Supporting Plain Language Summarization of Psychological Meta-Analyses with Large Language Models

Yarik Menchaca Resendiz[1,2], **Martin Kerwer**[1], **Anita Chasiotis**[1], **Marlene Bodemer**[1], **Kai Sassenberg**[1] and **Roman Klinger**[2]
[1]Leibniz-Institut für Psychologie (ZPID), Trier, Germany
[2]Fundamentals of Natural Language Processing, University of Bamberg, Germany
{ymr,mk,ac,mabo,ksa}@leibniz-pyschology.org, roman.klinger@uni-bamberg.de

Abstract

Communicating complex scientific findings to non-experts remains a major challenge in fields like psychology, where research is often presented in highly technical language. One effective way to improve accessibility, for non-experts, is through plain language summaries, which summarize key insights into simple and understandable terms. However, the limited number of institutions that produce lay summaries typically relies on psychology experts to create them manually – an approach that ensures high quality but requires significant expertise, time, and effort. In this paper, we introduce the KLARpsy App, a system designed to support psychology experts in creating plain language summaries of psychological meta-analyses using Large Language Models (LLM). Our system generates initial draft summaries based on a 37-criterion guideline developed to ensure clarity for non-experts. All summaries produced through the system are manually validated and edited by KLARpsy authors to ensure factual correctness and readability. We demonstrate how the system integrates LLM-generated content into an expert-in-the-loop workflow. The automatic evaluation showed a mean semantic-similarity score of 0.73 against expert-written summaries, and human evaluation on a 5-point Likert scale averaged above 3 (higher is better), indicate that the generated drafts are of high quality. The application and code are open source.

1 Introduction

Plain language summaries play an important role in making scientific findings accessible to broader audiences. In psychology and other disciplines, research findings are often communicated in technical language that can be difficult for non-experts to understand. This becomes especially challenging in the context of meta-analyses, where findings from multiple studies are summarized using specialized terminology and statistical information.

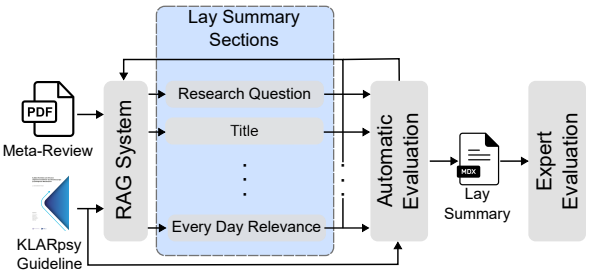

Figure 1: Workflow overview of the KLARpsy App, which follows the KLARpsy guidelines for summarizing psychological texts.

In recent years, there has been a growing initiative both in policy and practice – to promote more accessible scientific communication. For example, plain language summaries are now required for clinical trials in the European Union (under Regulation EU No. 536/2014; European Commission, 2023; Center for Drug Evaluation and Research et al., 2017; European Medicines Agency, 2022). These efforts aim to help non-experts understand and engage with research evidence.

Despite these developments, creating plain language summaries remains a time-consuming and expert-driven task. Summaries need to be both clear and accurate – especially in fields like psychology, where findings often inform health decisions or clinical practice. At the same time, new tools based on large language models offer promising support for generating initial drafts. These models are capable of producing fluent text (Shaib et al., 2023; Turbitt et al., 2023), but their reliability, particularly when it comes to factual accuracy, remains limited (Tomlin et al., 2024).

In this paper, we present KLARpsy App (Figure 1), a system developed to support psychology experts in creating plain language summaries of psychological meta-analyses[1]. Rather than re-

[1]We refer to scientific publications that include at least one meta-analysis as "meta-analyses".

Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations, pages 25–35

placing human expertise, the system is designed to work alongside it. It generates initial drafts based on a structured 37-point guideline to produce summaries for lay readers (Chasiotis et al., 2023). These drafts are then reviewed, edited, and finalized by domain experts, as ensuring factual accuracy and accessible language is essential. Even in full expert workflows, lay summaries are often checked by other experts or by laypersons.

Our approach highlights the potential of human-AI collaboration in scientific communication. By combining the efficiency of LLMs with expert oversight, KLARpsy App aims to reduce the load of producing summaries while maintaining high standards of quality. We also provide an evaluation, both automatic and human, of the system's outputs to better understand its strengths and limitations in practice. The code and an installable executable are available at https://github.com/leibniz-psychology/klarpsy-summarization-assistant.

2 Related Work

Automatic plain-text generation has been a long-standing area of research in natural language processing. Early work in this space focused on rule-based and statistical methods. Classical approaches include template-based generation and extractive summarization techniques that rely on word-frequency and cue-phrase detection (Tas and Kiyani, 2017; Reiter and Dale, 1997; El-Kassas et al., 2021; Kloehn et al., 2018).

In recent years, the field has increasingly shifted toward the use of large language models – such as GPT-Family (OpenAI, 2022, 2023), LLaMA-Family (Touvron et al., 2023; Grattafiori et al., 2024), and Mixtral (Jiang et al., 2024) – due to their ability to generate fluent, coherent text. This transition has significantly influenced the development of plain language summaries, especially in the scientific and biomedical domains. Recent work has evaluated LLMs in generating lay summaries of complex biomedical content, often highlighting their strengths in readability and fluency (Shaib et al., 2023; Turbitt et al., 2023; Tailor et al., 2024).

However, a limitation in LLM-based applications – not only summarization – is the risk of hallucination (i.e., the generation of inaccurate or fabricated information), which is particularly problematic in medical and scientific contexts where factual accuracy is critical (Tomlin et al., 2024). For instance, Fang et al. (2024) introduced the FAREBIO benchmark to evaluate factual consistency in biomedical lay summaries, showing that high fluency often correlates with diminished factual reliability.

To address this limitation, recent studies have explored human-in-the-loop frameworks where LLMs serve as assistive tools rather than standalone applications (Ovelman et al., 2024; Salazar-Lara et al., 2024; Tomlin et al., 2024; Chamberlain James, 2024). For example, Shyr et al. (2024) used engineered prompts with ChatGPT-4 to generate layperson-level summaries of clinical research abstracts. They validated performance using participant feedback from the ResearchMatch platform. Srivastava et al. (2024) introduced PIECE, a system that improves mental health dialogue summarization by first selecting important parts of the conversation using counseling knowledge. It then guides a language model with a structured plan to generate clearer and more accurate summaries. Showing the importance of humans in the loop, which is especially relevant in health care systems.

Our work builds on previous research by guiding LLM-based generation using a 37-point psychological guideline to produce lay summaries (Chasiotis et al., 2023). Rather than aiming to fully automate the process, our application is designed to function as a tool within a human-in-the-loop framework. The generated summaries are reviewed and validated by psychological experts to ensure accuracy and clarity.

3 KLARpsy App

This section introduces KLARpsy App, a system designed to assist psychology experts in generating plain language summaries of psychological meta-analyses. The tool uses a retrieval-augmented generation (RAG) system to produce initial summary drafts (in an MDX format, similar to a Markdown file), which are then reviewed and refined by domain experts and published[2]. The draft is guided by a 37-point criterion developed specifically for generating lay psychological summaries. Rather than replacing expert judgment, KLARpsy App supports a human-in-the-loop workflow that ensures both clarity and factual correctness. Section 3.1 provides a detailed overview of the system's architecture and workflow, while Section 3.2 outlines the psychological guideline that guides the system.

[2] https://klarpsy.de/klarpsytexte/

3.1 System Description

KLARpsy App is implemented as a retrieval-augmented generation system (Figure 1) using a client/server architecture. The client provides an easy-to-use interface for non-computer-science experts (e.g., psychologists), while the server performs the computationally intensive model inference. In this paper, we use OpenAI's GPT family as the backend because it eliminates the need for local GPU resources, enabling distribution as a standalone executable (e.g., a Windows .exe or a macOS .dmg) for non-technical users. The system can be easily updated to run local language models.

The application accepts as input one or more PDF files (meta-analyses), and optionally, a configuration file that specifies parameters such as the underlying model, API credentials, section lengths, and custom prompts.

The system is designed to generate 16 distinct sections that together form a plain language summary of a psychological meta-analysis. These sections include, for example, the Title, Background, and Research Question, and are structured according to the criteria defined in the KLARpsy guideline. Each section is generated independently, based on relevant content extracted from the meta-analysis. However, to maintain coherence and consistency across the full summary, the system models interdependencies between sections. For instance, the generation of the Title and Main Message depends on the content of the Research Question section. To handle these dependencies, KLARpsy App employs a chain-of-thought prompting strategy, where the content from earlier sections is explicitly passed as input into the generation of subsequent sections. For example, when generating the Title, the system first extracts the original scientific title from the meta-analysis and then paraphrases it to align with the identified research question.

The KLARpsy guideline defines two types of sections in the lay summary, each requiring a different generation strategy: (1) Free-text generation sections (e.g., Background, Interpretation of Results) are generated using zero-shot and few-shot prompting, guided by open-ended natural language instructions that incorporate the KLARpsy criteria (e.g., "Explain this in simple terms for a general audience."). (2) Information extraction and template filling sections follow fixed structures (e.g., Study Objective, Participant Information) that are generated using structured extraction and output generation techniques (Willard and Louf, 2023; OpenAI, 2024). Prompts for these sections include explicit output schemas to ensure the model adheres to predefined formats such as bullet points, labeled fields, or fixed templates (e.g., "The researchers searched for studies [topic_of_the_meta-analysis]. [Description_of_selection criteria]").

To further ensure the quality of the generated content, each section undergoes an automatic evaluation and optimization loop, which includes Structural checks (e.g., section length, formatting completeness) and readability assessments using established metrics such as Flesch Reading Ease (Flesch, 1948), Gunning Fog Index (Gunning, 1952), and SMOG Index (Mc Laughlin, 1969), to verify that the language is accessible to non-experts.

3.2 KLARpsy Guideline

The KLARpsy guideline (Chasiotis et al., 2023) provides a structured framework for writing lay summaries of psychological meta-analyses. Developed at the Leibniz Institute of Psychology (ZPID)[3], it aims to make complex research findings transparent and accessible to a non-specialist audience while maintaining neutrality and scientific accuracy.

Each KLARpsy text consists of 16 standardized content sections: Title, Key message, Authors and Affiliations, Background, Research Question, Study Selection, Study Selection Criteria, Study Approach, Study Variables, Key Results, Result Interpretation, Result Bias, Results Reliability, Every Day Relevance, and Funding and Conflicts of Interest (See Appendix A for an example). The guideline includes 37 criteria, organized into six categories: (1) *General content* ensures alignment with the original meta-analysis; (2) *Contextual attributes* address target audience and publication process; (3) *Linguistic attributes* cover tone, choice of words (e.g., handling of jargon and technical terms); (4) *Formal attributes* relate to structure, standardized formulation, and word limits, etc.; (5) *Presentation of results* guides the reporting of statistical findings; and (6) *Presentation of evidence quality* covers reliability of the evidence, such as reporting ratings or disclosing authors' conflicts of interest.

[3] https://leibniz-psychology.org/

4 Evaluation

We evaluate the initial drafts produced by KLARpsy App, which are intended to be refined through human-in-the-loop editing, using both automatic (Section 4.1) and human (Section 4.2) evaluations. In addition, in Section 4.3, we compare KLARpsy App against FlexRAG (Zhuocheng et al., 2025), an out-of-the-box RAG system.

4.1 Automatic Evaluation

The automatic evaluation consists of two parts: (1) assessing how well the KLARpsy App replicates expert-written summaries; and (2) evaluating the application's ability to generate lay summaries.

For the first part, we use 99 expert-written lay summaries as gold standards.[4] We then compare the model-generated summaries at the section level using two methods: BLEU scores and semantic similarity (we prompt GPT-4 to rate the similarity between corresponding sentences on a 5-level scale). BLEU provides a standard measure of word-level overlap, while the semantic-similarity score captures the overall idea even when phrasing differs. Table 1 reports the performance of the KLARpsy App replicating KLARpsy texts (expert-written). As expected, BLEU scores are generally low, since the same information can be phrased in many different ways. The exception is the *Authors* section, where BLEU scores are higher because author names must be directly extracted from the paper, leaving little room for paraphrasing. Semantic similarity (S. sim.), however, provides a more informative signal, showing stronger alignment between generated and human-written text. Sections involving information extraction (e.g., Authors, Conflict of Interest, Study Selection) achieve the highest similarity scores. In contrast, sections requiring more open-ended generation have a slightly lower performance (e.g., Research question, Background). Notably, the Results Reason section shows one of the lowest scores (0.56), likely reflecting the tendency of LLMs to generate assertive rather than uncertain statements – a limitation we discuss further in Section 4.2.

We evaluate readability using three standard indices: (1) Flesch Reading Ease (FRE), which ranges from 0–100 with higher values indicating easier text, (2) Gunning Fog Index, and (3) SMOG Index, both of which estimate the years of education required for comprehension (lower scores indicate simpler text). Table 2 reports scores for generated text, with expert-written references shown in parentheses. On average, KLARpsy App achieves a Gunning Fog score of 15.1 and a SMOG score of 13.8 – similar to expert-written texts – corresponding to a high school to early college reading level. This aligns with FRE values and suggests that readers do not need specialized training or professional expertise to understand the generated summaries.

Section	KLARpsy		FlexRAG	
	BLEU	S. sim.	BLEU	S. sim.
Title	0.02	0.70	0.01	0.13
Key message	0.03	0.62	0.01	0.10
First author	0.22	0.67	0.04	0.33
Affiliation	0.06	0.49	0.03	0.14
Authors	0.68	0.80	0.00	0.08
Background	0.02	0.62	0.01	0.13
Research Question	0.09	0.75	0.01	0.13
Selection criteria	0.04	0.65	0.00	0.18
Study selection	0.10	0.63	0.01	0.17
Study approach	0.04	0.62	0.01	0.15
Study variables	0.02	0.63	0.00	0.11
Key results	0.03	0.66	0.00	0.12
Results reason	0.07	0.53	0.00	0.14
Results bais	0.33	0.85	0.08	0.80
R. reliability	0.02	0.45	0.01	0.17
Everyday relevance	0.02	0.56	0.01	0.12
Funding	0.11	0.58	0.00	0.05
Conflict of interest	0.45	0.80	0.01	0.13
Average	0.15	0.73	0.01	0.20

Table 1: Evaluation of KLARpsy App and FlexRAG against 99 human-written texts across the KLARpsy sections.

4.2 Human Evaluation

For the human evaluation, three psychological experts evaluate the 10 generated lay summaries using a five-point Likert scale (Not at all, Slightly, Somewhat, Moderately, Extremely) across six statements: (1) the information in this slot correctly reflects the content/title; (2) the text avoids irrelevant information; (3) the language is understandable for laypeople; (4) the uncertainty in existing scientific findings is adequately reflected in the choice of words (or the language too assertive); (5) the text is neutral (i.e., without judgemental or advice); and (6) I would not recognize that the text has been written by an AI. These aspects were chosen based on the requirements of the KLARpsy guideline. Lay summaries are expected to use a neutral and non-directive tone to ensure that they are not mis-

[4]The 99 lay summaries were written by an expert and reviewed by another expert and a layperson to guarantee factual accuracy and accessible language, following the KLARpsy guideline. The summaries are available at https://klarpsy.de/klarpsytexte/)

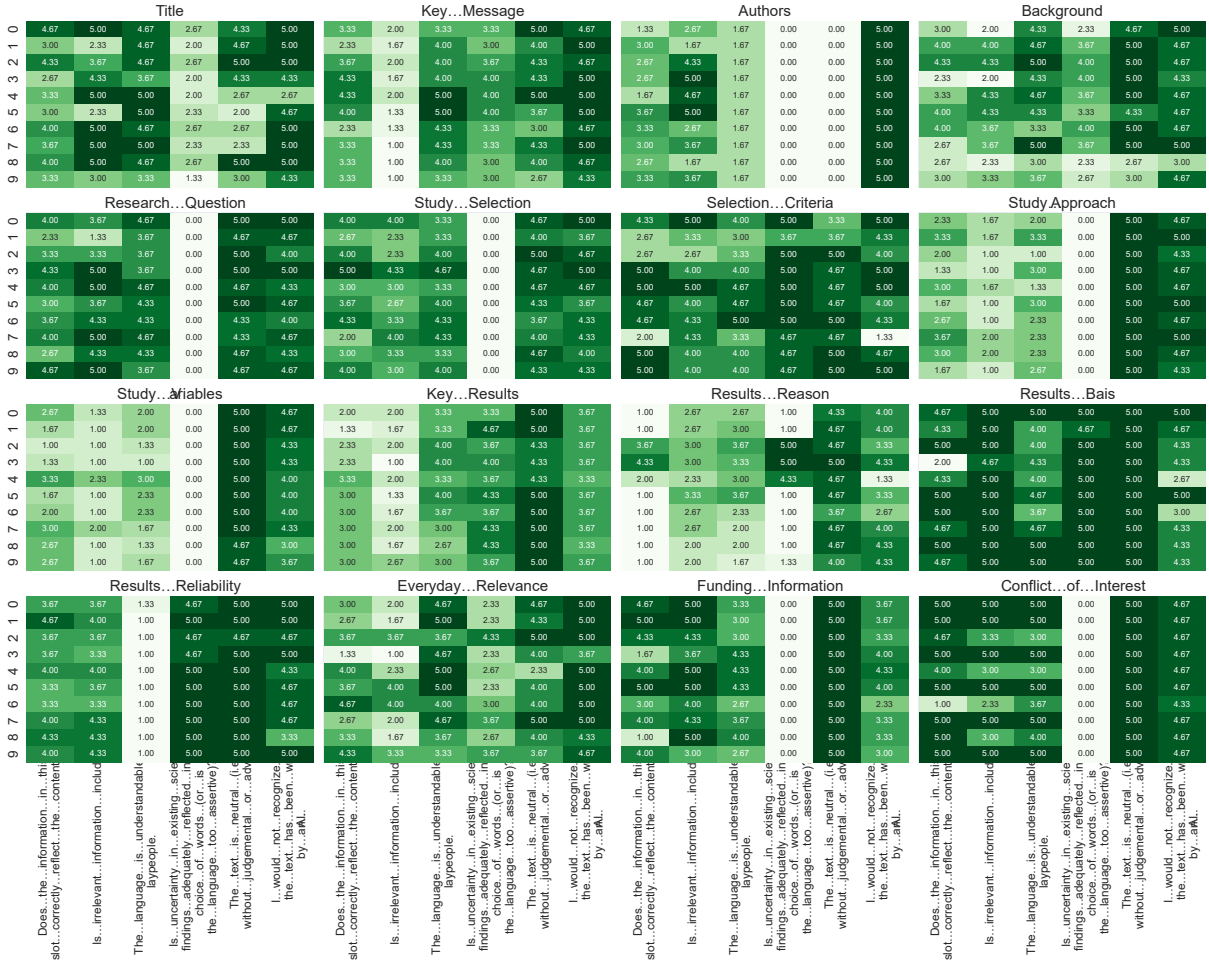

Figure 2: Section-level human evaluation results. Each heatmap shows the distribution of scores (1–5; higher/darker is better) across six evaluation criteria (x-axis) for the 10 papers (y-axis). Columns with a value of zero indicate that the criterion was not applicable (e.g., uncertainty does not apply to author names).

understood as guidelines or directions, but instead as evidence-based information. Table 3 reports the inter-annotator agreement among the three experts, with an average across the 16 sections of 0.60, indicating a high level of agreement.

Figure 2 presents heatmaps of section-level scores across the six criteria. Information extraction sections (e.g., Selection Criteria, Authors, and Funding Information) scored consistently high–predominantly dark green cells. In contrast, sections involving the presentation and interpretation of results (e.g., Study Variables, Research Question, and Results Reason) showed lower scores – primarily due to Criterion 4 (Is uncertainty in existing findings adequately reflected . . . ?) and Criterion 2 (Is irrelevant information included?). These results show that while LLMs handle information extraction reliably, they tend to produce verbose formulations and overstated claims about findings (e.g., "Treatment X helped patients" instead of "Treatment X benefited N out of 100 patients").

Table 4 shows the average score for each evaluation criterion across the 16 sections. Most criteria received scores above 3, indicating that the generated texts were generally of acceptable quality. The exception is Criterion 4, which averaged 2.05, reinforcing the automatic evaluation results reported in Section 4.1 and reflected in Figure 2.

4.3 Baseline Comparison

We compare KLARpsy App against FLEXRAG (Zhuocheng et al., 2025), an off-the-shelf RAG system, using the same set of 99 meta-reviews described in Section 4.1. To ensure a fair comparison, both systems rely on the same underlying model, GPT-4.0. For FLEXRAG, each summary was generated by providing only the meta-analysis and the KLARpsy guideline as prompts. Table 1 reports the BLEU scores and sentence similarity scores (against the 99

Section	FRE	Gunning	SMOG
Title	50.0 (28.5)	12.4 (15.8)	11.5 (12.9)
Key message	41.5 (25.2)	13.8 (16.6)	13.2 (14.7)
Background	37.5 (31.6)	14.9 (15.1)	13.4 (13.9)
Research Question	38.0 (38.6)	16.8 (16.2)	14.8 (14.5)
Selection criteria	44.5 (29.0)	14.3 (17.1)	13.4 (15.1)
Study selection	42.5 (54.7)	14.6 (12.3)	13.5 (11.8)
Study approach	38.1 (25.4)	15.5 (18.7)	14.7 (16.3)
Study variables	17.4 (15.1)	24.7 (28.4)	19.5 (23.3)
Key results	30.4 (31.3)	16.8 (15.0)	15.5 (13.5)
Results reason	29.2 (34.9)	17.1 (15.2)	15.3 (14.3)
Results bias	46.4 (47.5)	11.6 (11.0)	10.9 (10.6)
R. reliability	44.4 (33.1)	13.6 (14.7)	13.2 (13.7)
Everyday relevance	33.1 (31.3)	15.2 (16.2)	14.3 (14.8)
Funding	47.4 (36.0)	13.7 (17.2)	12.8 (15.1)
Conflict of interest	53.7 (60.7)	11.5 (9.8)	11.4 (10.1)
Average	39.6 (35.2)	15.1 (15.5)	13.8 (14.2)

Table 2: Evaluation of KLARpsy App generated text on readability using the Flesch Reading Ease (FRE), Gunning Fog (Gunning) Index, and SMOG Index across the KLARpsy sections. Readability scores for expert-written texts are shown in parentheses.

Section	Krippendorff's α
Title	0.31
Key message	0.53
Authors	0.59
Background	0.22
Research question	0.85
Study selection	0.77
Selection criteria	0.53
Study approach	0.86
Study variables	0.83
Key results	0.21
Results reason	0.53
Publication bias	0.33
Results reliability	0.77
Everyday relevance	0.53
Funding info	0.81
Conflict of interest	0.93
Average	0.60

Table 3: Inter-annotator agreement between three experts, reported as Krippendorff's α for each section and overall.

Criterion	Avg.
Does the information in this slot correctly reflect the content title?	3.32
Does the text avoid irrelevant information?	3.18
The language is understandable for laypeople.	3.41
Is uncertainty in existing scientific findings adequately reflected in the choice of words (or is the language too assertive)?	2.05
The text is neutral (i.e., without judgemental or advice)	4.31
I would not recognize that the text has been written by an AI.	4.36

Table 4: Average ratings (1–5 scale) for each criterion.

expert-written summaries) across the 16 sections. Although BLEU scores are low for both systems, KLARpsy App outperforms FLEXRAG across all sections. Sentence similarity scores show a similar trend. KLARpsy App outperforms FlexRAG with an average score of 0.73 compared to 0.20. These results suggest that lay summarization cannot be reliably achieved with an out-of-the-box RAG system.

5 Conclusion and Future Work

We introduced KLARpsy App, a RAG system that supports psychology experts in creating plain language summaries of meta-analyses. Guided and optimized by a 37-point guideline, the system produces layperson-friendly drafts that experts refine for factuality and clarity – in a human-in-the-loop workflow. Automatic evaluation shows that the model produces solid initial drafts, achieving a similarity score of 0.73 compared to expert-written texts. Human evaluations indicated strong overall performance, with scores exceeding 3 out of 5 on most criteria. The only exception was criterion 4 ("uncertainty expression"), which received a score of 2. This finding strengthens the need for experts in the loop, as LLMs tend to produce verbose formulations and overstated claims about findings.

In future work, we will explore automatic prompt-optimization techniques to refine prompts so they better match the domain of each meta-analysis (e.g., education, crime and law, media, mental health). In addition, we plan to investigate reinforcement learning with an expert in the loop and expand the KLARpsy App to additional languages.

6 Limitations

Summarising scientific texts for a lay audience inevitably involves a trade-off between accessibility and technical accuracy. Making complex findings easier to understand often requires simplifying terminology and concepts, but this can risk omitting important qualifiers or subtly changing the intended meaning. Our design addresses this by positioning KLARpsy App as a tool to assist, rather than replace, domain experts. However, this approach depends on experts being available to review and refine outputs, which may limit scalability in practice.

The system is built around a psychological guideline developed specifically for plain language sum-

maries of meta-analyses in psychology. While this framework could serve as a starting point for other fields, it would require adaptation to the terminology, conventions, and standards of each specific domain. This limits its immediate, "out-of-the-box" applicability beyond psychology.

7 Ethical Considerations

The presented work involves generating plain-language summaries of psychological meta-analyses using large language models. LLMs may introduce hallucinations or overly confident claims that could mislead non-expert readers. Since inaccurate information may affect public understanding of scientific findings, the KLARpsy App outputs are intended only as initial drafts for experts rather than as a fully automatic tool. Moreover, the KLARpsy App is designed specifically for summarizing psychological meta-reviews and has not been tested in other domains. Therefore, the use outside of psychology should be validated by domain experts.

We use GPT-4.0 as the underlying model, and results may vary with other LLMs or future model updates. Summary quality and accuracy may differ across models, so applying the system with alternative LLMs requires additional validation.

References

Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, and Center for Devices and Radiological Health. 2017. Draft fda guidance on provision of plain language summaries. Accessed: [August 1, 2025].

Lisa Chamberlain James. 2024. Plain language summaries of clinical trial results: What is their role, and should patients and ai be involved? *Medical Writing*, 33(3):34–37.

Anita Chasiotis, Gesa Benz, Martin Kerwer, Pawel Nuwaltzew, Marlene Stoll, and Mark Jonas. 2023. Klarpsy-richtlinie zum verfassen allgemeinverständlicher zusammenfassungen psychologischer metaanalysen.

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

European Commission. 2023. Clinical trials– regulation eu no 536/2014. Accessed: [August 1, 2025].

European Medicines Agency. 2022. Clinical trials information system. Accessed: [August 1, 2025].

Biaoyan Fang, Xiang Dai, and Sarvnaz Karimi. 2024. Understanding faithfulness and reasoning of large language models on plain biomedical summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9890–9911, Miami, Florida, USA. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Aaron Grattafiori, Abhimanyu Dubey, and Llama Team. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw–Hill.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Nicholas Kloehn, Gondy Leroy, David Kauchak, Yang Gu, Sonia Colina, Nicole P Yuan, and Debra Revere. 2018. Improving consumer understanding of medical text: development and validation of a new subsimplify algorithm to automatically generate term explanations in english and spanish. *Journal of medical Internet research*, 20(8):e10779.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

OpenAI. 2022. Gpt-3.5 model. Accessed: [30.03.2024].

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2024. Structured outputs. `https://platform.openai.com/docs/guides/structured-outputs`. Accessed: 2025-08-06.

Katharina Otten, Lara Keller, Andrei A. Puiu, Beate Herpertz-Dahlmann, Jochen Seitz, Nils Kohn, J. Christopher Edgar, Lisa Wagels, and Kerstin Konrad. 2022. Pre- and postnatal antibiotic exposure and risk of developing attention deficit hyperactivity disorder–a systematic review and meta-analysis combining evidence from human and animal studies. *Neuroscience and Biobehavioral Reviews*, 140:104776.

Colleen Ovelman, Shannon Kugley, Gerald Gartlehner, and Meera Viswanathan. 2024. The use of a large language model to create plain language summaries of evidence reviews in healthcare: A feasibility study. *Cochrane Evidence Synthesis and Methods*, 2(2):e12041.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Carolina Salazar-Lara, Andrés Felipe Arias Russi, and Rubén Manrique. 2024. Bridging the gap in health literacy: Harnessing the power of large language models to generate plain language summaries from biomedical texts. *medRxiv*.

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.

Cathy Shyr, Randall W Grout, Nan Kennedy, Yasemin Akdas, Maeve Tischbein, Joshua Milford, Jason Tan, Kaysi Quarles, Terri L Edwards, Laurie L Novak, Jules White, Consuelo H Wilkins, and Paul A Harris. 2024. Leveraging artificial intelligence to summarize abstracts in lay language for increasing research accessibility and transparency. *Journal of the American Medical Informatics Association*, 31(10):2294–2303.

Aseem Srivastava, Smriti Joshi, Tanmoy Chakraborty, and Md Shad Akhtar. 2024. Knowledge planning in large language models for domain-aligned counseling summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17775–17789, Miami, Florida, USA. Association for Computational Linguistics.

Prashant D. Tailor, Haley S. D'Souza, Clara M. Castillejo Becerra, Heidi M. Dahl, Neil R. Patel, Tyler M. Kaplan, Darrell Kohli, Erick D. Bothun, Brian G. Mohney, Andrea A. Tooley, Keith H. Baratz, Raymond Iezzi, Andrew J. Barkmeier, Sophie J. Bakri, Gavin W. Roddy, David Hodge, Arthur J. Sit, Matthew R. Starr, and John J. Chen. 2024. Utilizing ai-generated plain language summaries to enhance interdisciplinary understanding of ophthalmology notes: A randomized trial. *medRxiv*.

Oguzhan Tas and Farzad Kiyani. 2017. A survey automatic text summarization. *PressAcademia Procedia*, 5(1):205–213.

Holly R. Tomlin and 1 others. 2024. Challenges and opportunities for professional medical publications writers to contribute to plain language summaries (pls) in an ai/ml environment – a consumer health informatics systematic review. In *AMIA Annual Symposium Proceedings*, volume 2023, pages 709–717. Symposium held 11 Jan 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Oisín Turbitt, Robert Bevan, and Mouhamad Aboshokor. 2023. MDC at BioLaySumm task 1: Evaluating GPT models for biomedical lay summarization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 611–619, Toronto, Canada. Association for Computational Linguistics.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Zhang Zhuocheng, Yang Feng, and Min Zhang. 2025. FlexRAG: A flexible and comprehensive framework for retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 621–631, Vienna, Austria. Association for Computational Linguistics.

A Lay Summary Example from the KLARpsy App

In the next sections, we present the same lay summary produced by psychology experts (Section A.1) and by KLARpsy App (Section A.2).[5] All summaries are generated from the meta-analysis reported in Otten et al. (2022).

A.1 Human Generated Summary

```
---
title: "Is the risk of ADHD higher in babies who had early exposure to antibiotics?"

key_message: "Babies do not have an increased risk of ADHD if they receive antibiotics shortly after
    birth. However, babies do have an increased risk of ADHD if their mothers take antibiotics
    during pregnancy."
---
### Background:
"ADHD (Attention-Deficit/Hyperactivity Disorder) is one of the most common developmental disorders of
     the nervous system. The gut microbiota also plays an important role in the development of such
    disorders. It can be disrupted when antibiotics are taken. Therefore, researchers are examining
    whether early exposure to antibiotics can lead to a child developing ADHD."

### Research question:
"With their review, the researchers wanted to find out: Do babies who come into contact with
    antibiotics very early have a higher risk of developing ADHD later on?"

### Which studies did the researchers look for in the review?
"The researchers looked for studies that examined a connection between early exposure to antibiotics
    as a baby and the development of ADHD. The exposure to antibiotics had to occur during the
    mother's pregnancy or up to two years after birth."

### Which studies did the researchers find for the review?
"The researchers found a total of 8 studies from the years 2016 to 2021. Of these, 4 studies from the
     years 2019 to 2021 involved babies who received antibiotics after birth, which they could
    combine into a meta-analysis. These results pertain to 1,863,867 babies. Another 4 studies from
    the years 2019 to 2021 involved babies whose mothers took antibiotics during pregnancy, which
    they could combine into another meta-analysis. These results pertain to 2,398,475 babies."

### What did the researchers do in the review?
"In the 8 studies, the researchers examined whether early exposure to antibiotics increased the risk
    of infants developing ADHD later on."

### What did the researchers investigate in the review?
"The following characteristics of the babies were examined:

- Type and timing of contact with antibiotics
  - Before birth
    - The mother took antibiotics during pregnancy.
    - The mother did not take antibiotics during pregnancy.
  - After birth
    - The baby received antibiotics in the first two years of life.
    - The baby did not receive antibiotics in the first two years of life.
- Development of ADHD in childhood
  - The baby later developed ADHD in childhood.
  - The baby did not develop ADHD later in childhood."

## What are the most important results?
"- When the mothers of the babies took antibiotics during pregnancy, the risk of the babies
    developing ADHD later was significantly higher. The summary risk measure, Hazard Ratio, was 1.23.
     This means that the risk of these babies developing ADHD later was 1.23 times higher compared
    to babies whose mothers did not take antibiotics during pregnancy.
- When the babies received antibiotics after birth, the risk of them developing ADHD later was not
    significantly higher compared to babies who did not receive antibiotics after birth."

### What is the reason for the results?
"The review article observed a correlation between the intake of antibiotics during the mother's
    pregnancy and a subsequent ADHD diagnosis in the child. Due to the types of studies that were
```

[5] For space reasons, we omit all comments and interface-related code that is constant across summaries. The expert version was originally written in German, machine-translated into English, and then double-checked by the experts.

```
    considered, it is only known that this correlation exists. However, it cannot be definitively
    stated that the antibiotics caused the ADHD."

### Are the results biased by limited publication of studies?
- What is the issue? Clear research results are easier to publish than ambiguous results. This is
    problematic for reviews. They cannot take unpublished results into account.

"- **What is it about?** Clear research findings are easier to publish than ambiguous results. This
    poses a problem for review articles because they cannot take unpublished results into account.

- **What does this mean for the present review article?** The researchers do not specify whether
    there are indications of such biases. It remains unclear whether the connection between
    antibiotics and ADHD is actually smaller than calculated in this review article."

### How reliable are the results?
"All the studies examined were observational studies. From these results, it cannot be determined
    whether antibiotics actually cause ADHD. ADHD could also have been caused by something else,
    such as the mother's illness, which required her to take antibiotics."

## What relevance to everyday life do the researchers see in the review?
"The researchers say that the results are important for the study of ADHD. However, improved studies
    are needed in order to make clear statements. Experiments with test animals would also help in
    drawing reliable conclusions."

### Who funded the review?
"The creation of the review article was funded by the German Research Foundation. This is an
    organization that promotes science and research in Germany. It is financed by the federal and
    state governments."

### Do the researchers report their own conflicts of interest in the review?
"The researchers report that there are no conflicts of interest present."
```

A.2 KLARpsy App Lay Summary

```
   ---
title: "**Can Taking Antibiotics During and After Pregnancy Increase a Child's Risk of Developing
    ADHD?**"

key_message: "The study investigates whether using antibiotics around the time of birth could
    increase the risk of ADHD in children by affecting gut health and brain development. Animal
    studies have previously shown that early antibiotic use can change gut health and behavior,
    implying a connection between the gut and brain development."
---
### Background:
"The study investigates if taking antibiotics around the time of birth could increase a child's risk
    of developing ADHD. It suggests that antibiotics might alter gut bacteria, potentially impacting
     brain development and behavior. The research also considers other factors like genetics and the
     environment that could affect ADHD."

### Research question:
With their review, the researchers wanted to find out:"Does exposure to antibiotics before and
    shortly after birth increase the risk of a child developing ADHD, possibly by affecting gut
    health and brain development?"

### Which studies did the researchers look for in the review?
"Researchers looked at studies to see if taking antibiotics during pregnancy or early childhood is
    connected to being diagnosed with ADHD later on. They focused on studies involving humans in the
     first two years of life and some experiments with animals. Only studies written in English
    before January 2021 were considered, and two experts carefully checked each study to make sure
    it was relevant and met their criteria."

### Which studies did the researchers find for the review?
"The researchers found a total of 21 studies published between 1995 and 2017, involving 2,398,475
    subjects who participated in the analysis of prenatal antibiotic exposure and 1,863,867 subjects
     in the study on postnatal exposure within the first two years of life."

### What did the researchers do in the review?
"In the 21 studies, The researchers looked at how giving antibiotics to mothers before and shortly
    after birth could affect a child's chances of developing ADHD. They gathered information from
```

different human and animal studies. In human studies, they explored any possible link between antibiotic use and ADHD. The animal studies helped them understand how antibiotics might impact gut health and behavior. By using statistical techniques, they made sure their results were dependable, considering both the direct and indirect effects of antibiotics on the risk of ADHD ."

What did the researchers investigate in the review?
"- Investigating the impact of early antibiotic exposure on the risk of developing ADHD.
- Main variables investigated:
  - Prenatal Antibiotic Exposure
  - Postnatal Antibiotic Exposure
  - ADHD Diagnosis"

What are the most important results?
"The research found that taking antibiotics before birth is linked to a higher risk of a child developing ADHD, likely due to changes in gut health. However, giving antibiotics to young children after they're born doesn't generally increase ADHD risk, except in premature babies, where it might play a role. Scientists suggest that how antibiotics affect gut health could influence brain development, but more studies are needed to fully understand these connections."

What is the reason for the results?
"In the review, relationships between prenatal antibiotic exposure and ADHD risk were observed. Because of the type of studies that were considered, it can be said with a high degree of certainty,
- that prenatal antibiotic exposure is also the cause of an increased risk of developing ADHD in the general population.
- that postnatal antibiotic exposure does not cause increased susceptibility to ADHD, except possibly in preterm infants."

Are the results biased by limited publication of studies?
- What is the issue? Clear research results are easier to publish than ambiguous results. This is problematic for reviews. They cannot take unpublished results into account.

"What does this mean for this review? The researchers provide no information on whether there are indications of such biases. Whether the results indicating an increased risk for ADHD after prenatal antibiotic exposure are actually smaller than calculated in this review thus remains unclear."

How reliable are the results?
"The results of the study aren't very strong because there were only a few studies included, and they weren't all the same. Also, the way the study was designed might have added some confusion. The connections found might actually be due to the infections themselves rather than the antibiotics used."

What relevance to everyday life do the researchers see in the review?
"Taking antibiotics during pregnancy might increase the chances of the child developing ADHD by affecting gut health. After birth, antibiotics generally don't raise ADHD risk, except possibly in premature babies. Changes in gut health may impact brain development, but more research is needed to understand these links."

Who funded the review?
"The preparation of the review was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)."

Do the researchers report their own conflicts of interest in the review?
"The researchers report that they have no conflicts of interest."