LLM-Based Clinical Note Summarization Feasibility Study

Prakash Hampole, BMSc[1,2]; Dan Nguyen, PhD[1,2]; Michael Dohopolski, MD[1,2]; Ti Bai, PhD[1,2]; Steve B. Jiang, PhD[1,2]
[1]Medical Artificial Intelligence and Automation (MAIA) Lab
[2]Department of Radiation Oncology, UT Southwestern Medical Center

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

MAIA Laboratory
UT Southwestern Medical Center
Radiation Oncology

ABSTRACT

The extraction of clinically relevant details from detailed patient notes is essential for cancer treatment planning. As clinical documentation grows in volume and complexity, this review can become increasingly time-consuming and cognitively demanding. Large language model (LLM)-based multi-agent workflows offer a promising approach for streamlining information extraction. However, performance depends on design choices such as reflective reasoning and prompt configuration, and hallucinations remain unacceptable in clinical environments. This feasibility study evaluates the ability of a multi-agent system to generate accurate, evidence-grounded clinical summaries for cancer treatment planning, using adaptive radiotherapy (ART) clinical notes as a representative use case.

Twenty-one ART clinical notes written by radiation oncologists were used. Summaries were generated by an LLM and edited by a human reviewer to form the ground truth dataset. A four-agent pipeline was implemented: Agent 1 generated an initial summary; Agent 2 evaluated it against the ground truth; Agent 3 revised it using reflective reasoning; and Agent 4 evaluated the revised output. Evaluation agents classified each summary sentence into a confusion matrix based on agreement with the ground truth, from which precision, recall, and F1-score were computed. For 0-shot and 3-shot prompting, mean F1-scores before reflection were 85.3% and 86.5%, respectively. After reflection, mean F1-scores were 85.1% and 86.9%, respectively.

For clinical note summarization, multi-agent LLM performance improved modestly with 3-shot prompting and reflective reasoning, with the highest F1-score achieved using both. However, reflection lowered some metrics, possibly by introducing unnecessary edits. Limitations include dataset size and reflection prompt quality, which future work may address. Overall, these results demonstrate the practical potential of evidence-grounded LLM summaries to support cancer treatment planning and motivate evaluation in larger, clinician-reviewed datasets and broader clinical contexts.

CONTACT

Prakash Hampole
UT Southwestern Medical Center
Prakash.Hampole@utsouthwestern.edu

INTRODUCTION

- Cancer treatment planning often requires review of lengthy unstructured clinical notes to extract relevant details.
- These include diagnosis, prior irradiation, intent, and performance status.
- As clinical documentation expands, this review task becomes a bottleneck in time-sensitive treatment planning workflows.[1]
- Large language model (LLM)-based summarization has shown promise for condensing clinical text.[2]
- Thus, LLMs may offer a scalable approach for streamlining information retrieval directly at the point of care.
- The output quality of such an approach would be dependent on design choices such as the inclusion of reflective reasoning and in-context examples ("shots").[3]
- However, medical use requires careful attention to source grounding and hallucination risk.[4]
- This feasibility study used adaptive radiotherapy (ART) notes as a representative treatment-planning use case to test whether a multi-agent GPT-4o workflow could generate structured, evidence-grounded summaries.

METHODS

- Dataset: Twenty-one ART clinical notes were written by radiation oncologists.
- Ground truth: GPT-4o generated initial summaries, which were edited by a human reviewer to create the reference set.
- Summary format: Each summary was a JSON object with fields for diagnosis, prior irradiation, treatment intent, and Eastern Cooperative Oncology Group (ECOG) performance status. Each field included supporting evidence from the source note.
- Pipeline (Figure 1): Agent 1 generated Summary 1; Agent 2 evaluated Summary 1 against the ground truth; Agent 3 revised Summary 1 using reflective reasoning; and Agent 4 evaluated Summary 2.
- Prompting: 0-shot and 3-shot prompts were tested with and without reflection, yielding four configurations per note.
- Evaluation: Summary sentences were classified into a confusion matrix based on agreement with ground truth. Precision, recall, and F1-score were then calculated.

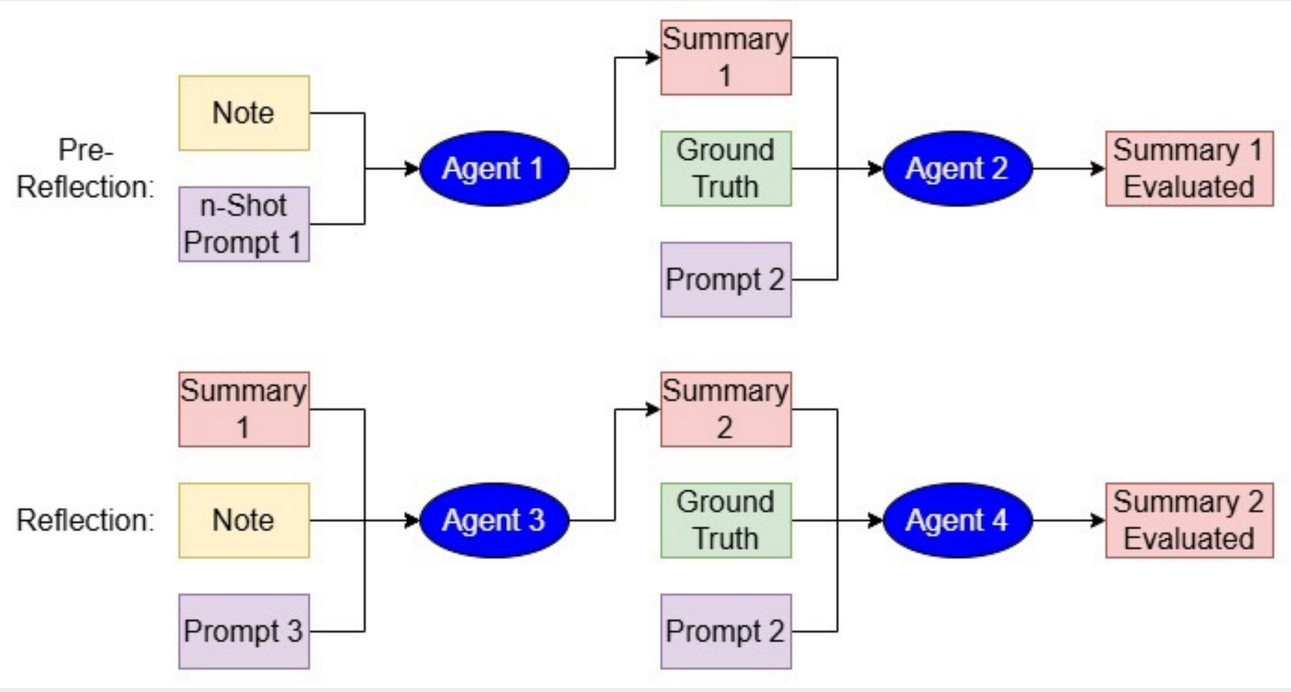

Figure 1. Four-agent pipeline. The clinical note and a 0- or 3-shot summarization prompt are inputted to Agent 1, which produces Summary 1. Alongside the ground truth and an evaluation prompt, this is inputted to Agent 2, which evaluates it. Alongside the note and a reflection prompt, Summary 1 is also inputted to Agent 3, which produces a revised summary, Summary 2. Alongside the ground truth and evaluation prompt, this is inputted to Agent 4, which evaluates it.

RESULTS

- Performance was usually higher with 3-shot prompting than with 0-shot prompting.
- However, without reflection, recall was lower with the 3-shot prompt.
- Performance was usually higher with reflection.
- However, precision was lower with reflection.
- The highest F1-score achieved was 86.9%, from 3-shot prompting with reflection.

Table 1. The mean precision of the summaries for each number of in-context examples, without and with reflection.

Mean Recall	Without Reflection (%)	With Reflection (%)
0-Shot	77.6	77.1
3-Shot	80.2	78.2

Table 2. The mean recall of the summaries for each number of in-context examples, without and with reflection.

Mean Recall	Without Reflection (%)	With Reflection (%)
0-Shot	97.7	98.6
3-Shot	96.7	99.5

Table 3. The mean F1-score of the summaries for each number of in-context examples, without and with reflection.

Mean F1-Score	Without Reflection (%)	With Reflection (%)
0-Shot	85.3	85.1
3-Shot	86.5	86.9

DISCUSSION

Interpretation:

- Additional in-context examples likely helped constrain the desired summary structure, contributing to the strongest F1-score with 3-shot prompting.
- Reflection may have helped recover missed information, as reflected by improved recall, but it could also have introduced unnecessary changes that reduce precision.
- The observed gains were modest, which is appropriate for a feasibility study and suggests that prompt and workflow design remain important.

Limitations:

- A small dataset was used (21 notes).
- Ground truth summaries were generated by an LLM and then human-edited, rather than independently authored by clinicians.
- Performance with reflection depended on the reflection prompt quality.
- Similarly, performance with 3-shot prompting depended on the quality of the in-context examples provided.
- All results were based on the GPT-4o deployment.

Future Work:

- Reproduce the study with a larger dataset and clinician-adjudicated reference summaries.
- Optimize reflection prompts and in-context examples.
- Compare additional LLMs and test multiple rounds of reflection.
- Evaluate usability in realistic treatment-planning workflows.

CONCLUSIONS

- This feasibility study supports the practicality of LLM-generated, evidence-grounded summaries for treatment-planning information extraction.
- In this dataset, 3-shot prompting with reflection achieved the highest F1-score, although improvements were modest and showed precision–recall tradeoffs.
- Larger clinician-reviewed studies are needed before clinical implementation.

REFERENCES

1. Wiest IC, Wolf F, Leßmann M-E, et al. A software pipeline for medical information extraction with large language models, open source and suitable for oncology. npj Precision Oncology. 2025;9(1). doi:10.1038/s41698-025-01103-4.
2. Bednarczyk L, Reichenpfader D, Gaudet-Blavignac C, et al. Scientific evidence for clinical text summarization using large language models: Scoping review. Journal of Medical Internet Research. 2025;27. doi:10.2196/68998.
3. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. arXiv. 2020;(2005). https://arxiv.org/abs/2005.14165.
4. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine. 2025;8(1). doi:10.1038/s41746-025-01670-7