The Manchester Bees at PerAnsSumm 2025: Iterative Self-Prompting with Claude and o1 for Perspective-aware Healthcare Answer Summarisation

Pablo Romero[1], **Libo Ren**[2], **Lifeng Han**[2,3*], and **Goran Nenadic**[2]
[1] Manchester Metropolitan University, UK
[2] The University of Manchester, UK
[3] LIACS & LUMC, Leiden University, Leiden, NL
[*] *corresponding author*
pablo2004romero,renlibo994@gmail.com
l.han@lumc.nl g.nenadic@manchester.ac.uk

Abstract

This system report presents an innovative approach to the PerAnsSumm2025 shared task at the Workshop CL4Health, addressing the critical challenges of perspective-aware healthcare answer summarization. Our method, Iterative Self-Prompting (ISP) with Claude and o1, introduces a novel framework that leverages large language models' ability to iteratively refine their own instructions, achieving competitive results without traditional model training. Despite utilising only API calls rather than computational-intensive training, our system "The Manchester Bees" secured 15th place among 23 leader board systems overall, while demonstrating exceptional performance in key metrics - ranking 6th in Strict-matching-F1 for span identification (Task A) and achieving the highest Factuality score for summary generation (Task B). Notably, our approach achieved state-of-the-art results in specific metrics, including the highest Strict-matching precision (0.2267) for Task A and AlignScore (0.5888) for Task B. This performance, accomplished with minimal computational resources and development time measured in hours rather than weeks, demonstrates the potential of ISP to democratise access to advanced NLP capabilities in healthcare applications. Our complete implementation is available as an open-source project on https://github.com/pabloRom2004/-PerAnsSumm-2025

1 Introduction

This system report presents our contribution to the PerAnsSumm 2025 shared task on perspective-aware healthcare answer summarization, organized in conjunction with the second edition of the CL4Health workshop (computational linguistics for healthcare) at NAACL 2025. The task addresses a critical challenge in modern healthcare: the growing reliance on online health forums where users seek medical advice from peers with similar experiences. While these forums provide valuable support, their unstructured nature necessitates effective methods for organizing and synthesizing the diverse perspectives they contain.

The PerAnsSumm shared task, based on the healthcare forum dataset developed by Naik et al. (2024), focuses on generating perspective-based summaries across five key categories: information, cause, suggestion, experience, and question. To address this challenge, this research proposes Iterative Self-Prompting (ISP), a novel approach utilising two decoder-only systems, Claude and o1. Our method leverages these models' capabilities to iteratively refine task-specific prompts through in-context learning from annotated training data. Notably, the systems demonstrated sophisticated analytical abilities, identifying patterns in data quality and autonomously adjusting prompts to handle edge cases and inconsistencies. Three versions of the system were submitted (ISP-claude/o1 v1, v2, v3), each showing strong performance across both primary tasks: span detection and classification (Task A) and summary generation (Task B). In the official evaluation among 23 top-performing systems, our approach achieved particularly notable results using the Strict-matching metric for Task A, ranking 6th in F1 score. For Task B, measured by Factuality metrics, our systems showed progressive improvement, with v1 ranking 6th (0.3545) and v3 achieving the top position (0.4277), primarily due to superior performance on the AlignScore submetric. Beyond these technical achievements, our method offers significant practical advantages in terms of computational efficiency and development time, suggesting a promising direction for future work in healthcare text analysis.

2 Related Work

2.1 Prompting Techniques

The evolution of prompt engineering for large language models (LLMs) has increasingly focused

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 340–348

on developing sophisticated methods that can fully leverage these models' inherent reasoning capabilities. Iterative Self-Prompting (ISP) follows naturally from research into various forms of model reasoning, including logical, common-sense, and symbolic reasoning, as explored by (Qiao et al., 2023). While researchers have made significant progress with techniques such as chain-of-thought (CoTs), in-context learning, and various prompting strategies (Cui et al., 2023), the field has increasingly recognized the potential of automated approaches. Notably, Automatic Prompt Engineering (APE) has demonstrated competitive performance compared to human-engineered prompts across several NLP tasks (Zhou et al., 2023), typically relying on evaluation scores for prompt refinement. Our work extends this paradigm by introducing a more sophisticated iterative framework that integrates multiple models in the automatic self-prompting process. This approach, inspired by recent advances in iterative refinement (Madaan et al., 2023), leverages sample-labeled data and self-feedback mechanisms to create a more robust and effective prompt engineering methodology.

2.2 Healthcare Data Summarisation

Healthcare data summarisation can be time consuming and costly, which has led to the automatic summarisation task in this domain. The data sources in this task can be electronic health records (EHRs) (Moen et al., 2016), clinical discharge summaries (Searle et al., 2023), medical papers (Sarker, 2014), and online forums (Naik et al., 2024), etc. The methodologies used for such tasks include extractive summarisation, abstractive summarisation, with/without (w/o) external domain knowledge base usage such as medical concepts. The models have included traditional training and fine-tuning paradigms and recent prompt engineering. The data this method utilizes is from perspective-aware online forum healthcare text by Naik et al. (2024).

3 ISP with Claude and o1

3.1 Methodology Overview

Iterative Self-Prompting (ISP) represents an advancement in approaches to prompt engineering and model instruction. At its core, the technique leverages a language model's ability to analyse, understand, and improve its own instructions through a structured feedback loop. This self-improving mechanism creates a powerful framework for developing highly effective prompts without the need for model training or extensive human intervention.

The process begins with a detailed description of the task provided to a language model. Rather than directly attempting to solve the problem, we ask the model to craft a prompt for completing the task. This meta-level approach allows the model to step back and think about how best to approach the problem systematically. The initial prompt generation phase is crucial, as it sets the foundation for all subsequent improvements.

Once we have an initial draft of the prompt, we enter the iterative refinement phase. This involves testing the prompt with training data and carefully analysing the results on another instance of the model with no other context for the task, just the prompt and the data. The key innovation here lies in how we use the model's own analytical capabilities. We present the model with its previous prompt, the outputs generated from the other model using that prompt, and the ground truth answer. The model then engages in a detailed analysis of what worked well and what needs improvement and refines the base prompt further, adding specific details to the prompt so that next time, the model does a little better on the task, this process is then repeated until the prompt is very detailed and outputs from the model are very high quality.

The power of this approach becomes apparent in how the model discovers and adapts to patterns in the data. For instance, when analysing outputs, the model might notice subtle patterns that weren't explicitly stated in the original task description. A concrete example of this meta-cognitive capability occurred during implementation when the model recognised the importance of handling empty categories in data classification tasks. The model observed that some categories naturally remain empty in certain cases and modified the prompt accordingly, without any human intervention. An example can be seen in Figure 2.

The theoretical implications of this technique extend beyond simple prompt engineering. It demonstrates a form of meta-learning, where the model learns to create better instructions through experience. This self-improving capability suggests interesting possibilities for autonomous systems that can optimise their own behaviour through structured self-reflection.

What makes ISPs particularly powerful is their universality. The technique doesn't depend on spe-

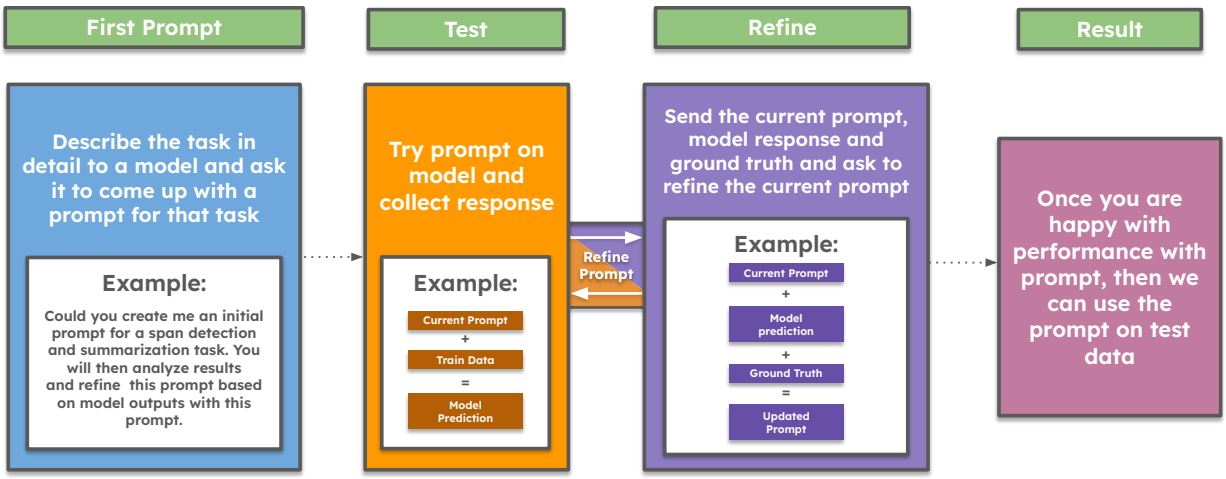

Figure 1: Iteration Cycle for ISP showing the process of prompt refinement through feedback loops.

cific model architectures or training approaches. Instead, it relies on the fundamental capabilities present in modern language models: understanding tasks, generating instructions, and analysing results. This makes it highly adaptable to different problems and domains.

3.2 ISP for PerAnsSumm Shared Task

Implementation Timeline:

- Hour 0 → Initial Setup
- Hour 1 → First Iterations
- Hour 2 → Refinement Cycles
- Hour 3 → Final optimisation
- Hour 4 → Submission

The practical effectiveness of Iterative Self-Prompting was demonstrated conclusively in the PerAnsSumm Shared Task at the CL4Health Workshop NAACL 2025. This implementation provides a compelling case study of how ISP can achieve competitive results with minimal resources and development time.

The task itself presented significant challenges in medical text analysis, requiring systems to identify and classify specific spans of text while generating comprehensive summaries across multiple categories. Traditional approaches typically involved fine-tuning models like BERT or T5, requiring substantial computational resources and development time. Instead, ISP was applied to develop a solution in just 3-4 hours, ultimately achieving a competitive performing system in the leader-board among top-23 teams.

The implementation began with a detailed task description provided to the model, including the **specific requirements** for *span identification and summary generation*. The model was then asked to **create a comprehensive prompt** that would *guide another model* through the analysis process. This initial prompt generation phase was crucial, as it needed to capture all the nuances of the task, from JSON formatting requirements to specific category definitions.

As demonstrated in Figure 2, our iteration cycle revealed remarkable self-improvement capabilities in the model's performance. Through systematic analysis of outputs against ground truth data, the model continuously identified specific areas for enhancement, leading to increasingly sophisticated prompt refinements. One particularly notable breakthrough emerged when the model began recognizing patterns in category distribution - specifically, how certain categories naturally remained empty in particular cases. This insight led to automated prompt adjustments that significantly improved the handling of these edge cases.

The successful development of these capabil-

<contemplator> Let me review the prompt and examples more carefully...

1. Checking Empty Categories Hmm... you're right. Looking at the example you provided (uri: 1564117), I notice:

- Empty arrays for EXPERIENCE and QUESTION spans
- Empty strings for EXPERIENCE and QUESTION summaries
- This is a crucial pattern I missed

Let me look through the other examples to understand empty category patterns:

- Some questions naturally have no QUESTION spans
- Some have no personal EXPERIENCE spans
- Others might lack CAUSE categorization
- Even high-quality answers might not cover all categories

This is an important clarification needed in the prompt. Let me think about where to add it...

Figure 2: Example of ISP process: reviewing the prompts

ities, however, presented us with an interesting technical challenge: managing the models' context window limitations. While our refined prompt demonstrated strong performance, implementing it across the full test set required careful consideration of these constraints. This challenge was addressed by splitting the 50-question test set into 10 manageable groups of 5 questions each. This segmentation allowed the o1 model to maintain sufficient context for each question while preserving the quality of analysis. Through a streamlined script, these individual results were then aggregated into the final submission, effectively balancing the computational and context constraints from the o1 model.

4 Submission to PerAnsSumm2025

Three systems were submitted to both shared tasks A and B, specifically the ISP-Claude/o1 versions 1, 2, and 3.

4.1 Submission outcomes

There are 155 submitted system outputs in the official shared evaluation sheet, however, only 23 systems were listed in the top-performing board from unique teams (no more than one system from each team). The system ranked 15th in the top-list by the 'Task A + B combined Average' score using Version 1 (out of three), scoring 0.3994 (as in Figure 7) (Agarwal et al., 2025). Using the official leaderboard scores from PerAnsSumm 2025, the advantages of the claude/o1 system are listed below for Task-A and B respectively.

For Task-A (span identification and classification) score, it is the average of classification weighted-F1, strict-matching-F1, and proportional matching F1. The system ranked 12th on Task-A using this overall average; however, the claude/o1 model performed much better on the Strict-matching category than the Proportional-matching. As shown in Figure 3, the system ranks 6th in the top-list of 23 systems for Strict-matching F1 (0.2092). Additionally, the system ranks **1st** out of 23 top systems on the **Strict-matching Precision** (**0.2267**). Interestingly, the highest Strict-matching Recall was achieved by the 10th system in this rank, the MediFact team, with score 0.3143 (bolded). For Task-B (summarisation), there are two aspect evaluations, Relevance and Factuality. Relevance score is averaged from automatic metrics of ROUGE, BERTscore, METEOR, and BLEU, which are originally machine translation (MT) evaluation metrics. For Factuality, there are the AlignScore and SummaC scores. Our system performed much better on the Factuality aspect in this task, espacially, in the **AlignScore** where we ranked the *second* with *0.4775* out of all top systems, and resulted as the 6th with overall Factuality score 0.3545 among the top 10, as in Figure 4.

4.2 Cost-Effectiveness Comparisons

Interestingly, the competition revealed some unexpected insights about the nature of the task itself. The baseline model, based on the Flan-T5 archi-

Final Ranking	Team	Submission Name	STRICT_MATCHING_P	STRICT_MATCHING_R	STRICT_MATCHING_F1
3	yxyx	sonnet	0.2205	0.2781	**0.2460**
5	KHU_LDI	0204_3	0.1868	0.3010	0.2305
13	NU-WAVE	k16	0.2048	0.2286	0.2160
14	Roux-lette	aa_version_3	0.2048	0.2286	0.2160
4	AICOE	submission_7	0.1765	0.2743	0.2148
15	The Manchester Bees	claude/o1	**0.2267**	0.1943	0.2092
6	LTRC@PerAnsSumm2025	submission-6	0.1915	0.2229	0.2060
2	YALENLP	250202_v3	0.1571	0.2857	0.2027
1	WisPerMed	WisPerMed-Finale	0.1726	0.2305	0.1974
12	MediFact	3	0.1383	**0.3143**	0.1921

Figure 3: Strict Matching Ranking on Task-A (Span Identification and Classification): the top 10 systems (highest score **bolded**, ours underlined)

Final Ranking	team	Submission Name	AlignScore	SummaC	TASK_B_FACTUALITY
11	HSE NLP	4o Mini NER	**0.5150**	0.2578	**0.3864**
8	Team Airi	Mistral + Lora	0.4728	0.2872	0.3800
3	yxyx	sonnet	0.4601	0.2834	0.3717
9	DataHacks	better_256	0.4427	*0.2899*	0.3663
10	UTSA-NLP	TrailNo6COT	0.4503	0.2620	0.3562
15	The Manchester Bees	claude/o1	*0.4775*	0.2316	0.3545
1	WisPerMed	WisPerMed-Finale	0.4085	**0.2958**	0.3521
20	TrofimovaMC	s_03	0.4679	0.2304	0.3491
4	AICOE	submission_7	0.4260	0.2701	0.3480
6	LTRC@PerAnsSumm2025	submission-6	0.4184	0.2701	0.3442

Figure 4: Task-B (Summarisation) Factuality Ranking: the top 10 systems (highest score **bolded**, second highest *italic*, ours underlined). This approach ranked are the 2nd highest in AlignScore.

tecture, established a foundation for comparison, though with performance metrics that left considerable room for improvement in this specialized task (Naik et al., 2024; Chung et al., 2024). This created an unusual situation where our model actually needed to "calibrate down" its responses to better match the expected output quality. This observation raises important questions about evaluation metrics and the balance between output quality and adherence to training data patterns.

The final results demonstrated the power of ISP: achieving top 15 placement out of 23 systems in the leaderboard (155 submissions overall) without any model training, using only prompt engineering and clever problem decomposition. This success challenges traditional assumptions about the necessity of model fine-tuning for competitive performance in specialized tasks. The entire process, from initial prompt generation to final submission, required only 3-4 hours of development time, showcasing the efficiency of the approach.

The implications of this success extend beyond the specific competition. It demonstrates that with well-crafted prompts and strategic task decomposition, existing language models can achieve competitive performance on specialized tasks without the need for additional training or fine-tuning. This suggests a promising direction for rapid development of AI solutions, particularly in domains where development time and computational resources are

Metric	Traditional Approach	ISP
Model Training	Hours/Days	None
Compute Resources	High	Minimal
Development Time	Days	3-4 Hours

Table 1: Comparison Between Traditional Approach and ISP Methods for Healthcare Summarization Tasks.

limited.

5 Discussion and Examples

5.1 On the dataset

There are some responses/questions that are just as funny or strange, which might affect the quality of the training data, but also may be true in the style of the online community forum as where the original data were extracted. Here are some examples:

- Unconventional medical category: "question": "Do women in the same house get period at the same time?"
- Not-really healthcare: "question": "Is there a way to make my voice deeper?" ⇒ "answers": ["You can modify your technique of speaking to include a deeper tone. Most people speak from the front of their mouth, ... "]
- Spelling and grammar: "txt": "nd, but these herbal remedies on the extremely rare occaission that they do work to help your bust, the results are only temporary."
- Not-meaningful: "question": "How thin is too thin?" ⇒ "SUGGESTION_SUMMARY": "To determine if your weight is too low, use the BMI chart. It is also advised to release not all guys want skin and bones."

5.2 On system rankings and metrics

It is interesting to see so many metrics reported in the overall categories and subcategories for Task A and B in the official evaluation (Agarwal et al., 2025). However, observations reveal that the metrics and ranking results do not always agree with each other, spacially, between tasks (A vs B). For instances, among our three submissions (v1, v2, v3), even though our system-v1 achieved the highest Task A + B combined average score (0.3993) in comparison to the other two systems (0.3928 and 0.3496), system-v2 and v3 have produced better scores for individual metrics and tasks, respectively.

As in Figure 5, for Task A (span identification and classification), our **system 2** produced **better scores on macro F1, weighted F1, and strict matching precision**, in comparison to the version 1 system. However, it lost to the strict matching recall value, leading to a lower strict matching F1.

For Task B (summarisation) Factuality ranking, our **system 3** boosted both AlignScore and SummaC scores, leading to the **highest Factuality score** (**0.4277**) among the top 10 systems in the leader board as in Figure 6, referring to Figure 4 for the top 10 (highest Factuality score 0.3864).

6 Conclusions and Future Work

In conclusion, we submitted three system outputs using the method Iterative Self-Prompting (ISP) with Calude and o1, ISP-claude/o1, to perspective-aware healthcare answer summarisation shared task (PerAnsSumm2025). The vesion 1 output of ISP-claude/o1 is officially ranked 15th in the leaderboard of top 23 teams, using the combined average scores of Task A and B. Task specifically, the ISP-claude/o1 performs better on Strict-matching for Task A (the 6th in Figure 3), span-identification and classification, versus proportial-matching. For Task B summarisation, it performs better on AlignScore for Factuality (the 1st via ISP-claude/o1-system3, 0.4277 in Figure 6), instead of Relevance (ROUGE, BERTscore, METEOR, and BLEU, much lower scores). In the future work, it is worthy to explore the reasons on such contradiction scores across metrics, i.e., Strict-matching vs Proportional-matching, and Relevance vs Factuality. Our complete implementation is available as an open-source project on https://github.com/pabloRom2004/-PerAnsSumm-2025

Limitations

The present study faced several constraints that suggest directions for future research. Due to time limitations, only decoder models employing prompting techniques were evaluated in this shared task. For a more comprehensive analysis, future work

claude/o1	macro F1	CLASSIFICATION_Weighted_F1	STRICT_MATCHING_P	STRICT_MATCHING_R	STRICT_MATCHING_F1
v1	0.8268	0.8769	0.2267	**0.1943**	**0.2092**
v2	**0.8664**	**0.9031**	**0.2327**	0.1733	0.1987
v3	0.6760	0.7581	0.1526	0.0724	0.0982

Figure 5: The Manchester Bees 3 systems comparisons on Task A

Team	Submission Name	AlignScore	SummaC	TASK_B_FACTUALITY
HSE NLP	4o Mini NER	*0.5150*	0.2578	*0.3864*
DataHacks	better_256	0.4427	*0.2899*	0.3663
The Manchester Bees	claude/o1-v1	0.4775	0.2316	0.3545
The Manchester Bees	claude/o1-v2	0.4119	0.2291	0.3205
The Manchester Bees	claude/o1-v3	**0.5888**	0.2666	**0.4277**
WisPerMed	WisPerMed-Finale	0.4085	**0.2958**	0.3521

Figure 6: Task-B (Summarisation) Factuality Ranking: including three systems of our submissions, keeping the highest and the 2nd highest scores in the top-10 list (highest score **bolded**, second highest *italic*, ours underlined). Our system 3 (claude/o1-v3) gets the highest in AlignScore and Factuality.

should include comparisons with traditional fine-tuned approaches, particularly encoder-decoder architectures such as T5-variants for span detection tasks. Such comparisons would provide valuable benchmarks against established methodologies in the literature (Belkadi et al., 2023; Cui et al., 2023).

A significant technical challenge encountered during the ISP-claude/o1 implementation involved context window limitations of the models. This necessitated dividing the test dataset into smaller chunks for processing. Further research could explore efficient solutions to these context constraints, potentially through advanced chunking strategies or more context-efficient prompting techniques.

While fine-tuning smaller models represents a potentially more cost-effective approach for production deployment, the ISP method demonstrated distinct advantages in rapid development scenarios. The implementation required only 3-4 hours without GPU training resources, model optimization, or hyperparameter tuning. This approach prioritized development efficiency and exploration of state-of-the-art models' few-shot learning capabilities, though future work could investigate quantized versions of fine-tuned models for production environments with comparable performance at reduced computational cost.

Ethical Statement

While the ISP method demonstrates its effectiveness in the summary task of healthcare with perspectives, there are many concerns about the use of commercial chatbots, for example ChatGPT, Claude, etc. for personal data (Ray, 2023; Ren et al., 2024). It is still challenging on how to safeguard private health information with the usage of AI models. For the current shared task, the organisers have prepared annonymised online forum data for system development purposes.

Acknowledgments

LH and GN are grateful for the grant "Integrating hospital outpatient letters into the healthcare data space" (EP/V047949/1; funder: UKRI/EPSRC). LH is grateful to the 4D Picture EU Project https://4dpicture.eu.

References

Siddhant Agarwal, Md Shad Akhtar, and Shweta Yadav. 2025. Overview of the peranssumm 2025 shared task on perspective-aware healthcare answer summarization. In *Proceedings of the Second Workshop on*

Patient-Oriented Language Processing (CL4Health) @ NAACL 2025, Albuquerque, USA. Association for Computational Linguistics.

Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. 2023. Generating medical instructions with conditional transformer. In *SyntheticData4ML Workshop at NeurIPS 2023*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Yang Cui, Lifeng Han, and Goran Nenadic. 2023. MedTem2.0: Prompt-based temporal classification of treatment events from discharge summaries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 160–183, Toronto, Canada. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial intelligence in medicine*, 67:25–37.

Gauri Naik, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. 2024. No perspective, no perception!! perspective-aware healthcare answer summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15919–15932, Bangkok, Thailand. Association for Computational Linguistics.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Libo Ren, Samuel Belkadi, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. 2024. Synthetic4health: Generating annotated synthetic clinical letters. *Preprint*, arXiv:2409.09501.

Abeed Sarker. 2014. *Automated Medical Text Summarisation to Support Evidence-based Medicine*. Ph.D. thesis, Macquarie University, Centre for Language Technology, Department of Computing.

Thomas Searle, Zina Ibrahim, James Teo, and Richard JB Dobson. 2023. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. *Journal of Biomedical Informatics*, 141:104358.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

Final Ranking	Team Name	Submission Name	Task A + B Combined Average
1	WisPerMed	WisPerMed-Finale	0.4571
2	YALENLP	250202_v3	0.4548
3	yxyx	sonnet	0.4526
4	AICOE	submission_7	0.4495
5	KHU_LDI	0204_3	0.4492
6	LTRC@PerAnsSumm2025	submission-6	0.4395
7	MNLP	v3_4	0.4321
8	Team Airi	Mistral + Lora	0.4238
9	DataHacks	better_256	0.4203
10	UTSA-NLP	TrailNo6COT	0.4112
11	HSE NLP	4o Mini NER	0.4081
12	MediFact	3	0.4077
13	NU-WAVE	k16	0.4046
14	Roux-lette	aa_version_3_20250204_004205	0.3996
15	**The Manchester Bees**	**claude/o1**	**0.3994**
16	Abdelmalak	sub2	0.3907
17	umb	umba	0.3824
18	massU	1	0.3815
19	RVK_Med	Run_1	0.3750
20	TrofimovaMC	s_03	0.3698
21	TeamENSAK@PerAnsSumm2025	Azzedine	0.3641
22	CaresAI	submission_1	0.3405
23	LMU	llama 70b_8b	0.1726

Figure 7: Official Ranking Task A+B from Top 23 Systems (Agarwal et al., 2025)

A The Official Ranking

B The original prompt

Here is the original prompt describing the task:

(Examples from the test set here in-context)

" *Could you write me a prompt that takes a test set answer and provides the format that is expected in the output, could you look very carefully at how the spans are structured and what the labels are/what they represent in this specific database and be able to detect spans and create reasonable spans and summarization. Make sure to look very closely at the data I have provided and come up with a good prompt that captures the essence of each label and how to pick it up accordingly, this prompt will be used for another model with no previous knowledge about the task so you will need to make sure you explain it all thoroughly*

Before completing the task, just talk out loud about the task and how you will complete it, and ask me any questions you may have before writing this prompt, this prompt will just be the first version, I will give you more examples so you are able to refine it more and I will test it with the model and bring back the results so you can tweak the prompt to see better behaviour, I will give you the original input with the prompt you will create, then give you the model output along with the ground truth so you are able to tweak it."

Detailed ISP used for this task is shared on our open-source project page https://github.com/pabloRom2004/-PerAnsSumm-2025