Towards a Real-time Clinical Agenda Setting System for Enhancing Clinical Interactions in Primary Care Visits

Kuk Jin Jang,[*1, 2] **Sameer A. Bhatti,**[*1] **Sydney Pugh,**[*2]
Chimezie Maduno,[1] **Sarang Sridhar,**[1] **Sriharsha Mopidevi,**[2] **Eric Eaton,**[1] **Kevin B. Johnson**[1,2]

[1] Department of Computer and Information Science, University of Pennsylvania
[2] Perelman School of Medicine, University of Pennsylvania
{jangkj, sabhatti, cjmaduno, saresri, eeaton}@seas.upenn.edu,
{sydney.pugh, sriharsha.mopidevi, kevin.johnson1}@pennmedicine.upenn.edu

Abstract

Technology has increasingly hindered meaningful engagement between patient and providers during primary care visits, often detracting from effective communication. However, artificial intelligence (AI) advancements present new opportunities to enhance and improve patient-provider communication. A promising application is the use of AI to identify and highlight agenda items for discussion during visits and to summarize relevant clinical details in real-time. This study explores the feasibility, potential, and challenges of developing a real-time automated agenda-setting system leveraging generative AI, specifically large language models (LLMs). From a dataset of recorded and annotated simulation visits, we evaluate the performance of LLMs in identifying agenda items and capturing associated clinical details within the conversation flow. In particular, we focus on the impact of real-time constraints and contextual factors on the ability to detect and summarize relevant items. Our findings suggest that optimizing performance requires a balance between providing contextual information through both summaries and the actual conversation. Based on these results, we discuss the challenges involved in developing a real-time agenda-setting system and offer recommendations for future advancements.

1 Introduction

Effective communication in healthcare is crucial for delivering quality care. Research has shown that communication attributes significantly influence interaction quality and clinical outcomes; for instance, negative provider discourse has been associated with lower post-visit medication adherence in young patients (Glenn et al. 2021). Despite the importance, a significant number of patient and clinician questions remain unanswered during a typical primary care visit (Del Fiol, Workman, and Gorman 2014; Ely et al. 2007; Hood-Medland et al. 2021). Furthermore, agenda setting—a structured approach to addressing patient concerns—has been shown to improve visit outcomes without adversely affecting the overall experience, making it a valuable tool for enhancing engagement and satisfaction (Singh Ospina et al. 2019).

However, several challenges impede effective communication. The increasing use of technology in clinical settings, while intended to streamline processes, often detracts from clinician engagement during patient interactions. Poor communication resulting from technology distractions has been well-documented, highlighting the need for solutions that support, rather than hinder, meaningful exchanges (Liu et al. 2024).

Recent advancements in artificial intelligence (AI) offer promising approaches to address these challenges. Automated tools and systems for analyzing clinical interactions hold significant potential for identifying and mitigating barriers to effective communication. In particular, large language models (LLMs) have opened new opportunities to explore clinical interactions in depth, enabling the development of previously infeasible systems. For instance, AI-driven technologies like Nuance DAX (Nuance Communications 2023) have emerged to assist with post-visit clinical documentation. However, these systems primarily focus on summarization after the visit, leaving an unmet need for real-time solutions that enhance engagement during the interaction.

To address this gap, this study explores the feasibility and challenges of leveraging generative AI for real-time clinical agenda setting and summarization. Agenda setting, a critical component of effective communication, plays a key role in ensuring that both patient and clinician concerns are systematically addressed during a visit. A real-time system capable of keeping track of key issues, facilitating engagement, and ensuring comprehensive documentation could transform patient-clinician interactions. This work evaluates the performance and limitations of utilizing state-of-the-art (SOTA) generative AI-based approaches for agenda-setting. Our contributions are as follows:

1. We define the agenda-setting problem in a clinical interaction and propose a system that uses generative AI to address the issue
2. We curate a dataset of simulated clinical visits and annotate the conversations for evaluation in real-time agenda setting.
3. Through our quantitative evaluation and qualitative analysis, we highlight the potential for generative AI's use in

[*]These authors contributed equally.

real-time agenda setting and the challenges that must be addressed.

2 Related Work

2.1 AI in Healthcare Communication

Recent advancements in clinical natural language processing (NLP) have shown significant potential for handling medical text through approaches such as training new models (Singhal et al. 2022), (Tu et al. 2023), fine-tuning pre-existing ones (Toma et al. 2023), (Veen et al. 2023), or integrating task-specific examples into model prompts [(Mathur et al. 2023), (Veen et al. 2023)]. (Van Veen et al. 2024) applies various adaptation techniques to eight open-source and proprietary Large Language Models (LLM) across four summarization tasks using six datasets. The findings highlight the substantial benefits of model adaptation over zero-shot prompting while also exploring trade-offs related to model size, novelty, and domain specificity. Additionally, the findings suggest the clinical summaries from adapted LLMs can be preferred over the summaries from medical experts.

With greater availability in patient-provider conversation datasets, there is also the opportunity to develop medical dialogue systems to act as virtual medical consultants to the physician. (He et al. 2022) develop a dataset and method for analyzing patient-provider conversations to recommend medications. (Xu et al. 2024) proposed a framework that generates a response that relies on abductive and deductive reasoning to align with the clinician's diagnostic reasoning process.

2.2 Technology in Clinical Communication

Given the significant amount of time clinicians spend on Electronic Health Record tasks (Arndt et al. 2017), several tools have been developed to aid in clinical documentation, also called ambient scribing. Tools like Dragon Ambient eXperience (DAX) and Abridge listen into the patient-provider conversation and generate structured clinical notes, alleviating the documentation burden on clinicians. These tools have been shown to present potential time-saving and prevent physician burnout (Liu et al. 2024). However, since these tools transcribe the conversation and create the note offline post-visit, there is no real-time tracking of questions and concerns raised by both clinician and patient. Therefore, if certain questions or issues are only mentioned and not discussed by the clinician or the patient, they will not be reported thoroughly in the clinical note.

This study aims to leverage LLMs in clinical interaction analysis for real-time summarization to address the above limitations. Real-time summarization can allow physicians to capture information without writing down or remembering all patient information. We explore LLMs' capabilities in capturing relevant clinical events, termed as agenda items and details, with various amounts of context to simulate an active conversation.

3 Problem Setting and Study Design

3.1 Real-time Clinical Agenda Setting

A clinical agenda item refers to a specific issue, concern, or goal that a patient or clinician identifies as a priority for discussion and resolution during a medical appointment. These items can include symptoms the patient is experiencing, questions about medications or treatment plans, or broader health goals such as weight management or mental health support. Properly identifying and addressing these agenda items ensures a focused and productive clinical encounter. Agenda setting in a clinical context is a collaborative effort between the patient and the clinician. It involves identifying key issues to be discussed, prioritizing them based on urgency or importance, and organizing the discussion points and action items in a structured manner.

To support the agenda-setting process, an automated system must meet specific functional and performance criteria:

- Accuracy: The system must accurately recognize and categorize agenda items discussed during a clinical encounter and associate them with pertinent clinical details, ensuring that no critical information is overlooked.
- Timeliness: Real-time systems must process information quickly to avoid disrupting the natural flow of conversation
- Robustness: Clinical conversations are non-linear and may include interruptions and topic shifts. The system must be able to identify proper agenda items and details within the dynamics of the conversation
- Context-awareness: Clinical dialogues frequently involve specialized medical terminology, implicit references, and context-dependent nuances. The system must interpret the context accurately to ensure meaningful contributions to the interaction and avoid misinterpretation of critical details.

This work explores the feasibility of developing a real-time agenda setting system based on generative AI that satisfies these requirements.

3.2 Key Questions

This study focuses on the following questions:

- What is the baseline performance of existing LLMs for identifying agenda items and details?
- What is the form of the context window that is required?

4 Experiments and Results

In order to assess the key questions, we curate a dataset of clinical conversations and design experiments which we describe in the following section.

4.1 Dataset and Metrics

Data source and preprocessing We use a dataset of 16 simulated patient-provider interaction videos. Each interaction features a medical provider's conversation with a standardized patient following a given case description describing the primary concerns, associated symptoms, family history, vital signs, and possible differential diagnoses. The 16

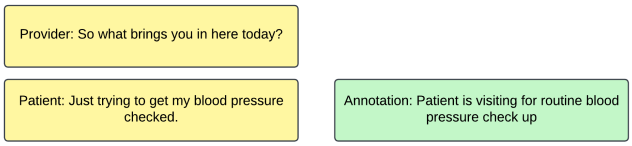

Figure 1: Example of an annotated transcript.

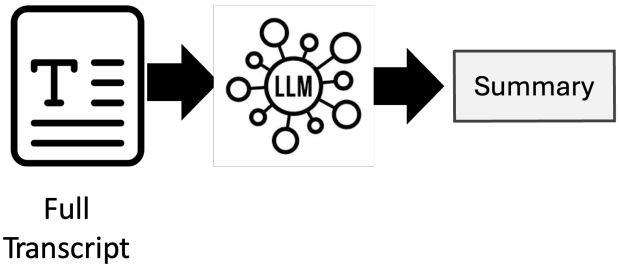

Figure 2: Baseline experiment. Full clinical conversation is input into LLM to generate a summary

simulated interactions had a total duration of about 300 minutes, with an average of 18.7 minutes per interaction.

The audio from each interaction video was transcribed using the automatic speech recognition (ASR) model WhisperX (Bain et al. 2023). The resulting transcript was then diarized with patient and provider speaker labels using GPT-4o (OpenAI 2023b).We prompt GPT to annotate the speaker for 50 lines of a transcript at a time.

Data annotation protocol We annotate the diarized patient-provider interaction transcripts for *agenda items* and other relevant *details* that a real-time agenda-setting system should detect. Agenda items are specific issues, concerns, or goals that a patient or provider identifies as a priority to address during the clinical visit. For example, if a patient expresses they have a cough and shortness of breath, the agenda item would be "Patient is experiencing cough with shortness of breath." Relevant details that could arise later in the interaction could include how long the symptoms have persisted or if the patient is a smoker.

Annotators would be given the case description for each transcript to provide all important information to look for in the transcript. Annotators would then read through the transcript for agenda items and details related to the case and summarize these in-line where the event came up. Annotators would label each annotation as an "agenda item" or "detail."

Several difficulties arose in this annotation process. First, details could be repeated several times in the conversation, making it difficult to decide whether to include redundant information. We made the decision to include annotations for the repeated information as this would still need to be captured in a real-time system. Second, it is often difficult to find the exact place in the transcript where the main issue for the patient is expressed i.e. the agenda item occurs in several lines. In these cases, the event was labeled as an agenda item only in the first instance it came up in the conversation since this would be considered the main issue of why patient came in.

A total of 688 agenda items and clinical details were annotated across the 16 clinical conversation recordings.

Evaluation metrics We evaluate the performance of detecting agenda items and details, as well as the quality of their summaries, across our various experiments.

Metrics for agenda and detail detection To assess our proposed system's ability to detect agenda items and relevant detailsin a clinical visit, we use precision and recall measures. Precision evaluates the likelihood that system detected agenda items and details are actually relevant (for agenda-setting), whereas recall measures how well the system detects all relevant events.

Suppose a conversation transcript consists of N lines. Let y_i represent the ground truth for line i, where $y_i = 1$ if an agenda *or* detail was annotated for line i, and $y_i = 0$ otherwise. Let $\hat{y}_i$ represent the system's detection output for line i, where $\hat{y}_i = 1$ if the system detects an event, and $\hat{y}_i = 0$ otherwise. We compute precision and recall as follows,

$$\text{Precision} = \frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1 \text{ and } y_i = 1)}{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1)} \quad (1)$$

$$\text{Recall} = \frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1 \text{ and } y_i = 1)}{\sum_{i=1}^{N} \mathbf{1}(y_i = 1)} \quad (2)$$

Quantitative metrics for summarization quality To evaluate the quality of the summaries of agenda items and details, we employed several well-established quantitative metrics:

- **ROUGE-L**: Rouge-L(Lin 2004) is a recall-oriented metric that looks for the longest common subsequence between the reference and the candidate.
- **BLEU**: The BLEU (BiLingual Evaluation Understudy) metric is a metric that was originally developed for the automatic quality evaluation of machine-translated texts. The BLEU metric is a corpus-level metric based on the modified n-gram precision measure with a length penalization for the candidate sentences that are shorter than the reference ones. (Papineni et al. 2002)
- **BERTScore**: BERTScore(Zhang et al. 2020), leverages contextual BERT embeddings to evaluate the semantic similarity of the generated and reference texts.
- **SemScore**: SemScore (Aynetdinov and Akbik 2024): SemScore is an evaluation metric for assessing LLM outputs by measuring semantic similarity to reference responses, offering a closer alignment with human judgment compared to traditional metrics like BLEU or ROUGE.

4.2 Experiment: Baseline LLM performance

To establish a baseline for the performance of an agenda-setting LLM, we prompt a model to summarize the agenda items and details using an entire diarized patient-provider

Model	Rouge-L	BLEU	BERTScore	SemScore
Llama 2	7.87 ± 7.28	0.81 ± 1.89	76.88 ± 3.13	31.26 ± 29.27
Llama 3	23.57 ± 4.83	6.99 ± 2.70	**86.74 ± 1.11**	**83.65 ± 3.98**
Vicuna	23.01 ± 4.98	4.51 ± 3.87	85.86 ± 1.28	82.19 ± 7.28
GPT 3.5 Turbo	**28.89 ± 5.39**	**8.79 ± 3.20**	85.88 ± 1.04	80.73 ± 4.45

Table 1: Summarization scores for baseline agenda-setting LLM that receives the entire transcript as input. We denote the best scores in bold and the second best scores are underlined.

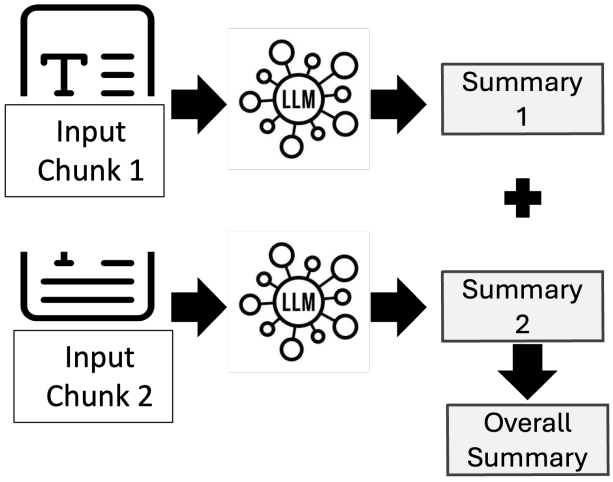

Figure 3: Diagram of input lines experiment.

interaction transcript as input. We conduct this experiment using models Llama 2 and 3 (Touvron et al. 2023), Vicuna (Team 2023), and GPT 3.5 Turbo (OpenAI 2023a). Table 1 reports the average scores for the aforementioned summarization quality metrics across all 16 patient-provider interactions. We observe that GPT 3.5 Turbo achieved the highest ROGUE and BLEU scores which indicates the model has better ability to summarize agenda items and details similarly to the annotators. Llama 2 had the highest BERTScore and SemScore suggesting this model's summaries captured more semantically meaningful and relevant information for agenda-setting. These results suggest that more advanced models, such as GPT 3.5 Turbo and Llama 3, are better suited for agenda-setting tasks.

4.3 Experiment: Varying number of input lines

Input Lines	Rouge-L	BLEU	BERTScore	Semscore
2	27.99 ± 4.49	8.00 ± 2.69	84.81 ± 0.75	79.38 ± 6.03
5	**29.59 ± 4.88**	**8.82 ± 3.19**	**85.32 ± 0.86**	82.00 ± 4.91
10	28.60 ± 3.31	7.69 ± 2.05	85.17 ± 0.77	**83.38 ± 3.40**
20	27.80 ± 4.36	8.20 ± 2.28	85.09 ± 1.11	82.29 ± 6.19

Table 2: Summarization scores for agenda-setting LLM that receives a fixed number of transcript lines at a time. We denote the best scores in bold and the second best scores are underlined.

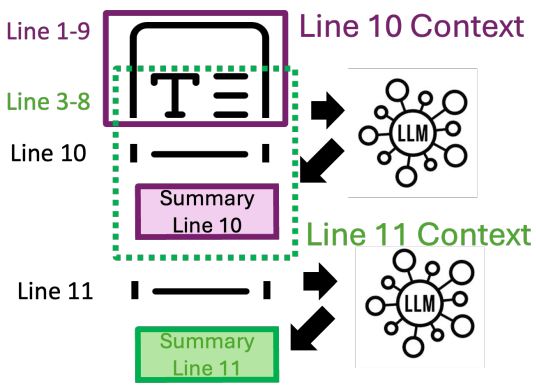

Figure 4: Diagram of real-time simulation experiment.

In a real-time setting, an agenda-setting LLM would receive the transcript of a patient-provider interaction incrementally, rather than all at once. In this experiment, we assess the impact of the number of transcript lines provided to the LLM at a time on agenda-setting performance. We use the GPT 3.5 Turbo model to summarize agenda items and details from each fixed-size chunk of lines from the transcript. Each chunk input and the corresponding summary output are kept in the model's context window. We conduct our experiment by providing 2, 5, 10, and 20 lines at a time. The experiment results are presented in Table 2. The results indicate that the best-quality summaries are generated when 5 lines are provided. Providing fewer lines may not provide enough context for the conversation, while more lines may overwhelm the model's capacity to summarize the high-priority agenda items and details within the chunk.

4.4 Experiment: Real-time simulation

To evaluate the feasibility of a real-time agenda-setting LLM we conduct an experiment where the model processes transcripts line-by-line, simulating the flow of a live clinical visit conversation. The real-time simulation experiment is depicted in Figure 4. We initialize an LLM with a system prompt describing the type of input it will receive and instructions for the agenda-setting task. The system prompt is as follows.

> You are a clinical agenda-setting assistant. You will receive a transcript of a clinical visit, provided line by line. Each line begins with a speaker tag, either "[Provider]" or "[Patient]". Your task is to summarize all clinically relevant details mentioned in each line in a single concise sentence. Use the context of previous lines to understand the conversation when needed. If the line is spoken by the provider and contains a question (e.g., "[Provider] How bad is the pain from 1 to 10?"), respond with "None." Wait for the corresponding patient response (e.g., "[Patient] It's like a 9.") before summarizing any details. If a line does not mention any clinically relevant details, respond with "None."

Following the system prompt, the LLM receives each line of

Model	Context Size	Precision	Recall	Rouge-L	BLEU	BERTScore	SemScore
Llama 3.1 8B	0	27.79 ± 7.93	46.21 ± 13.49	29.9 ± 4.93	4.6 ± 1.95	**86.02 ± 0.79**	79.52 ± 4.4
	1	**35.7 ± 9.84**	68.61 ± 12.54	**30.61 ± 4.96**	**5.58 ± 1.99**	85.85 ± 0.6	**79.68 ± 4.56**
	20	25.55 ± 8.2	91.45 ± 12.72	20.4 ± 5.01	3.38 ± 1.27	83.64 ± 1.04	75.73 ± 4.69
	50	24.66 ± 7.42	97.13 ± 3.98	18.67 ± 5.09	3.0 ± 1.06	83.58 ± 1.29	75.41 ± 5.62
	100	23.31 ± 7.3	95.9 ± 10.81	17.39 ± 4.72	2.66 ± 1.01	83.43 ± 1.45	75.21 ± 5.0
	max	23.88 ± 7.24	**98.07 ± 2.94**	17.63 ± 4.43	2.71 ± 0.94	83.49 ± 1.21	75.38 ± 5.99
GPT 3.5 Turbo	0	**39.31 ± 10.16**	48.33 ± 13.5	**28.67 ± 4.32**	**5.68 ± 2.12**	**86.56 ± 0.93**	**77.57 ± 6.2**
	1	35.72 ± 11.28	59.43 ± 11.93	25.35 ± 4.8	4.76 ± 1.75	86.12 ± 1.2	77.36 ± 6.32
	20	25.59 ± 7.43	**92.64 ± 9.79**	17.18 ± 4.38	2.88 ± 1.09	83.64 ± 1.6	70.17 ± 8.17
	50	26.98 ± 11.87	90.49 ± 11.02	18.13 ± 6.93	3.11 ± 1.81	83.64 ± 1.54	71.7 ± 8.66
	100	25.91 ± 7.22	88.26 ± 14.84	18.05 ± 4.81	2.91 ± 1.15	83.64 ± 1.81	70.92 ± 8.72
	max	26.39 ± 8.98	89.2 ± 14.95	19.86 ± 5.3	3.45 ± 1.52	83.86 ± 1.98	71.01 ± 8.43

Table 3: Real-time agenda-setting simulation performance scores. We denote the best metric scores for each large language model (LLM) in bold. We denote the best scores in bold and the second best scores are underlined for each LLM.

the transcript and responds with a summary of any clinically relevant details. As our prior experiments suggest, providing previous lines for context can help improve the quality of the LLM's summaries. Therefore, we store each line and corresponding summary in the context window until the maximum context window size is reached. When the size exceeds the limit, the earliest pairs of transcript lines and corresponding summaries are removed until the context fits within the limits.

We conduct our experiment using models Llama 3.1 with 8 billion parameters and GPT 3.5 Turbo. For both models, we set the temperature hyperparameter to zero to receive more focused and deterministic responses. We try 0, 1, 20, 50, and 100 for the maximum context size. Typically, the length of the context window is determined in tokens rather than number of messages (lines) and responses (summaries).[1] Hence we also try including the maximum possible previous lines and summaries that fits the context window in terms of tokens.

Table 3 presents the results of the experiment. Llama 3.1 performs best with a context size of 1 while GPT 3.5 Turbo performs best with no context (size 0). These findings suggest that Llama 3.1 benefits from incorporating a minimal amount of context, while GPT 3.5 Turbo achieves its best performance when it processes each line of the transcript independently. Additionally, the results demonstrate a trade-off between precision and recall as the context size increases. For both models, larger contexts tend to improve recall, but this comes at the expense of precision. This implies that while providing more context may help the LLM capture more agenda items and relevant details from the transcript, it can also introduce significant noise into the summarization.

4.5 Experiment: Real-time simulation with context aggregation

Our experiments have suggested that for an agenda-setting LLM, providing a few lines from the transcript at a time and maintaining a small context window enhances the model's ability to capture relevant agenda items and details while minimizing noisy detections. Based on these findings, our final experiment combines these ideas with two real-time simulations.

The first simulation aims to limit the information provided as context. Instead of maintaining each transcript line and corresponding summary in the context window, which can potentially overwhelm the model with irrelevant information, we use a context comprising the last K summaries. We modify the real-time simulation from Section 4.4 to maintain only a single context summary at all times. After processing every K line, the context summary is updated using one of two aggregation strategies. The first method replaces the current context summary with concatenating the last K LLM-generated summaries. We refer to this aggregation strategy as a sliding window because each context summary covers disjoint chunks of the transcript. The second method appends the concatenation of the last K summaries to the current context summary. This is the growing window strategy because each context summary covers an increasing proportion of the transcript. We run this experiment using GPT 3.5 Turbo with (input size of 1) context sizes of 20 and 50.

The second simulation aims to combine findings from the first simulation and also utilize results from Section 4.3 by adding additional input lines from the conversation. Given the summarization scores of the various input lines in Table 2, we use the growing window strategy with an input and context size of 5 lines.

The experiment results are presented in Table 4. Overall, context aggregation yields significant improvements in precision scores and marginal improvements in the summarization metrics. This result suggests that the context summaries were useful for the LLM in having more accurate detections while maintaining the quality of the resulting summary. The results of the second simulation demonstrate the best performance in terms of balancing precision and recall. Further evaluations will be needed to determine the optimal set of input size and context size. Moreover, for real-time sys-

[1]The maximum context window length for Llama 3.1 8B and GPT 3.5 Turbo are 128K and 4096 tokens, respectively.

Aggregation	Input Size	Context Size	Precision	Recall	Rouge-L	Bleu	BERTScore	SemScore
sliding window	1	20	40.09 ± 11.32	44.07 ± 9.77	25.62 ± 6.54	4.59 ± 2.41	86.48 ± 1.26	79.24 ± 3.55
	1	50	47.41 ± 14.83	38.0 ± 11.61	29.41 ± 5.08	6.64 ± 2.74	87.33 ± 1.0	80.15 ± 6.78
growing window	1	20	52.17 ± 10.41	34.88 ± 12.71	27.68 ± 6.12	5.99 ± 2.41	86.99 ± 0.93	79.94 ± 6.4
	1	50	53.73 ± 12.13	37.01 ± 12.5	30.09 ± 3.62	6.23 ± 2.52	87.8 ± 1.09	85.83 ± 4.14
growing window	5	5	66.7 ± 0.130	77.8 ± 7.1	25.1 ± 3.7	5.9 ± 2.0	$84.6 \pm 1.$	82.1 ± 5.2

Table 4: Real-time with context aggregation performance scores.

tems that may possibly have limited computational budget, improving efficiency while maintaining reasonable performance will be critical. A context size of 20 may be unrealistic for an interface system that would not obstruct the natural flow of the conversation.

5 Discussion

Challenges in defining and annotating agenda items. The annotation process for identifying agenda items and clinical details highlighted several challenges. Annotators encountered difficulty with repeated details, as deciding whether to annotate redundant information required balancing the need for comprehensive capture against efficiency. Additionally, defining the exact line where key agenda items first appeared in the transcript proved challenging, particularly when discussions spanned multiple lines. These issues highlight the complexity of clinical interactions, which should be explored further in future work.

Limitations of current models Current large language models (LLMs) exhibit real-time clinical agenda setting limitations. While effectively capturing semantic meaning, they struggle with identifying context-specific nuances in clinical conversations. Issues such as imprecise handling of interruptions and shifts in topics and challenges in understanding implicit context or medical jargon reduce their reliability. Furthermore, our results demonstrate that the model often trades off precision for recall when given larger context windows, leading to noisy outputs.

Considerations for real-time systems Real-time implementation of an agenda-setting system will require addressing several technical challenges. Optimizing the chunk size for processing is critical. As our experiments showed, further exploration will be needed to balance the requirement to provide sufficient context without overwhelming the system's capacity or introducing excessive delays. Effective post-processing mechanisms are needed to refine and structure the summaries generated in real-time. This ensures that information remains concise and actionable while aligning with clinical priorities.

Future Work Several areas of improvement and exploration are necessary to advance the development of real-time agenda-setting systems. First, collaborating with clinicians to refine the definition of agenda items and identify the aspects most relevant to clinical decision-making is vital for system effectiveness. Next, combining real-time agenda setting with tools for pre-visit preparation and post-visit summarization can create a more cohesive clinical documentation workflow. Developing user-friendly interfaces that seamlessly integrate into existing clinical workflows is crucial for adoption. These interfaces should prioritize accessibility and minimize disruptions during patient-provider interactions. Creating metrics designed explicitly for real-time summarization, where the entire transcript is unavailable for comparison, will enable better system performance evaluation. Expanding the dataset to include a broader range of clinical scenarios and diverse patient demographics will enhance the system's generalizability and robustness. Finally, assessing how a real-time system would be integrated into clinical workflows is crucial for clinical utility.

6 Conclusion

This study explored the feasibility and challenges of developing a real-time automated agenda-setting system leveraging generative AI to enhance patient-provider interactions. By analyzing annotated simulated clinical conversations, we evaluated the performance of large language models (LLMs) in identifying agenda items and summarizing clinical details in real-time. Our findings highlight LLMs' potential to support clinical workflows but also present critical limitations, such as difficulties handling dynamic conversations and trade-offs between precision and recall. Addressing these challenges and advancing the proposed system will require interdisciplinary efforts by AI researchers, clinicians, and human-computer interaction experts. With real-time agenda-setting systems, we can enable more efficient, engaging, and effective patient-provider interactions, ultimately improving healthcare outcomes.

Acknowledgments

This research was partially supported by the National Institutes of Health (NIH) under award #DP1-LM014558 (PI: Johnson, 09/01/2023-07/31/2028) for the project "Helping Doctors Doctor: Using AI to Automate Documentation and 'De-Autonomate' Health Care," the National Science Foundation (#NSF-1915398), the Institute for Translational Medicine and Therapeutics, the National Center for Advancing Translational Sciences of the NIH (#UL1TR001878), and by the Collaborative Research in Trustworthy AI for Medicine grant from ASSET at the University of Pennsylvania.

References

Arndt, B. G.; Beasley, J. W.; Watkinson, M. D.; Temte, J. L.; Tuan, W.-J.; Sinsky, C. A.; and Gilchrist, V. J. 2017. Teth-

ered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *The Annals of Family Medicine*, 15(5): 419–426.

Aynetdinov, A.; and Akbik, A. 2024. SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity. ArXiv:2401.17072.

Bain, M.; Huh, J.; Han, T.; and Zisserman, A. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023*.

Del Fiol, G.; Workman, T. E.; and Gorman, P. N. 2014. Clinical Questions Raised by Clinicians at the Point of Care: A Systematic Review. *JAMA Internal Medicine*, 174(5): 710–718.

Ely, J. W.; Osheroff, J. A.; Maviglia, S. M.; and Rosenbaum, M. E. 2007. Patient-Care Questions that Physicians Are Unable to Answer. *Journal of the American Medical Informatics Association*, 14(4): 407–414.

Glenn, T. W.; Riekert, K. A.; Roter, D.; Eakin, M. N.; Pruette, C. S.; Brady, T. M.; Mendley, S. R.; Tuchman, S.; Fivush, B. A.; and Eaton, C. K. 2021. Engagement and affective communication during pediatric nephrology clinic visits: associations with medication adherence. *Patient education and counseling*, 104(3): 578–584.

He, Z.; Han, Y.; Ouyang, Z.; Gao, W.; Chen, H.; Xu, G.; and Wu, J. 2022. DialMed: A Dataset for Dialogue-based Medication Recommendation. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 721–733. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Hood-Medland, E. A.; White, A. E. C.; Kravitz, R. L.; and Henry, S. G. 2021. Agenda setting and visit openings in primary care visits involving patients taking opioids for chronic pain. *BMC Family Practice*, 22(1): 4.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, T.-L.; Hetherington, T. C.; Dharod, A.; Carroll, T.; Bundy, R.; Nguyen, H.; Bundy, H. E.; Isreal, M.; McWilliams, A.; and Cleveland, J. A. 2024. Does AI-Powered Clinical Documentation Enhance Clinician Efficiency? A Longitudinal Study. *NEJM AI*, 1(12): AIoa2400659.

Mathur, Y.; Rangreji, S.; Kapoor, R.; Palavalli, M.; Bertsch, A.; and Gormley, M. R. 2023. SummQA at MEDIQA-Chat 2023:In-Context Learning with GPT-4 for Medical Summarization. arXiv:2306.17384.

Nuance Communications. 2023. Nuance Dragon Ambient eXperience (DAX). https://www.nuance.com/healthcare/ambient-clinical-intelligence/dragon-ambient-experience.html. [Online; accessed October 3, 2023].

OpenAI. 2023a. ChatGPT (GPT-3.5). https://chat.openai.com/. Large language model.

OpenAI. 2023b. GPT-4 Technical Report.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Singh Ospina, N.; Phillips, K. A.; Rodriguez-Gutierrez, R.; Castaneda-Guarderas, A.; Gionfriddo, M. R.; Branda, M. E.; and Montori, V. M. 2019. Eliciting the Patient's Agenda-Secondary Analysis of Recorded Clinical Encounters. *Journal of General Internal Medicine*, 34(1): 36–40.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; Payne, P.; Seneviratne, M.; Gamble, P.; Kelly, C.; Scharli, N.; Chowdhery, A.; Mansfield, P.; y Arcas, B. A.; Webster, D.; Corrado, G. S.; Matias, Y.; Chou, K.; Gottweis, J.; Tomasev, N.; Liu, Y.; Rajkomar, A.; Barral, J.; Semturs, C.; Karthikesalingam, A.; and Natarajan, V. 2022. Large Language Models Encode Clinical Knowledge. arXiv:2212.13138.

Team, V. 2023. Vicuna: An Open-Source Chatbot. https://lmsys.org/blog/2023-03-30-vicuna/. A fine-tuned large language model based on LLaMA.

Toma, A.; Lawler, P. R.; Ba, J.; Krishnan, R. G.; Rubin, B. B.; and Wang, B. 2023. Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. arXiv:2305.12031.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; Mustafa, B.; Chowdhery, A.; Liu, Y.; Kornblith, S.; Fleet, D.; Mansfield, P.; Prakash, S.; Wong, R.; Virmani, S.; Semturs, C.; Mahdavi, S. S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Singhal, K.; Florence, P.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Generalist Biomedical AI. arXiv:2307.14334.

Van Veen, D.; Van Uden, C.; Blankemeier, L.; Delbrouck, J.-B.; Aali, A.; Bluethgen, C.; Pareek, A.; Polacin, M.; Reis, E. P.; Seehofnerová, A.; Rohatgi, N.; Hosamani, P.; Collins, W.; Ahuja, N.; Langlotz, C. P.; Hom, J.; Gatidis, S.; Pauly, J.; and Chaudhari, A. S. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4): 1134–1142.

Veen, D. V.; Uden, C. V.; Attias, M.; Pareek, A.; Bluethgen, C.; Polacin, M.; Chiu, W.; Delbrouck, J.-B.; Chaves, J. M. Z.; Langlotz, C. P.; Chaudhari, A. S.; and Pauly, J. 2023. RadAdapt: Radiology Report Summarization via Lightweight Domain Adaptation of Large Language Models. arXiv:2305.01146.

Xu, K.; Cheng, Y.; Hou, W.; Tan, Q.; and Li, W. 2024. Reasoning Like a Doctor: Improving Medical Dialogue Systems via Diagnostic Reasoning Process Alignment. In Ku, L.-W.;

Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 6796–6814. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675.