 0.17, 95% CI 0.05 to 0.29; participants = 1339; studies = 9) and 12 months (SMD 0.17, 95% CI 0.03 to 0.31; participants = 728; studies = 5). PCT had approximately 14% lower treatment dropout rates compared to TF-CBT (RD -0.14, 95% CI -0.18 to -0.10; participants = 1542; studies = 10). We assessed the quality of this evidence as moderate. There was no evidence of meaningful differences on self-reported PTSD (MD 4.50, 95% CI 3.09 to 5.90; participants = 983; studies = 7) or depression symptoms (MD 1.78, 95% CI -0.23 to 3.78; participants = 705; studies = 5) post-treatment.

Authors' conclusions

Moderate-quality evidence indicates that PCT is more effective in reducing PTSD severity compared to control conditions. Low quality of evidence did not support PCT as a non-inferior treatment compared to TF-CBT on clinician-rated post-treatment PTSD severity. The treatment effect differences between PCT and TF-CBT may attenuate over time. PCT participants drop out of treatment at lower rates relative to TF-CBT participants. Of note, all of the included studies were primarily designed to test the effectiveness of TF-CBT which may bias results away from PCT non-inferiority. The current systematic review provides the most rigorous evaluation to date to determine whether PCT is comparably as effective as TF-CBT. Findings are generally consistent with current clinical practice guidelines that suggest that PCT may be offered as a treatment for PTSD when TF-CBT is not available.

Figure 2: Original abstract from [Belsher et al. \(2019\)](#) (ID: CD012898.PUB2)

PICO Framework Summary:

Population (P): The research focused on adults diagnosed with post-traumatic stress disorder (PTSD).

Intervention (I): The primary intervention in the study was Present-centered therapy (PCT), a non-trauma, manualized psychotherapy technique designed for PTSD patients.

Comparison (C): PCT was compared to two conditions: trauma-focused cognitive-behavioral therapy (TF-CBT) and control conditions, which included wait-list or minimal attention controls.

Outcomes (O): The main outcomes measured were the severity of PTSD symptoms (based on clinician-administered measures) and treatment dropout rates.

Methodology:

The study was a systematic review and meta-analysis of randomized clinical trials (RCTs), following Cochrane standards. Researchers screened multiple databases for relevant studies up to February 2019, including grey literature and trial portals. They selected RCTs that assessed PCT versus TF-CBT or control conditions for adults with PTSD. Data analyses included random-effects meta-analyses, subgroup analyses, and sensitivity analyses.

Main Findings

- PCT showed superior effectiveness in reducing PTSD symptoms compared to wait-list/minimal attention controls with moderate-quality evidence.
- PCT was not found to be non-inferior to TF-CBT in terms of reducing PTSD symptom severity, based on a clinician-administered measure, with low-quality evidence. However, differences between PCT and TF-CBT reduced over time at 6 and 12-month follow-ups.
- PCT had a lower treatment dropout rate (approximately 14% lower) than TF-CBT, supported by moderate-quality evidence.
- No significant differences were observed in self-reported PTSD or depression symptoms post-treatment.

Conclusions

The research concluded that PCT is an effective treatment for reducing PTSD severity compared to control conditions, although it does not match TF-CBT's effectiveness according to clinician ratings. PCT's lower dropout rates suggest it may be a preferable option for patients likely to disengage from therapy. The study highlights the limitations due to the included studies being primarily designed to test TF-CBT, potentially biasing against PCT. Overall, it aligns with clinical guidelines suggesting PCT as an alternative when TF-CBT is unavailable.

Significance in Patient Care:

The research underscores PCT as an effective treatment alternative for PTSD, especially when access to TF-CBT is limited. Its lower dropout rates are clinically significant for maintaining patient engagement and adherence to therapy. However, the evidence supporting its non-inferiority to TF-CBT is of low quality, suggesting a need for further research focusing specifically on PCT's standalone efficacy.

Limitations and Recommendations for Future Research

The study acknowledges potential bias due to the primary design focus on TF-CBT effectiveness in the included trials. Future research should prioritize directly evaluating PCT's efficacy against other evidence-based treatments with distinct emphasis on PCT itself, employing high-quality RCTs to solidify its position in PTSD treatment protocols.

Figure 3: Summary of abstract from [Belsher et al. \(2019\)](#) (ID: CD012898.PUB2) generated by GPT-4o

Dimension	Score	Action	Rationale	Discussion
Methodology	1	Add info	Missing <i>number of studies</i>	Relevant methodological information is contained within multiple original abstract sections (i.e., <i>Search Methods</i> provides details of the types of resources searched; <i>Selection Criteria</i> specifies studies of interest were randomised clinical trials(RCTs); <i>Data Collection and Analysis</i> contains details about the standards used for data screening and collection, along with the data analysis methods employed; and <i>Main Results</i> contains information about the number of included studies). While most of this information was extracted and appropriately summarised, the score of 1 was assigned according to the summary's omission of a single piece of information considered essential for professional acceptability (i.e., the number of included studies).
Main Findings	2	None	-	All results reported in the <i>Main Results</i> section of the original abstract correctly extracted and appropriately summarised in the generated text, covering the comparison of the intervention of interest (i.e., <i>PCT</i>) with the target comparators (i.e., <i>TF-CBT</i> and the control conditions <i>wait-list</i> and <i>minimal attention</i>) in the context of all of the outcomes of interest (i.e., <i>PTSD severity</i> (determined by a clinician-administered measure), <i>treatment dropout rates</i> , <i>self-reported PTSD</i> and <i>depression symptoms</i>).
Conclusion	1	Remove info	<i>The summary should remove "Future research should prioritize directly evaluating PCT's efficacy against other evidence-based treatments with distinct emphasis on PCT itself, employing high-quality RCTs to solidify its position in PTSD treatment protocols."</i>	In the generated summary, concluding information is split between the <i>Conclusions</i> , <i>Significance in Patient Care</i> and <i>Limitations and Recommendations for Future Research</i> sections. Among these generated sections, the content of the <i>Conclusions</i> section is fully consistent with the information provided in the <i>Authors' conclusion</i> section of the original abstract. Appropriate content regarding implications for patient care and study limitations from the <i>Authors' conclusion</i> section of the original abstract is also used to generate accurate statements in the <i>Significance in Patient Care</i> and <i>Limitations and Recommendations for Future Research</i> sections of the summary. However, the <i>Authors' conclusion</i> section of the original abstract does not mention any potential future work. Accordingly, the final sentence in the <i>Limitations and Recommendations for Future Research</i> section of the generated summary, which recommends future work, is unsupported by the original abstract. Thus, a score of 1 is assigned.

Table 7: Evaluation scores, corrective actions, rationales (where appropriate) and associated discussions for the *Methodology*, *Main Findings* and *Conclusions* dimensions