Beyond One-Size-Fits-All: Multi-Agent Refinement Framework for Persona-Based Biomedical Summarization

Rohan Charudatt Salvi[1], **Chirag Chawla**[2], **Md. Shad Akhtar**[3], **Shweta Yadav**[1]
[1]University of Illinois Chicago [2]Indian Institute of Technology Varanasi
[3]Indraprastha Institute of Information Technology Delhi
{rcsalvi2, shwetay}@uic.edu

Abstract

Biomedical lay summarization aims to make biomedical research accessible to non-experts, but most approaches assume a uniform audience, overlooking variation in medical literacy and information needs. We present MAPS (Multi-Agent Persona-based Summarization), a framework that generates persona-specific summaries through iterative cross-agent feedback. Human evaluation shows MAPS improves quality over single-agent baselines, while automatic metrics fail to capture these gains. LLM-based judges also exhibit limited sensitivity, assigning inflated scores and misdetecting errors. These findings highlight the need for improved evaluation methods for persona-based summarization.

Keywords: Biomedical Text Summarization, Controlled Text Generation, Multi-Agent System, Summary Evaluation

1. Introduction

Lay summarization aims to summarize jargon-heavy texts for non-experts in plain language (King et al., 2017). The lay summaries make scientific knowledge accessible beyond specialized communities (Stableford and Mettger, 2007). Prior lay summarization work has primarily targeted a uniform non-expert audience (Guo et al., 2021; Goldsack et al., 2023). In the real world, however, information needs vary widely across audiences. A pre-medical student benefits from more technical depth than a lay person, while a researcher looks for methodological detail that neither would expect (Salvi et al., 2025a). These differences motivate *persona-based summarization*, which requires controlling not only linguistic complexity but also content selection, such as what to explain, retain, and how much detail to include.

Prior controllable summarization work has focused on linguistic attributes such as sentence length or reading level (Luo et al., 2022; Tran et al., 2025b), without modeling how different audiences require different content, not just different wording. The PERCS dataset (Salvi et al., 2025a) addresses this by providing summaries tailored to four personas that vary in both readability and information depth, but effective generation methods for producing such persona-specific summaries remain underexplored.

Persona-based summarization requires balancing competing objectives such as readability, faithfulness, and persona-appropriate detail that are difficult to jointly optimize (Mullick et al., 2024; Song et al., 2025). Multi-agent LLM frameworks offer a natural solution by decomposing generation into specialized subtasks (Fang et al., 2025; Mo and Hu, 2024), and have shown promise for text simplification (Fang et al., 2025) and lay summarization (Lyu and Pergola, 2024). However, applying multi-agent LLMs to persona-based summarization raises design questions. Feedback must reflect persona-specific criteria, since what counts as an error for a lay reader (e.g., unexplained jargon) differs from what matters for a researcher (e.g., omitted statistics). Adjudication must similarly balance readability and detail differently per persona.

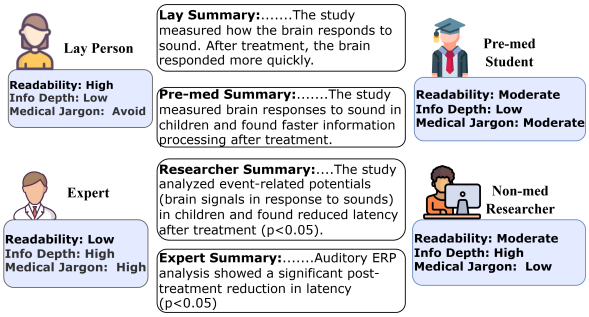

Figure 1: Persona-specific profiles highlighting differences in objectives and communication needs.

To address these challenges, this paper presents **MAPS** (Multi-Agent Persona-based Summarization), a framework for generating persona-specific biomedical summaries through iterative cross-agent feedback. Using PERCS, we study how design choices affect persona alignment by comparing feedback strategies and adjudication mechanisms. We also examine whether automatic metrics and LLM-based judges reliably evaluate persona-specific summaries. Our results show MAPS improves summary quality in human evaluation, while automatic metrics and LLM judges exhibit weak alignment with human judgments, highlighting a key evaluation challenge.

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 320–332
12 May 2026.

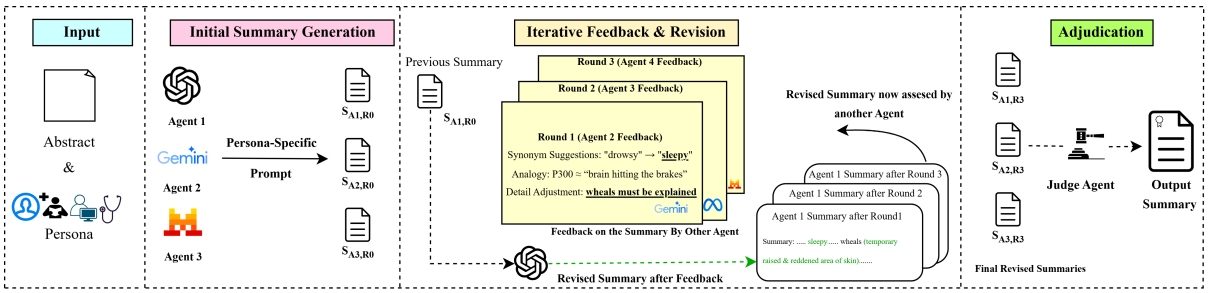

Figure 2: Overview of the MAPS pipeline ($S_{i,j}$: summary from agent i in feedback round j).

2. Background

Prior work on biomedical lay summarization has explored in-context learning and prompting-based methods (Goldsack et al., 2025; Ming et al., 2025), largely focusing on controlling linguistic attributes such as reading level (Luo et al., 2022; Tran et al., 2025a). Persona-dependent content selection remains underexplored.The PERCS dataset (Salvi et al., 2025a) addresses this gap by enabling persona-based evaluation and establishing single-LLM baselines, which we extend by studying multi-agent generation. Multi-agent LLM frameworks decompose generation into specialized subtasks to better address various objectives such as faithfulness, readability, and attributes like length (Fang et al., 2025)(Fang et al., 2025; Mo and Hu, 2024; Lyu and Pergola, 2024), though prior multi-agent work reports mixed results on automatic metrics (Zhu et al., 2025). In contrast, we evaluate summaries using human evaluation, LLM judges, and automatic metrics to better assess persona alignment.

3. Method

3.1. MAPS Framework

Given a biomedical abstract D and target persona P, MAPS generates a summary S that is faithful to D while matching persona-specific readability and information depth. MAPS uses k agents $\mathcal{A} = \{A_1, \ldots, A_k\}$ to generate, iteratively refine, and adjudicate persona-specific summaries (Figure 2).

Stage 1: Initial Generation Each agent generates a persona-conditioned summary, yielding diverse candidates $\{S_i^{(0)}\}$.

Stage 2: Iterative Feedback & Revision Initial summaries contain factual errors, omit information, or exhibit persona mismatches (Fang et al., 2024). MAPS refines them through T rounds of cross-agent feedback: in each round, agent A_i receives critique from agent $A_{(i \bmod k)+1}$ in round-robin scheme and revises its summary. Separating critique from generation reduces error reinforcement common in self-refinement (Wadhwa et al., 2024; Xie et al., 2025). We explore two feedback strategies:

Persona-Centered Feedback: This strategy explicitly specifies expectations for each target persona in terms of content depth, readability level, assumed biomedical knowledge, and summary length. Rather than leaving agents to infer audience needs implicitly, we define persona-specific feedback categories covering simplification requirements, level of detail, key point coverage, and faithfulness. For instance, feedback for a researcher persona emphasizes methodological and numerical details, whereas feedback for a lay persona prioritizes accessible definitions and reduces complexity. Agents evaluate summaries against these criteria and suggest specific revisions to improve persona alignment and faithfulness. As shown in Figure 3, persona-centered feedback highlights a missing definition of "ERP" and notes omitted p-values essential for a research audience. Based on prior work (You et al., 2024; Ming et al., 2025), we also developed a variant *Persona-Centered-Wiki* that provides Wikipedia as a tool to feedback agents, enabling them to verify definitions for faithfulness.

Error-Structured Feedback: Rather than aligning summaries through explicit persona profiles,

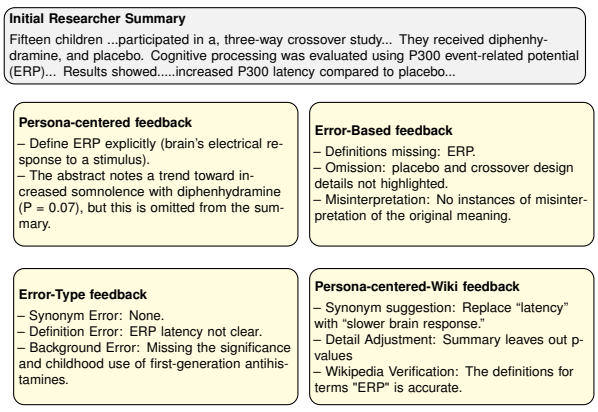

Figure 3: Illustrative examples of feedback strategies employed in the MAPS framework.

this strategy draws on research documenting common errors in lay summaries (Guo et al., 2024; Joseph et al., 2024). The feedback agent assumes the role of a medical expert reviewing summaries against a structured error taxonomy. Unlike persona-centered feedback, this approach is diagnostic rather than prescriptive: agents flag errors without suggesting corrections, leaving revision to the generating agent. We adopt the PERCS taxonomy (Salvi et al., 2025a), comprising 11 error types across four categories: new information errors (incorrect definitions, synonyms, background), inference errors (contradictions, omissions, misinterpretations), structure and readability issues, and hallucinations. We test two variants: *Error-based* flags all categories simultaneously, while *Error-type* targets one category per round.As shown in Figure 3, this approach identifies similar issues as persona-centered feedback but categorizes them systematically (e.g., "Definition Error. ERP latency not fully clear").

Stage 3: Adjudication After T rounds, an adjudication step selects a summary among candidates based on comprehensiveness, readability, and faithfulness. We explore three strategies: (a) **Judge Agent**, where a dedicated agent selects the best candidate, (b) **Voting**, where agents vote independently with ties resolved by the judge, and (c) **Combined Generation**, where a judge synthesizes a new summary by merging strengths across all candidates.

3.2. Experimental Setup

We evaluate on the PERCS dataset (Salvi et al., 2025a), which contains 500 biomedical abstracts paired with 2,000 expert-curated summaries across four personas: layperson, pre-medical student, non-medical researcher, and medical expert. We use the 150-abstract test split with k=4 agents and T=3 feedback rounds, comparing against zero-shot, few-shot, and self-refine baselines using the same LLMs as PERCS (GPT-4o, Mixtral-8x7B-Instruct, Gemini-2.0 Flash Lite, and LLaMA-3 70B). For MAPS, we test single-model (agents from the same LLM) with persona-based feedback only, and multi-model (agents from diverse LLMs) settings across all feedback strategies and adjudication methods. We include full prompt templates, agent instructions, and evaluation guidelines in the appendix, where we provide the Researcher persona as a representative example. Prompts for all other personas, along with the full codebase, are publicly available at GitHub. A token consumption and cost comparison between baselines and MAPS is provided in Table 13 in the Appendix.

3.3. Evaluation

Automatic Metrics. We use ROUGE-1,2,L (Lin, 2004) and SARI (Xu et al., 2016) for comprehensiveness, and FKGL, DCRS, CLI, and LENS (Maddela et al., 2022) for readability.

Human and LLM Evaluation. Both human raters and LLM evaluate summaries across criteria following Salvi et al. (2025b) guidelines on a 5-point Likert scale: i) *Comprehensiveness*: the extent to which the summary contains information necessary for the target persona to understand the research, ii) *Readability*: how easy the summary is to read based on appropriate medical jargon simplification. iii) *Factuality*: the extent to which the summary is factually consistent with the source abstract.

For human evaluation, we recruited two raters per persona via Upwork, selecting annotators whose background aligned with the target persona. Each annotator rated summaries from three methods (GPT-4o zero-shot as a consistent baseline across all personas, the best-performing PERCS model per persona, and MAPS) for 30 abstracts in their matched persona. To ensure consistent annotation, we measured inter-rater agreement using Krippendorff's α, which exceeded 0.84 for all personas. For LLM-based rating, we prompt LLaMA-3 with persona definitions and criteria to rate the same summaries. Additionally, we evaluate whether LLMs can reliably identify errors by prompting LLaMA-3, Mixtral, and Gemini to classify errors according to the PERCS taxonomy on 50 summaries and comparing against medical expert judgments. Notably, we selected LLaMA-3 as the LLM judge based on error detection performance, as it showed the closest alignment with human judgments, particularly for omission errors.

4. Results

(1) Does MAPS Setup Improves Performance? We compare MAPS (persona-specific feedback with combination adjudication) against GPT-4o zero-shot and the best-performing PERCS model per persona. Table 1 reports results for the selected configurations. On automatic metrics, MAPS does not consistently outperform baselines. However, human evaluation shows a contrasting trend where it substantially outperforms both PERCS models and GPT-4o for Pre-med, Researcher, and Expert personas. For example, MAPS improves Researcher comprehensiveness from 3.36 (Mixtral) to 4.74, and Expert comprehensiveness from 4.33 (GPT-4o) to 4.85. For Layperson, MAPS performs comparably to GPT-4o. Overall, multi-agent feedback markedly improves persona alignment for advanced personas while yield-

Persona	Method Category	Method	R-1 ↑	R-2 ↑	R-L ↑	SARI ↑	FKGL	DCRS	CLI	LENS
Layperson	Single LLM	GPT-4o Few-shot	0.6016	0.2476	0.3554	53.5134	11.22	8.94	11.99	75.0276
		GPT-4o Zero-shot	0.5972	0.2439	0.3390	53.0256	10.37	8.54	11.30	51.7573
	MAPS	Combine	0.5478	0.2067	0.2914	50.6332	12.80	11.54	13.60	75.6658
Premed	Single LLM	Llama-3.1 Few-shot	0.6149	0.2926	0.3823	51.9053	16.72	11.78	17.69	53.6619
		GPT-4o Zero-shot	0.5916	0.2380	0.3434	49.5828	17.91	12.77	19.44	46.2652
	MAPS	Combine	0.5962	0.2657	0.3469	50.7323	14.96	10.35	15.72	50.1168
Researcher	Single LLM	Mixtral Few-shot	0.6122	0.3072	0.3978	49.4988	15.05	10.65	15.93	59.3144
		GPT-4o Zero-shot	0.6207	0.2742	0.3831	43.9918	15.87	11.00	16.48	51.2367
	MAPS	Combine	0.6106	0.2815	0.3677	49.8121	16.21	11.01	16.99	52.0884
Expert	Single LLM	GPT-4o Zero-shot	0.6000	0.2905	0.3997	42.1944	17.91	12.77	19.44	51.2367
	MAPS	Combine	0.6148	0.3576	0.4213	47.8559	17.24	11.98	18.26	52.0884

Table 1: Selected performance on PERCS across personas using automatic evaluation metrics.

Lay

J	Method	C	R	F
H	GPT-4o-ZS	4.83	**4.97**	4.87
H	GPT-4o-FS	**4.87**	**4.97**	4.93
H	MAPS	4.83	4.93	4.83
L	GPT-4o-ZS	**4.83**	4.57	4.94
L	GPT-4o-FS	**4.83**	4.53	**4.96**
L	MAPS	4.71	**4.63**	**4.96**

Pre-med

J	Method	C	R	F
H	GPT-4o-ZS	4.65	4.45	4.93
H	Llama3-FS	4.52	4.12	4.8
H	MAPS	**4.81**	**4.71**	**5.00**
L	GPT-4o-ZS	**4.96**	4.75	**4.97**
L	Llama3-FS	4.6	4.8	4.96
L	MAPS	4.86	**4.94**	4.96

Researcher

J	Method	C	R	F
H	GPT-4o-ZS	4.69	**4.5**	**4.93**
H	Mixtral-FS	3.36	3.21	3.71
H	MAPS	**4.74**	4.47	**4.93**
L	GPT-4o-ZS	**4.30**	**5.00**	**4.97**
L	Mixtral-FS	4.1	4.83	4.77
L	MAPS	4.18	**5.00**	**4.97**

Expert

J	Method	C	R	F
H	GPT-4o-ZS	4.33	**4.06**	4.36
H	MAPS	**4.85**	4.15	**4.91**
L	GPT-4o-ZS	**4.99**	**5.00**	**4.99**
L	MAPS	4.90	**5.00**	4.90

Table 2: LLM and human evaluation scores across methods and personas measured by Comprehensiveness (C), Readability (R), and Faithfulness (F), with Judge (J) indicating Human (H) or LLM (L).

Error Type	Gemini	Llama	Mixtral	Experts
Background Information	0	6	3	0
Definition	2	1	3	0
Synonym	0	1	0	0
Entity	4	5	7	0
Contradiction	0	0	0	0
Misinterpretation	1	2	6	3
Omission	2	26	36	74
Jumping to Conclusions	0	2	0	1
Structure	0	0	2	0
Persona Relevance	0	2	2	0
Hallucination	0	0	0	2

Table 3: Comparison between LLMs as judges and human experts for error detection on a set of 50 summaries

ing comparable performance for lay summarization.

(2) Are Automatic Metrics Suitable? Our results show automatic metrics fail to capture persona-appropriate quality.

Comprehensiveness: Weak ROUGE-human alignment is documented in biomedical summarization (Wang et al., 2023). Our human evaluation further shows that information depth varies across personas, a distinction human raters capture but lexical metrics do not. Mixtral Few-shot achieves the best ROUGE for Researcher yet receives drastically lower human Comp. (3.36 vs. 4.74 for MAPS), indicating lexical overlap does not capture persona-appropriate information. SARI, designed for simplification, is misaligned with expert personas where preserving terminology is desirable, decreasing monotonically from Layperson (53.51) to Expert (42.19) and penalizing summaries that retain medical language.

Readability: Lower readability scores align with lay personas and higher scores with experts, but the metrics fail to distinguish intermediate personas. Pre-med (14.96-17.91) and Researcher (15.05-16.21) overlap substantially and differ in information depth while using similar levels of jargon (medical and scientific respectively). LENS, trained to predict simplification quality, exhibits similar limitations, inversely correlating with human judgments for non-lay personas. For Pre-med, LLaMA-3.1 achieves higher LENS (53.66) than MAPS (50.12), yet humans rate MAPS substantially better (Comp. 4.81 vs. 4.52, Read. 4.71 vs. 4.12). This highlights that simplification-oriented metrics actively misrank summaries when simplification is not the sole objective.

(3) How Reliable are LLM-Based Judges? LLM-based judges diverge substantially from human evaluations. For summary rating (Table 2), LLMs assign higher scores with limited variance and frequently favor baselines over MAPS, even when humans strongly prefer MAPS. Strong annotator agreement across personas ($\alpha > 0.84$) suggests this divergence reflects LLM limitations in capturing persona requirements rather than subjective persona definitions. Similarly, for error detection (Table 3), LLMs substantially underdetect omissions (Human: 74, LLaMA: 26, Mixtral: 36, Gemini: 2) while flagging unsupported errors and missing hallucinations. These results indicate the need for improved LLM-based evaluation techniques tailored to persona-specific summarization.

(4) Which Multi-Agent Setup Performs Best? MAPS with persona-centered prescriptive feedback and combination adjudication consistently performs best across personas. Full results across all configurations are reported in Tables 6-13 in the appendix. We make several observations. First,

multi-model setups outperform single-model configurations, particularly for the Researcher and Expert personas, suggesting that diversity across agents produces more balanced and persona-appropriate summaries. Second, among adjudication strategies, combination adjudication outperforms both judge-only and voting approaches, indicating that synthesizing strengths across candidate summaries is more effective than selecting a single best one.

Third, persona-centered feedback outperforms error-structured variants, which we attribute to its prescriptive nature. Rather than simply flagging problems, it provides agents with concrete and actionable revision targets that are more consistently incorporated. Finally, comparing MAPS against the self-refine baseline shows that cross-agent feedback outperforms single-agent iterative refinement, especially on comprehensiveness for researcher and expert personas, highlighting that multi-agent collaboration drives gains beyond what self-correction alone can achieve. Together, these findings also explain why error-structured feedback is less effective, as LLM judges have limited sensitivity to fine-grained errors, making diagnostic feedback a weaker signal than explicit persona-centered guidance.

5. Conclusion

We introduced MAPS, a multi-agent framework for persona-based biomedical summarization. Human evaluation shows MAPS improves summary quality, but current evaluation methods fail to capture these gains. Unlike lay summarization where lower readability signals quality, persona-based summaries require varying objectives, making standard metrics unreliable. LLM judges struggle, assigning higher scores and missing errors like omissions. These findings highlight the need for evaluation frameworks that model persona expectations.

6. Limitations

Our work has several limitations that suggest directions for future research. First, we evaluate only on the PERCS dataset, which includes four predefined biomedical personas. While these personas capture meaningful variation in medical literacy, real users may exhibit hybrid or more fine-grained needs. Extending this study to more granular personas, additional domains such as law, climate science, or multilingual settings would help assess its generalizability. Second, although we show that automatic metrics and LLM judges fail to capture persona-specific quality, we do not introduce new evaluation methods to address this gap. Developing persona-aware metrics and reference-free evaluation approaches remains an important direction for future work. Moreover, our multi-agent framework also considers only round-robin feedback. Exploring alternative feedback structures, such as hierarchical or fully connected schemes, may offer different trade-offs between feedback diversity and coherence. Lastly, our study centers on prompting-based methods, which are effective in this setting, while the exploration of persona-based fine-tuning and reinforcement learning is left to future work.

7. Bibliographical References

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.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912.

Tomas Goldsack, Carolina Scarton, and Chenghua Lin. 2025. Leveraging large language models for zero-shot lay summarisation in biomedicine and beyond. *arXiv preprint arXiv:2501.05224*.

Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023. Enhancing biomedical lay summarisation with external knowledge graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8016–8032, Singapore. Association for Computational Linguistics.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9194–9211, Miami, Florida, USA. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

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.

Stuart RF King, Emma Pewsey, and Sarah Shailes. 2017. An inside guide to elife digests. *ELife*, 6:e25410.

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. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen Lyu and Gabriele Pergola. 2024. Society of medical simplifiers. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 61–68, Miami, Florida, USA. Association for Computational Linguistics.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2022. Lens: A learnable evaluation metric for text simplification. *arXiv preprint arXiv:2212.09739*.

Shufan Ming, Yue Guo, and Halil Kilicoglu. 2025. Towards knowledge-guided biomedical lay summarization using large language models. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 285–297, Albuquerque, New Mexico. Association for Computational Linguistics.

Kaijie Mo and Renfen Hu. 2024. Expertease: A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099.

Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Bhowmick, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. 2024. On the persona-based summarization of domain-specific documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14291–14307, Bangkok, Thailand. Association for Computational Linguistics.

Rohan Charudatt Salvi, Chirag Chawla, Dhruv Jain, Swapnil Panigrahi, Md Shad Akhtar, and Shweta Yadav. 2025a. Percs: Persona-guided controllable biomedical summarization dataset. *arXiv preprint arXiv:2512.03340*.

Rohan Charudatt Salvi, Swapnil Panigrahi, Dhruv Jain, Shweta Yadav, and Md. Shad Akhtar. 2025b. Towards understanding LLM-generated biomedical lay summaries. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 260–268, Albuquerque, New Mexico. Association for Computational Linguistics.

Junjie Song, Yiwen Liu, Dapeng Li, Yin Sun, Shukun Fu, Siqi Chen, and Yuji Cao. 2025. Balancing rewards in text summarization: Multi-objective reinforcement learning via hypervolume optimization. *arXiv preprint arXiv:2510.19325*.

Sue Stableford and Wendy Mettger. 2007. Plain language: a strategic response to the health literacy challenge. *Journal of public health policy*, 28(1):71–93.

Hieu Tran, Zonghai Yao, Won Seok Jang, Sharmin Sultana, Allen Chang, Yuan Zhang, and Hong Yu. 2025a. Medreadctrl: Personalizing medical text generation with readability-controlled instruction learning. *arXiv preprint arXiv:2507.07419*.

Hieu Tran, Zonghai Yao, Lingxi Li, and Hong Yu. 2025b. ReadCtrl: Personalizing text generation with readability-controlled instruction learning. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 19–36, Albuquerque, New Mexico, US. Association for Computational Linguistics.

Manya Wadhwa, Xinyu Zhao, Junyi Jessy Li, and Greg Durrett. 2024. Learning to refine with fine-grained natural language feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12281–12308, Miami, Florida, USA. Association for Computational Linguistics.

Lucy Lu Wang, Julia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey Kuehl, Erin Bransom, and Byron C Wallace. 2023. Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9871–9889.

Zhihui Xie, Jie Chen, Liyu Chen, Weichao Mao, Jingjing Xu, and Lingpeng Kong. 2025. Teaching language models to critique via reinforcement learning. *arXiv preprint arXiv:2502.03492*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zhiwen You, Shruthan Radhakrishna, Shufan Ming, and Halil Kilicoglu. 2024. UIUC_BioNLP at BioLaySumm: An extract-then-summarize approach augmented with Wikipedia knowledge for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 132–143, Bangkok, Thailand. Association for Computational Linguistics.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. 2025. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. *arXiv preprint arXiv:2505.12371*.

A. Human Evaluation Guidelines

Goal. The goal of this evaluation is to assess the quality of model-generated summaries. Evaluators will rate summaries based on their **comprehensiveness**, **layness**, **usefulness**, and **factuality**.

Materials Provided for Evaluation.

- Abstract of the article
- Model-generated plain-text summary for the abstract

Evaluation Criteria.

- We use a 1-5 Likert scale where 1 indicates poor quality and 5 indicates excellent quality.
- The interpretation of each rating for each evaluation facet is defined below.

Comprehensiveness. This criterion evaluates how well the summary contains information necessary for a non-expert reader to understand the high-level topic and the significance of the research.

1. The summary is incomplete; an evaluator cannot understand the topic or the significance of the research.
2. The summary is partially complete; an evaluator gains a vague idea of the topic but cannot grasp the significance due to missing key details.
3. The summary allows an evaluator to understand the topic but lacks important details that convey the research's significance.
4. The summary enables an evaluator to understand both the topic and significance, missing only minor details that could enhance understanding.
5. The summary thoroughly covers all necessary information, allowing an evaluator to fully understand the topic and the significance of the research.

Readability. Readability is measured based on the use of medical jargon, sentence structure, and whether explanations are provided for technical terms.

1. There is little difference between the plain-text summary and the original abstract.
2. The summary omits a few jargon-heavy sentences or removes some technical words. It is easier to read but does not meaningfully simplify the content.
3. The summary contains a mix of jargon and simple terms, simple and complex sentences, and some definitions. Lay readers may understand the main points but may find certain terms or sentences confusing.
4. The summary is generally easy to understand, with occasional complex sentences or unexplained medical terms.
5. The summary avoids jargon or replaces it with simple synonyms. When this is not possible, it provides sufficient context or explanations. Sentences are simple and clear, making the summary accessible to a general audience.

Factuality. Factuality measures the extent to which the plain-text summary is consistent with the information presented in the abstract.

1. The summary alters the findings or methodology, misrepresenting the study.
2. The summary alters parts of the study in ways that may lead to misinterpretation of the methods or results.
3. The summary is largely accurate but contains frequent minor inconsistencies, such as incorrect figures, typos, or omission of key findings.
4. The summary is accurate with one or two minor exceptions.
5. The summary is fully factual and aligns completely with the abstract.

Procedure.

Error	Description
Incorrect Definitions	Wrong or misleading explanations of medical terms or concepts.
Incorrect Synonyms	Replacing medical words with inaccurate or oversimplified terms.
Incorrect Background	False or irrelevant contextual details about prevalence or treatment.
Entity Errors	Wrong factual details like numbers, names, or dosages.
Contradiction	Summary directly opposes the abstract's results or claims.
Omission	Missing key findings or results from the abstract.
Jumping to Conclusions	Overstating results beyond what data supports.
Misinterpretation	Misstating or oversimplifying meaning of the abstract.
Structure Error	Disorganized layout or mixing sections like methods and results.
Persona Relevance	Language complexity unsuitable for target persona.
Hallucination	Adding fabricated or irrelevant content not in abstract.

Table 4: Error types and descriptions for persona-aware biomedical summarization.

1. Each evaluator independently reviews the abstract, the reference lay summary, and the model-generated summaries.
2. Evaluators rate each summary using the specified evaluation facets and the provided Likert scale.

Method Category	Method	Model	R-1	R-2	R-L	SARI
Single LLM	Zero-shot	GPT-4o	0.5916	0.2380	0.3434	49.5828
		Gemini	0.5960	0.2495	0.3424	48.3962
		Mixtral	0.5819	0.2629	0.3622	49.3190
		LlaMA3	0.5900	0.2568	0.3429	49.0921
	Few-shot	GPT-4o	0.6089	0.2550	0.3629	50.4423
		Gemini	0.6024	0.2683	0.3648	48.3962
		Mixtral	0.5911	0.2686	0.3688	49.2427
		Llama3	0.6149	0.2926	0.3823	51.9053
	Self-Refine	GPT-4o	0.5624	0.2100	0.3126	47.8304
		Gemini	0.5564	0.2084	0.3011	46.4857
		Mixtral	0.4980	0.2028	0.2800	46.6206
		Llama3	0.4973	0.2073	0.2748	48.7969
Single LLM + Multi-Agent	Debate + Judge	GPT-4o	0.5121	0.1724	0.2626	45.8277
		Gemini	0.5135	0.1940	0.2740	47.0927
		Mixtral	0.5031	0.1999	0.2742	45.3938
		Llama3	0.5081	0.1991	0.2706	46.6280
Multi LLM + Multi-Agent	Feedback + Judge		0.4422	0.1610	0.2354	42.5489
	Feedback + Voting		0.5499	0.2064	0.2918	46.7492
	Feedback + Combine		0.5962	0.2657	0.3469	50.7323
	Wiki-Feedback + Judge		0.5691	0.2344	0.3189	45.8352
	Error Type + Judge		0.5730	0.2317	0.3236	46.2927
	Error Type + Voting		0.5680	0.2284	0.3204	45.9240
	Error Type + Combine		0.5756	0.2352	0.3263	46.7815
	Error Based + Judge		0.5653	0.2370	0.3143	45.3778
	Error Based + Voting		0.5601	0.2331	0.3108	44.9326
	Error Based + Combine		0.5669	0.2396	0.3161	45.8129

Table 5: Comprehensiveness scores for Pre-med Persona

Method Category	Method	Model	FKGL	DCRS	CLI	LENS
Single LLM	Zero-shot	GPT-4o	17.91	12.77	19.44	46.2652
		Gemini	15.47	12.69	18.34	49.8001
		Mixtral	16.08	11.80	17.53	50.3914
		Llama3	17.32	12.28	18.21	49.5902
	Few-shot	GPT-4o	15.74	11.12	16.73	45.8277
		Gemini	15.43	12.57	18.03	53.2864
		Mixtral	16.04	11.80	17.46	48.1884
		Llama3	16.72	11.78	17.69	53.6619
	Self-Refine	GPT-4o	15.83	11.43	17.59	47.0772
		Gemini	14.82	10.76	16.42	45.4320
		Mixtral	15.42	10.52	17.01	45.9425
		Llama3	16.42	10.07	15.98	46.0638
Single LLM + Multi-Agent	Debate + Judge	GPT-4o	13.85	10.70	15.83	49.7191
		Gemini	14.57	11.01	16.26	49.5062
		Mixtral	14.15	10.23	15.72	48.0412
		Llama3	14.82	10.22	15.62	47.7128
Multi LLM + Multi-Agent	Feedback + Judge		12.67	10.19	14.26	48.9614
	Feedback + Voting		14.51	10.50	15.60	48.8772
	Feedback + Combine		14.96	10.35	15.72	50.1168
	Wiki-Feedback + Judge		15.15	10.58	16.24	47.3505
	Error Type + Judge		16.14	10.52	16.62	49.5623
	Error Type + Voting		15.80	10.40	16.40	49.0148
	Error Type + Combine		16.25	10.60	16.75	50.0431
	Error Based + Judge		14.15	10.63	15.85	45.1387
	Error Based + Voting		13.95	10.55	15.65	44.7042
	Error Based + Combine		14.30	10.70	15.95	45.6129

Table 6: Readability Scores for Pre-med Persona

Method Category	Method	Model	R-1	R-2	R-L	SARI
Single LLM	Zero-shot	GPT-4o	0.6207	0.2742	0.3831	43.9918
		Gemini	0.6093	0.2838	0.3881	43.1912
		Mixtral	0.6165	0.3039	0.4086	49.5102
		Llama3	0.6120	0.2880	0.3719	48.4270
	Few-shot	GPT-4o	0.6259	0.2778	0.3878	49.1682
		Gemini	0.6242	0.2956	0.4000	44.9244
		Mixtral	0.6122	0.3072	0.3978	49.4988
		Llama3	0.6164	0.2987	0.3838	48.2722
	Self-Refine	GPT-4o	0.5624	0.2100	0.3126	46.3540
		Gemini	0.5564	0.2084	0.3011	46.6575
		Mixtral	0.4980	0.2028	0.2800	46.6433
		Llama3	0.4973	0.2073	0.2748	48.1174
Single LLM + Multi-Agent	Debate + Judge	GPT-4o	0.5521	0.1982	0.2905	43.9918
		Gemini	0.5792	0.2382	0.3308	43.1912
		Mixtral	0.5606	0.2392	0.3166	44.1105
		Llama3	0.5380	0.2192	0.2827	46.2658
Multi LLM + Multi-Agent	Feedback + Judge		0.5570	0.2161	0.3006	45.7067
	Feedback + Voting		0.5610	0.2182	0.3045	45.8344
	Feedback + Combine		0.6106	0.2815	0.3677	49.8121
	Wiki-Feedback + Judge		0.5651	0.2283	0.3128	45.7165
	Error Type + Judge		0.5730	0.2317	0.3236	45.8457
	Error Type + Voting		0.5680	0.2284	0.3204	45.4820
	Error Type + Combine		0.5756	0.2352	0.3263	46.3290
	Error Based + Judge		0.5653	0.2370	0.3143	45.5971
	Error Based + Voting		0.5601	0.2331	0.3108	45.1187
	Error Based + Combine		0.5669	0.2396	0.3161	45.9640

Table 7: Comprehensiveness scores for Researcher Persona

Method Category	Method	Model	FKGL	DCRS	CLI	LENS
Single LLM	Zero-shot	GPT-4o	15.87	11.00	16.48	51.2367
		Gemini	15.53	11.19	16.88	53.5571
		Mixtral	15.34	10.79	16.39	55.6439
		Llama3	15.19	10.02	14.95	57.7103
	Few-shot	GPT-4o	16.14	11.10	16.86	59.1560
		Gemini	15.28	11.01	16.42	54.1410
		Mixtral	15.05	10.65	15.93	59.3144
		Llama3	15.40	10.14	15.23	60.6989
	Self-Refine	GPT-4o	15.83	11.43	17.59	51.3194
		Gemini	14.82	10.76	16.42	50.6804
		Mixtral	15.42	10.52	17.01	51.2360
		Llama3	16.42	10.07	15.98	49.0160
Single LLM + Multi-Agent	Debate + Judge	GPT-4o	15.81	11.27	17.55	52.7774
		Gemini	14.93	11.03	16.53	48.8293
		Mixtral	15.80	10.63	17.21	50.2164
		Llama3	17.33	10.34	16.55	48.5189
Multi LLM + Multi-Agent	Feedback + Judge		15.79	10.80	16.95	52.7070
	Feedback + Voting		15.74	10.92	16.92	52.0788
	Feedback + Combine		16.21	11.01	16.99	52.0884
	Wiki-Feedback + Judge		16.08	10.99	17.28	53.6723
	Error Type + Judge		15.87	11.19	17.49	55.5326
	Error Type + Voting		15.78	11.08	17.40	55.0210
	Error Type + Combine		15.95	11.22	17.55	55.8870
	Error Based + Judge		16.58	10.98	17.40	52.7774
	Error Based + Voting		16.40	10.92	17.32	52.3012
	Error Based + Combine		16.62	11.04	17.46	53.0585

Table 8: Readability scores for Researcher Persona

Method Category	Method	Model	R-1	R-2	R-L	SARI
Single LLM	Zero-shot	GPT-4o	0.5972	0.2439	0.3390	53.0256
		Gemini	0.6029	0.2624	0.3691	53.0180
		Mixtral	0.5235	0.2068	0.3186	48.3885
		Llama3	0.5900	0.2585	0.3527	52.0627
	Few-shot	GPT-4o	0.6016	0.2476	0.3554	53.5134
		Gemini	0.6018	0.2611	0.3773	53.2864
		Mixtral	0.5073	0.1892	0.3014	47.2191
		Llama3	0.5841	0.2542	0.3499	53.6619
	Self-Refine	GPT-4o	0.5288	0.1727	0.2696	48.3749
		Gemini	0.5422	0.1874	0.2756	47.9891
		Mixtral	0.4750	0.1609	0.2523	47.1213
		Llama3	0.4233	0.1366	0.2008	47.4832
Single LLM + Multi-Agent	Debate + Judge	GPT-4o	0.4614	0.1212	0.2217	46.4322
		Gemini	0.2751	0.0268	0.1274	37.2884
		Mixtral	0.4341	0.1316	0.2237	48.3885
		Llama3	0.2724	0.0288	0.1269	37.7294
Multi LLM + Multi-Agent	Feedback + Judge		0.4721	0.1380	0.2344	46.6069
	Feedback + Voting		0.4810	0.1489	0.2402	47.0179
	Feedback + Combine		0.5478	0.2067	0.2914	50.6332
	Wiki-Feedback + Judge		0.5158	0.1729	0.2665	48.2856
	Error Type + Judge		0.4687	0.1350	0.2355	45.1380
	Error Type + Voting		0.4640	0.1322	0.2320	44.7021
	Error Type + Combine		0.4706	0.1384	0.2371	45.6214
	Error Based + Judge		0.4837	0.1390	0.2415	45.9380
	Error Based + Voting		0.4791	0.1361	0.2380	45.4027
	Error Based + Combine		0.4854	0.1415	0.2430	46.3886

Table 9: Comprehensiveness scores for Lay Persona

Method Category	Method	Model	FKGL	DCRS	CLI	LENS
Single LLM	Zero-shot	GPT-4o	10.37	8.54	11.30	51.7573
		Gemini	10.04	8.71	11.66	52.5033
		Mixtral	12.38	9.37	12.93	62.7845
		Llama3	12.07	8.59	11.86	53.1642
	Few-shot	GPT-4o	11.22	8.94	11.99	75.0276
		Gemini	11.00	8.97	12.17	69.1026
		Mixtral	13.08	9.81	13.84	64.7531
		Llama3	12.06	8.70	11.99	78.4790
	Self-Refine	GPT-4o	11.24	9.47	12.84	71.6981
		Gemini	9.44	8.45	11.14	71.5003
		Mixtral	12.49	9.55	13.71	57.7433
		Llama3	11.84	8.32	12.07	62.4744
Single LLM + Multi-Agent	Debate + Judge	GPT-4o	15.15	13.08	16.81	51.7573
		Gemini	12.54	11.86	14.15	52.5033
		Mixtral	14.70	12.72	15.57	57.8722
		Llama3	15.55	12.09	14.80	53.1642
Multi LLM + Multi-Agent	Feedback + Judge		14.64	12.73	15.87	54.6001
	Feedback + Voting		14.56	12.50	15.35	58.2415
	Feedback + Combine		12.80	11.54	13.60	75.6658
	Wiki-Feedback + Judge		13.66	12.19	14.93	66.8700
	Error Type + Judge		14.32	12.63	15.85	56.5742
	Error Type + Voting		14.18	12.42	15.62	56.0216
	Error Type + Combine		14.40	12.71	15.94	56.9820
	Error Based + Judge		14.52	12.83	16.25	58.0742
	Error Based + Voting		14.39	12.62	16.02	57.4185
	Error Based + Combine		14.61	12.90	16.31	58.6034

Table 10: Readability scores for Lay Persona

Method Category	Method	Model	R-1	R-2	R-L	SARI
Single LLM	Zero-shot	GPT-4o	0.6000	0.2905	0.3997	42.1944
		Gemini	0.6348	0.3451	0.4511	45.1591
		Mixtral	0.6478	0.3911	0.4758	48.2700
		Llama3	0.6594	0.3928	0.4738	47.8625
	Few-shot	GPT-4o	0.5418	0.2416	0.3555	41.4523
		Gemini	0.6599	0.3815	0.4912	44.6463
		Mixtral	0.6625	0.4035	0.4946	48.4338
		Llama3	0.6769	0.4198	0.5037	45.4880
	Self-Refine	GPT-4o	0.5577	0.2547	0.3535	42.4489
		Gemini	0.6010	0.3126	0.4184	44.1335
		Mixtral	0.6552	0.3973	0.4852	48.3519
		Llama3	0.4732	0.2345	0.2981	43.1135
Single LLM + Multi-Agent	Debate + Judge	GPT-4o	0.5418	0.2416	0.3555	42.1944
		Gemini	0.6599	0.3815	0.4912	41.1892
		Mixtral	0.6625	0.4035	0.4946	43.3271
		Llama3	0.6769	0.4198	0.5037	30.5329
Multi LLM + Multi-Agent	Feedback + Judge		0.5954	0.2974	0.3924	43.0934
	Feedback + Voting		0.5946	0.3007	0.3920	43.5686
	Feedback + Combine		0.6148	0.3576	0.4213	47.8559
	Wiki-Feedback + Judge		0.6057	0.3048	0.3975	42.3177
	Error Type + Judge		0.6012	0.2884	0.3938	40.5707
	Error Type + Voting		0.5964	0.2852	0.3905	40.1189
	Error Type + Combine		0.6028	0.2920	0.3953	41.0024
	Error Based + Judge		0.5827	0.2843	0.3633	41.7574
	Error Based + Voting		0.5782	0.2812	0.3601	41.2036
	Error Based + Combine		0.5841	0.2865	0.3650	42.1298

Table 11: Comprehensiveness scores for Expert Persona

Method Category	Method	Model	FKGL	DCRS	CLI	LENS
Single LLM	Zero-shot	GPT-4o	17.91	12.77	19.44	51.2367
		Gemini	15.47	12.69	18.34	53.5571
		Mixtral	16.08	11.80	17.53	55.6439
		Llama3	17.32	12.28	18.21	57.7103
	Few-shot	GPT-4o	15.57	11.39	16.84	59.1560
		Gemini	15.43	12.57	18.03	54.1410
		Mixtral	16.04	11.80	17.46	59.3144
		Llama3	16.72	11.78	17.69	60.6989
	Self-Refine	GPT-4o	17.30	12.76	19.39	51.3194
		Gemini	15.36	12.49	17.91	50.6804
		Mixtral	16.06	11.80	17.50	57.4792
		Llama3	17.33	11.95	18.10	49.0160
Single LLM + Multi-Agent	Debate + Judge	GPT-4o	15.57	11.39	16.84	53.5064
		Gemini	15.43	12.57	18.03	48.8293
		Mixtral	16.04	11.80	17.46	50.2164
		Llama3	16.72	11.78	17.69	48.5189
Multi LLM + Multi-Agent	Feedback + Judge		15.44	12.18	17.78	52.7070
	Feedback + Voting		16.34	12.39	18.49	52.0788
	Feedback + Combine		17.24	11.98	18.26	52.0884
	Wiki-Feedback + Judge		16.45	12.31	18.44	52.8652
	Error Type + Judge		17.23	12.57	19.10	55.5326
	Error Type + Voting		17.10	12.42	18.92	55.0214
	Error Type + Combine		17.28	12.60	19.14	55.8871
	Error Based + Judge		17.10	12.37	18.90	52.7774
	Error Based + Voting		16.97	12.22	18.72	52.3189
	Error Based + Combine		17.14	12.40	18.96	53.0642

Table 12: Readability scores for Expert Persona

Method	Model	Tokens	Cost (USD)
Few-shot	LLaMA-3	3.5K	$0.0010
	Mixtral	4.2K	$0.0020
	Gemini	3.5K	$0.0003
	GPT-4o	3.2K	$0.0104
Self-Refine	LLaMA-3	6.3K	$0.0020
	Mixtral	8.7K	$0.0049
	Gemini	5.5K	$0.0007
	GPT-4o	6.7K	$0.0218
MAPS	4 agents	49.7K	$0.0575

Table 13: Average token usage and estimated cost per sample across methods and models, estimated on 10 samples per persona (40 total).

B. Researcher Persona Prompts

Stage 1: Initial Generation

You are a knowledgeable science communicator with interdisciplinary expertise. Your task is to write a clear and comprehensive summary of a scientific abstract for researchers outside the specific field of the study but with a solid understanding of general scientific principles.

When summarizing, follow these key principles:

The audience consists of researchers in fields other than biology and medicine who may not be familiar with domain-specific terminology or methods. Use clear and accessible language. Avoid unnecessary technical jargon; when technical terms are necessary, explain them briefly in context. Accurately represent the study's main findings, methodology, and results. Maintain a professional but approachable tone. Aim for clarity and precision without overly simple or domain-specific language. Structure the summary as a paragraph. Do not use bullet points, numbered lists, or Q&A formats. Focus on the scientific content only. Do not include commentary about the summarization process or subjective judgment of the study's importance.

The final summary should not exceed 350 words.

ABSTRACT

{abstract}

Stage 2: Persona-specific Feedback

Task: Review the provided summary compared to the original abstract. Suggest improvements to make the summary better intended for researchers with a general scientific background who are not specialists in the specific field.

Feedback categories include definition enhancement, synonym suggestions, simplification needs, detail adjustment, factuality check, overall coherence, and key point highlighting.

For each category, provide specific examples from the summary and propose clear, constructive improvements. Where applicable, explain why the changes would enhance understanding for non-specialist researchers.

ABSTRACT

{abstract}

SUMMARY

{summary}

Stage 3: Feedback-Guided Revision

Refine the following summary intended for researchers with a general scientific background, but who are not specialists in the specific field, based on the feedback provided below.

Do not use numbered lists to explain terminology, methods, or findings, as this negatively impacts readability. Ensure that the revised summary does not exceed 350 words.

SUMMARY

{last_summary}

FEEDBACK

{feedback}

B.1. Adjudication Prompts for the Researcher Persona

Judge Strategy

You are a knowledgeable science communicator with strong interdisciplinary expertise evaluating summaries of a scientific abstract. Each summary is intended for researchers outside the original field but with a solid scientific background.

Evaluate each summary based on accuracy and faithfulness, clarity and accessibility, comprehensiveness, and professional tone. After reviewing each summary step by step, report your reasoning and state your final judgment as **Summary X**.

ABSTRACT

{abstract}

SUMMARIES

{summaries_text}

Combine Strategy

You are a skilled science communicator with deep interdisciplinary expertise. Your task is to produce a single clear, accurate, and accessible summary for non-medical researchers using multiple candidate summaries as input.

The final summary must maintain scientific faithfulness, avoid excessive jargon, and clearly explain findings and methods to non-specialists.

ABSTRACT

{abstract}

CANDIDATE SUMMARIES

{summaries_text}