HealthTrajectory: Patient Journey Summaries and Visualizations for Patient-Clinician Communication Support

Rohmah Hidayah, Tomohiro Nishiyama, Shoko Wakamiya, Eiji Aramaki
Nara Institute of Science and Technology, Japan
rohmah.hidayah.ri5@naist.ac.jp, {nishiyama.tomohiro.ns5, wakamiya, aramaki}@is.naist.jp

Abstract

In recent years, patient narratives have been used to understand subjective experiences that are not recorded in clinical notes. However, narratives tend to be long and unstructured, requiring summarization. However, text-based summaries often require a lot of clarification from patients and make it difficult for clinicians to review events and changes in symptoms over time. In this study, we expanded the summary output by presenting a visualization of the patient's journey to facilitate communication between patients and medical staff. Referring to the widespread use of LLM for summarization, we compared GPT-4.1 and Gemini-2.5-pro, and used Gemini-3-pro-image-preview for visualization. Data was collected from DIPEx-Japan, then the quality of the summaries was evaluated quantitatively and the visualizations qualitatively. Quantitative evaluation using BLEU and ROUGE metrics showed that Gemini-2.5-pro achieved higher summary scores than GPT-4.1, and Japanese summaries scored higher than English ones. Conversely, English performed better than Japanese in temporal expression extraction using precision, recall, and F1 metrics, and the Gemini-2.5-pro model consistently outperformed GPT-4.1. In qualitative evaluation using the pairwise method, the timetable-based model was far superior with an overall win rate of 0.865 in Japanese and 0.969 in English compared to the baseline.

Keywords: LLMs, Summarization, Visualization, Patient Journey, Timetable

1. Introduction

Sharing information between a patient and a clinician is crucial. The World Health Organization emphasizes the importance of patient and public involvement, positioning patients and families as partners to improve safety, service quality, and patient experience (World Health Organization, 2021). This is especially critical in acute consultations, where clinicians must immediately understand the chronology of symptoms and examinations under tight time constraints.

Patient narratives offer valuable insights into subjective healthcare experiences (Oluoch et al., 2023), yet they are frequently lengthy, fragmented, and unstructured. In particular, chronological events are often expressed inconsistently across languages and contexts, making it difficult for clinicians to synthesize them efficiently (Spitale et al., 2023). In practice, this cognitive load increases the risk of delays and errors in clinical decision-making (Lenz et al., 2025).

Prior work has explored time-based and visualization-oriented approaches. Ledesma et al. (2019) demonstrated that time-based representations assist clinicians in reviewing patient histories and supporting decision-making. Similarly, Torsvik et al. (2025) showed that integrating narrative notes with timelines enables clinicians to quickly access relevant context. Other studies have highlighted the value of patient narratives for understanding healthcare experiences (Ma et al., 2024), time-oriented visualization in clinical contexts (Scheer et al., 2022), and information extraction for clinical decision-making support (Elgaar et al., 2025). However, these approaches primarily rely on structured clinical data from healthcare systems, leaving a gap in transforming long, unstructured, patient-generated narratives into outputs tailored for patient-clinician communication.

Our approach focuses on overcoming two primary challenges inherent in this gap:

Extracting dispersed clinical events: Clinically relevant events, ranging from subjective symptom changes to formal medical consultations, are scattered throughout lengthy texts, with their corresponding temporal expressions often being implicit, inconsistent, or context-dependent.

Communication-oriented visualization: While structured tables provide concise data, text-only formats often lack the intuitive clarity required for rapid understanding in clinical settings.

To address these challenges, we introduce **HealthTrajectory**, an LLM-based pipeline designed to convert lengthy, unstructured patient narratives, such as interview transcripts, into an intuitive visual format. The system has two core stages as shown in Figure 1:

Timetable-based summarization: This stage identifies temporal expressions and maps summarized clinical events to these time

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 149–158

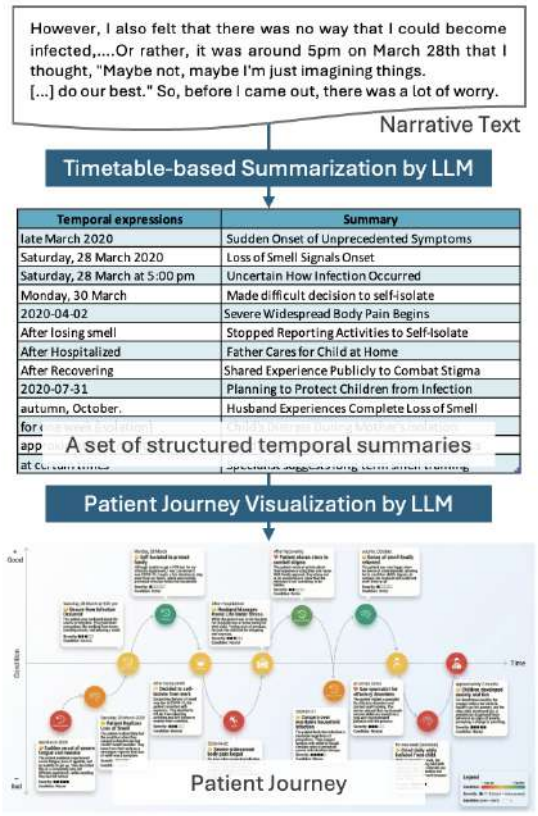

Figure 1: Overview of HealthTrajectory. A patient narrative is processed via Timetable-based Summarization to ensure chronological structure. The resulting summaries are then visualized as a patient journey, transforming raw text into an intuitive trajectory. This intermediate structured step ensures high fidelity, temporal accuracy, and content completeness in the final output.

points to create a structured temporal summaries.

Patient journey visualization: This stage converts the structured summaries into a chronological graphical representation to facilitate effective two-way communication.

The main contributions of this study are as follows:

- We propose **timetable-based summarization**, which identifies temporal expressions and maps summarized clinical events to them, distilling a patient's fragmented narratives into a single structured table.

- We propose a **patient journey visualization** generated from the structured table, enabling a concise and interpretable presentation that facilitates more effective clinical communication than text alone.

2. Related Work

2.1. Medical Summarization with LLMs

The development of LLM has triggered a surge in medical summarization research because these models can summarize long texts such as visit summaries, emergency room summaries, and discharge summaries. For instance, a study by Williams et al. (2025) evaluating GPT-4 and GPT-3.5 for emergency department visit summaries found that while LLM summaries were reasonably accurate, they remained prone to hallucinations and sometimes omitted relevant clinical information.

Other studies have explored the automation of discharge summaries using open-source models, emphasizing an editable draft approach to ensure clinician supervision and patient safety (Ganzinger et al., 2025). Additionally, Grazhdanski et al. (2025) focused on generating synthetic discharge summaries for training and evaluation purposes to bypass the constraints of real patient data. However, conventional paragraph-based summaries often do not clearly highlight the temporal structure and sequence of events, so the chronology still has to be manually rearranged. Consequently, studies such as Williams et al. (2025) encourage more structured summary formats that facilitate traceability to the original source.

2.2. LLM-based Visualization

Research related to converting summaries into images or visualizations is still limited, especially in the medical field. However, clinical communication often requires quick understanding, and visualization is an important complement to text summaries. In the field of natural language to visualization, studies such as Wu et al. (2024) show that LLMs can help convert sentence instructions into visualizations and empirically evaluate the accuracy of the results.

On the other hand, Ouyang et al. (2025) presents an agent-based workflow approach for NL2Vis and highlights the challenges of cross table reasoning, showing that the ability of LLMs to generate visuals depends on input structure and control mechanisms. At the benchmark level, Chen et al. (2024) also confirms that NL2Vis remains challenging and requires rigorous evaluation in the era of LLMs. However, none of the studies have yet created chronological patient summary visualizations that integrate time and medical event information in a structured manner for ease of communication between patients and clinicians, making the visuals not only appealing but also accountable for clinical discussions.

Entry (Excerpts with annotated temporal expressions)	Gold Summary	Theme
However, I also felt that there was no way that I could become infected, and so, as someone in a reporting position, at the time I was already working from home as much as possible, avoiding rush hour traffic, trying to avoid crowds, and wearing a mask, so I had no idea how I had been infected. Or rather, it was around **5pm on March 28th** (**Saturday, 28 March at 5:00 pm**) that I thought, "Maybe not, maybe I'm just imagining things." [...] So, I called my boss and contacted my husband, and since I was at work for a bit, I went home, and we decided that I had no choice but to contact the public health center and follow their instructions, so I went home by train, wearing a mask and being careful.	I felt there was no way I could get infected. I worked from home as much as possible, avoided rush hour and crowded places, and wore a mask, so I had no idea how I was infected.	Route of infections
After I was hospitalized (**After hospitalized**), my husband (who was in close contact with me and is under health observation) stayed at home for two weeks, but we couldn't avoid going shopping, and I couldn't keep my kids confined to the house from morning until night. [...] But even more than that, people whose wives have COVID might be wondering if it's okay for them to be in the park at that moment, or if it's okay to go shopping, but if they don't, they won't be able to eat, and they have children, and so I heard that he was probably thinking about things like that out in the field.	While she was in hospital, her husband was waiting at home and couldn't keep the child locked up all the time, so he chose a time to go shopping and take the child to exercise in the field. I think her husband was also under a lot of pressure.	Young children and coronavirus infection

Table 1: Sample entries from a single patient interview (H01). The table presents English translations of excerpts from the Japanese transcripts, where annotated temporal expressions are highlighted in bold, followed by their normalized labels in parentheses.

3. Dataset

3.1. Materials

Our primary data source comprises patient narratives curated by DIPEx-Japan[1], a national branch of DIPEx International[2]. Adhering to the qualitative research methodology developed at the University of Oxford (Torishima et al., 2020), DIPEx-Japan emphasizes "patient-centered medical care" by documenting the lived experiences of health and illness as a vital social resource.

For this study, we utilized a specialized database of COVID-19 narratives in DIPEx-Japan, comprising interview transcripts that document the lived experiences of 14 individuals: 12 patients (6 men and 6 women, aged 30–70) and 2 patient family members. The interviews were conducted in Japanese from January to July 2021. Note that we excluded two interviews conducted with family members to focus exclusively on first-person patient narratives, resulting in a final dataset of 12 patients.

To ensure thematic consistency, DIPEx-Japan categorized the interview content into 9 themes: "routes of transmission", "children and coronavirus infection", "onset of symptoms", "main symptoms in the acute phase", "anxiety and suffering of infected patients", "impact on work and the workplace", "long-term symptoms and side effects", reporting infections", and "views on life and the world after COVID-19". Since not every patient discussed all nine themes, we treated each available thematic

Interview	Entries	Words (ja)	Words (en)
H01	13	7,847	5,289
H03	6	3,773	2,086
H04	7	3,395	2,217
H05	10	4,638	2,712
H07	4	1,915	1,208
H08	13	6,570	3,997
H09	7	3,026	1,866
H10	8	4,444	2,653
H11	8	4,588	2,955
H12	8	3,618	2,290
H13	7	3,283	2,284
H14	10	5,789	3,630

Table 2: Statistics of the COVID-19 narrative dataset. The dataset based on 12 interviews (one per patient), with a total of 101 thematic sections (entries) identified across all interviews. While the original database included 14 individuals, two interviews with family members (H02 and H06) were excluded to focus on patient narratives. For each interview, the number of available entries and total text length (word counts) are reported.

[1] https://www.dipex-j.org/

[2] http://www.dipexinternational.org/

section as an individual entry as shown in Table 1. Consequently, each patient is represented by a varying number of entries as shown in Table 2, each paired with a summary that serves as a gold standard for text summarization.

3.2. Dataset Construction and Annotation

We translated the original Japanese transcripts into English to evaluate the system's cross-language utility and to ensure consistent performance across different languages. Table 2 provides a statistical overview of the dataset, including the number of patients, segments, and the average text length in both Japanese and English.

Furthermore, to support the evaluation of our patient journey visualizations, one of the authors manually annotated temporal expressions within each segment to form a gold standard timeline. This reference allows for a direct comparison between human-annotated temporal structures and LLM-generated outputs.

4. Method

HealthTrajectory systematically transforms unstructured patient narratives into structured visualizations through a two-stage pipeline. First, we perform timetable-based summarization (Section 4.1), in which an LLM extracts key clinical events and aligns them along chronologically to create structured temporal summaries. Second, these summaries serve as the input for patient journey visualization (Section 4.2), where the LLM renders a graphical representation of the patient's health trajectory.

4.1. Timetable-based Summarization

The initial stage generates summaries and temporal expressions that maintain chronological coherence, thereby producing a stable and structured foundation for our workflow. The process consists of the following steps:

- **Narrative processing**: The input text (e.g., the patient narrative shown in Table 1) is converted into structured XML format along with a timeline. Our system is implemented in both Japanese and English. Both datasets are processed using the same procedure to ensure a fair bilingual comparison of the LLM's performance.
- **Summary Generation**: The model generates concise text summaries from the narratives, ensuring a consistent format throughout the entire timeline.
- **Timeline generation**: The LLM generates time-based structured output, specifically identifying events and their corresponding temporal expressions. For temporal expressions, the model extracts exactly as it appears in the narrative if the time is an expression such as "a week later, before testing positive" in order to maintain time consistency and avoid causing confusion.
- **Chronological ordering**: The model arranges events from the earliest to the latest. It prioritizes the inherent temporal order of the narrative without interpolating or inventing nonexistent time points.

Finally, the output is checked against the predefined schema and required chronological structure. If inconsistencies are found, such as empty columns, invalid events, repeated content, or incorrect event ordering, the model is prompted to regenerate the output until it matches the required format and temporal sequence.

4.2. Patient Journey Visualization

In this stage, we employ an LLM to generate patient journey visualizations. The structure and components of the prompt used for the visualization are shown in Figure 4. Specifically, the prompt is designed to consider the following points:

- **Chronological order**: The timeline is maintained from the earliest point to the present. This is essential because the present state is often implicit rather than explicitly stated in the narrative.
- **Linear connectivity**: Connections between nodes are restricted to a single, one-way chain without branching to ensure a clear and unambiguous narrative flow.
- **Node composition**: Each node features a card-like display incorporating a temporal label, a concise description, an icon/emoji, and a "severity" indicator to reflect changes in the patient's condition over time.
- **Readability and layout**: Explicit rules are established to optimize visual clarity, such as the vertical arrangement of nodes (top to bottom), the prevention of overlapping, and the inclusion of legends. These constraints facilitate rapid information processing for clinical decision-making and help patients intuitively understand their disease progression.

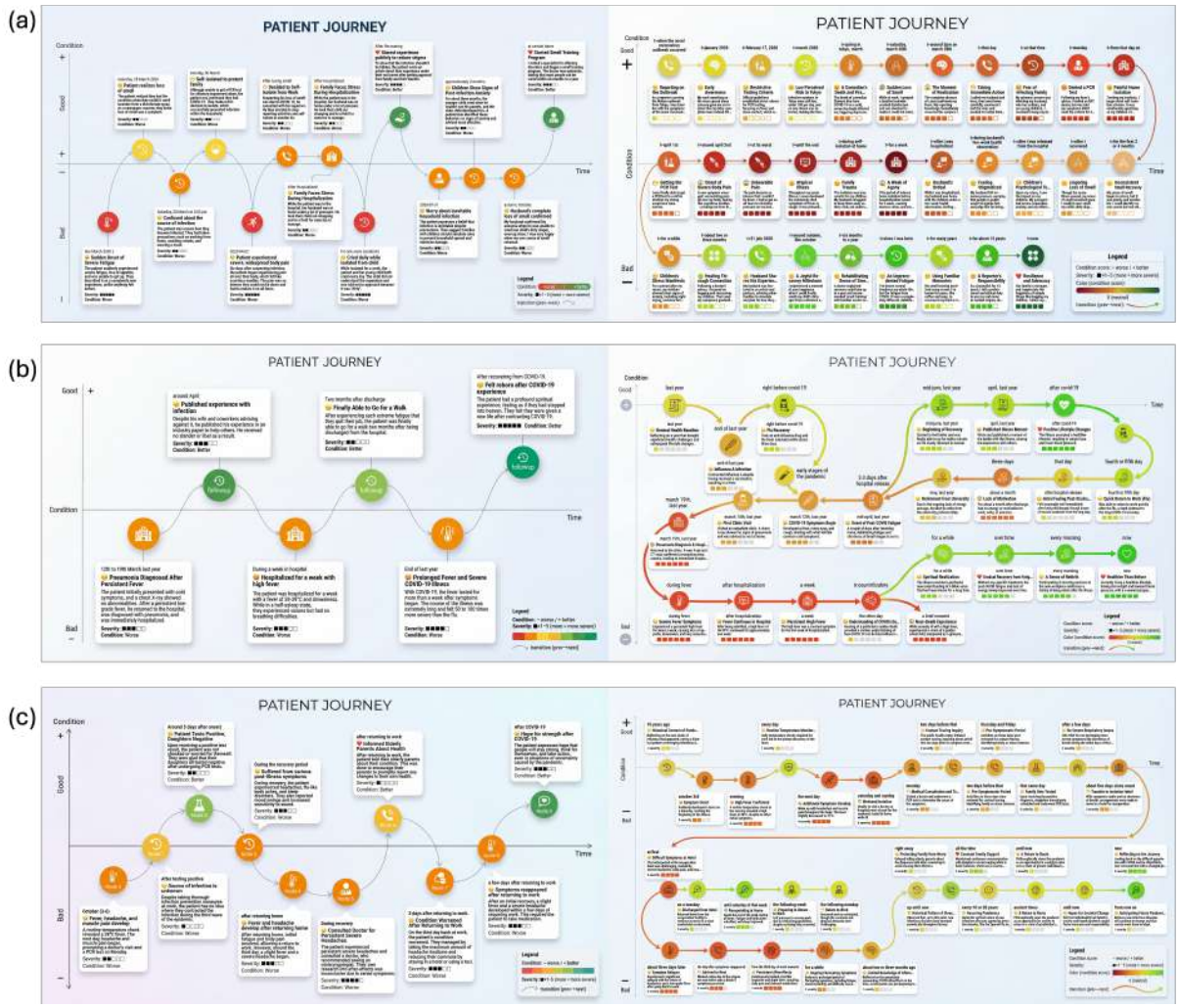

Figure 2: Comparison of patient journey visualizations across three cases. Each pair shows the visualization generated by HealthTrajectory (left) and the baseline (right). This visualization shows the patient's condition score, where a higher score on the positive y-axis indicates a more positive change in the patient's condition, meaning the patient is improving. We also use color to facilitate a quicker understanding of the patient's condition. On the patient card, we also add a number of boxes to the severity scale, indicating a more severe condition if the severity level is high, and a lower number indicates improvement. Transition markers are used to indicate the chronological order of events.

5. Experimental Setup

Our system consists of two steps: experimental setup for timetable-based summarization (Section 5.1) and timetable-based patient journey visualization (Section 5.2).

5.1. Timetable-based Summarization Setup

5.1.1. Models and Prompts

We chose the Gemini-2.5-pro and GPT-4.1 models with a zero-shot setting. We selected GPT-4.1 because medical evaluations based on rubrics indicate that GPT-4.1 is a high-performing model, especially for clinical text comprehension tasks that require accuracy and can perform consistent temporal extraction (not fabricating time and maintaining the sequence of events) (Arora et al., 2025). We chose Gemini-2.5-Pro because official technical reports demonstrate its advanced capabilities in reasoning, multimodality, and long context (Comanici et al., 2025). Both are highly suitable for our workflow, which processes long patient narrative texts into timetable-based summarization.

We used a prompt that instructs the model to divide a single interview transcript from one patient into several segments in chronological order, considering the aspects found in the patient's narrative data as shown in Figure 3. For each segment, the model is instructed to extract a timeline with one to two TIMEX3-style temporal expressions, with con-

System prompt: "You are a time-label sorter, specialized in ordering timeline labels into a strict chronological sequence (oldest→newest)."
Instruction: You receive one interview represented as multiple timeline entries, where each entry contains a timeline label and an optional aspect field. Sort the entries strictly from **oldest to newest**, ignoring the original table order and narrative/story order. If a timeline label is vague, you may use the aspect field only as supporting evidence, but you must not hallucinate or infer dates/events that are not stated in the input. Expressions such as "after N years ..." are not concrete dates and should be placed after entries with explicit dates, while "now/recently" must be placed as the newest.
Output: Return a **JSON** object that provides the **chronological ordering** of all entry indices from oldest to newest.

System prompt: "You are a medical summarizer, specialized in producing faithful point-level summaries for patient-journey tracking."
Instruction: You receive a time label (which must be used **as-is**), an optional aspect field (supporting cue only), and the corresponding content (a text segment or an existing summary, when available). Generate **exactly one** timetable point for each timeline entry. Do not add any facts that are not present in the input (**no hallucination**). Write a short medical summary consisting of a concise `heading` and `detail`. The `heading` must be short ($\leq$ 7 words) and complete (no ellipsis). The `detail` must be concise and complete, written as 1–2 full sentences (no ellipsis, no truncation). Select `emoji` from a predefined emoji pool, and `icon_hint` from a predefined icon inventory. Assign a `severity` level on a 1–5 scale, where `1` indicates the **worst** condition and `5` indicates the **best** condition. If the generated text appears cut off or contains ellipses ("..."/"..."), rewrite it to be shorter but still complete, while preserving the original meaning.
Output: Return a **JSON** object that contains the medical summary and its **severity** level (1–5), along with auxiliary fields used for visual rendering.

Figure 3: Prompts for timetable-based summarization.

System prompt: "You are a medical infographic designer, specialized in rendering a constrained patient-journey timeline as a clean, readable 16:9 vector infographic."
Instruction: You receive a chronologically ordered list of patient-journey points and must produce a professional **16:9** infographic in a clean **vector** style with sharp, readable text (no blur). The infographic must contain exactly `n` nodes (colored circles) and exactly `n-1` arrows, forming a single chain `0->1->2->...->n-1` with no branching and no extra icons/circles. Draw thin axes: a thin x-axis across the middle labelled *Time* near the right end and a thin y-axis on the left labelled *Condition*. Reserve a clean left margin lane (∼10% of the canvas width) exclusively for the y-axis label and polarity markers; no nodes, arrows, cards, legend, or other objects may overlap this lane. Place nodes above/below the x-axis based on the qualitative condition label (e.g., Better/Worse/Neutral), and never print any hidden numeric variables used for layout. Use the provided `severity` (1–5) to encode visual emphasis (e.g., color intensity or a 5-block bar), where 1 represents the **worst** condition and 5 represents the **best** condition. Every node must have exactly one attached card connected by a thin leader line, and cards must not overlap each other, nodes, arrows, axes, the reserved lane, or the legend. Each card must include the time label (as provided), bold `emoji + heading`, the `detail` text (1–2 complete sentences), the `severity` indicator (five blocks), and a qualitative label *Better/Worse/Neutral*. Include a compact legend at the bottom-right showing the severity key (1–5 blocks) and one sample thin curved arrow labelled *transition (prev->next)*; do not show hex codes and do not show any hidden layout scores.
Output: Return one **16:9** infographic image satisfying all constraints above.

Figure 4: Prompt for patient journey visualization.

tent length adjusted to the medical context of each segment. Additionally, the model is asked to summarize the narrative into one to three sentences that align with the medical context of each segment without adding new information.

5.1.2. Evaluation Metrics

Automatic evaluation consists of two components: summary evaluation and timeline evaluation.

Summary Evaluation: We use ROUGE-1/2/L to capture diverse aspects of summary quality. ROUGE-1 (unigram) evaluates how much information from the reference is captured, while ROUGE-2 (bigram) evaluates more strictly and ensures that the model not only captures information but also retains phrases. ROUGE-L (longest common subsequence) assesses similarity in terms of event chronology more flexibly (Zhang et al., 2025). Additionally, we use BLEU to provide a complementary perspective on n-gram overlap (Chu et al., 2024).

Timeline Evaluation: For the timeline generation task, we use precision, recall, and F1-score. These metrics are essential for ensuring that the model accurately extracts temporal expressions while minimizing the inclusion of incorrect time points. This approach is often used for evaluating

temporal information extraction, which are often referred to as TIMEX3 (Yao et al., 2025).

5.2. Patient Journey Visualization Setup

5.2.1. Models and Prompts

We used the Gemini-3-pro-image-preview model[3][4]. This model was launched on November 20, 2025, as the final iteration of the Nano Banana model for image generation. This model is designed to follow complex instructions with more accurate image results and is specifically tailored for the production of high-quality visual assets, which is why we chose this model over others.

We use a prompt that instructs the model to visualize the patient's journey with a timeline infographic design that includes visual elements such as icons, nodes, and chronological order, as shown in Figure 4. The prompt also instructs the model to create a time axis (x-axis) and a condition axis (y-axis), and each event has a time label, an icon/emoji as a marker, a brief description, and a severity level indicator (1–5) shown using different colors. To keep the visualization readable and avoid overlapping when many events are present, the layout is allowed to extend downward. Overall, the instructions aim to keep all elements structured and clearly presented.

5.2.2. Evaluation Metrics

To quantify the effectiveness of the generated visualizations, we conducted a human evaluation based on pairwise comparisons over the entire dataset. This method provides a more reliable measure of human preference than independent scoring, as it directly compares the outputs of two systems given the same input narrative.

We evaluated the visualizations using three criteria, each consisting of two evaluation aspects summarized from our questionnaire:

- **Correctness:** Assesses *clinical relevance* (whether the visualization accurately reflects the medical context) and *hallucination avoidance* (whether the visualization avoids adding non-existent information).
- **Coverage:** Evaluates *completeness* (the extent to which information from the table is captured) and *faithfulness* (the degree to which the visualization adheres to the table).
- **Visual design:** Measures *readability* (ease of understanding the content) and *layout quality* (effectiveness of the design for following the timeline).

To determine the relative performance, we calculated the winning proportion for the proposed method. Let W_{base} denote the number of votes for the baseline, and W_{ours} denote the votes for the proposed method. The winning proportion is defined as:

$$\mathrm{WinProp} = \frac{W_{ours}}{W_{base} + W_{ours}}, \tag{1}$$

where $\mathrm{WinProp} > 0.5$ indicates a preference for the proposed method.

The evaluation was conducted by four Japanese evaluators, including three NLP researchers and one medical professional, to ensure both technical and clinical validity. For each patient case, they were presented with the gold standard timeline table (Section 3.2) and asked to determine which visualization better represented the structured information.

6. Results and Discussion

6.1. Timetable-based Summarization

In timetable-based summarization, we evaluated two viewpoints: the results of the Gemini-2.5-Pro model and the GPT-4.1 model on the timeline, as well as the results of the Gemini-2.5-Pro model and the GPT-4.1 model on summarization.

Summarization Results: Table 3 shows the summarization evaluation results. In the table, the Gemini-2.5-pro model achieved higher scores in both Japanese and English compared to the GPT-4.1 model in the narrative summarization task. Specifically, the Gemini-2.5-pro model scored the highest in all measurements for Japanese, with a score of 0.506 on ROUGE-1, followed by a score of 0.230 on ROUGE-2, then 0.316 on ROUGE-L, and 0.474 on BLEU. This indicates that the Gemini-2.5-pro model is more stable in maintaining core content and word choices that are closer to the reference.

Timeline Results: Table 3 shows the performance of the timeline generation. The results in English are higher compared to Japanese, which is in contrast to the summary evaluation that shows higher scores for Japanese compared to English. The scores for precision, recall, and F1 are consistently highest for the Gemini-2.5-pro model, with a score of 0.251 for precision and 0.242 for recall and 0.246 for F1 in English time extraction. From these results, the Gemini-2.5-pro model appears more conservative in extracting times that closely match the reference compared to the GPT-4.1 model.

[3] www.ai.google.dev/gemini-api/docs/image-generation

[4] www.ai.google.dev/gemini-api/docs/changelog

Lang	Model	Summary Evaluation				Timeline Evaluation		
		ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	BLEU↑	P↑	R↑	F1↑
JA	Gemini-2.5-pro	**0.506**	**0.230**	**0.316**	**0.474**	0.157	0.157	0.157
JA	GPT-4.1	0.472	0.204	0.294	0.433	0.179	0.159	0.168
EN	Gemini-2.5-pro	0.276	0.077	0.186	0.253	**0.251**	**0.242**	**0.246**
EN	GPT-4.1	0.256	0.065	0.176	0.233	0.250	0.231	0.240

Table 3: Performance of timetable-based summarization in Japanese (ja) and English (en) text.

Criterion	Aspect	Japanese			English		
		W_{base}	W_{ours}	WinProp	W_{base}	W_{ours}	WinProp
Correctness	Clinical relevance	8	40	0.833	1	47	0.979
	Hallucination avoidance	3	45	0.938	0	48	1.000
Coverage	Completeness	1	47	0.979	0	48	1.000
	Faithfulness	3	45	0.938	0	48	1.000
Visual design	Readability	13	35	0.729	5	43	0.896
	Layout quality	11	37	0.771	3	45	0.938
	Overall	**39**	**249**	**0.865**	**9**	**279**	**0.969**

Table 4: Results of Patient Journey Visualization using the pairwise method. W_{base} and W_{ours} show the vote counts for the baseline and the proposed method by aspect and language, respectively; WinProp is the winning proportion of our method.

6.2. Patient Journey Visualization

As shown in Table 4, the evaluators consistently preferred the proposed method over the baseline across all evaluation aspects for both Japanese and English. The proposed method achieved substantially higher preference scores, with 0.865 for Japanese and 0.969 for English, indicating a clear advantage over the baseline. These results suggest that structuring patient narratives into timetable-based visualizations improves the perceived quality of the output in terms of clinical relevance, completeness, faithfulness, and hallucination reduction.

in the English setting, the proposed method achieved perfect scores (1.000) for hallucination avoidance, completeness, and faithfulness. In contrast, improvements in visual aspects such as readability, layout quality, and clinical relevance were more moderate, particularly in the Japanese setting. This discrepancy may indicate that while the proposed method effectively constrains content generation, visual presentation quality is influenced by additional factors, such as language-specific phrasing and formatting preferences.

Overall, the proposed method (left panel of Figure 2) appears to be well-suited for bilingual Patient Journey visualization, supporting clearer and more structured explanations of patient trajectories. The results based on Gemini-2.5-pro further suggest that high-quality upstream generation is critical for producing coherent and interpretable visualizations, particularly in maintaining alignment between clinical events and temporal information.

7. Conclusion

We proposed a bilingual timetable-based Patient Journey system that generates structured visualizations from patient narratives to support clinician–patient communication. The evaluation results demonstrate that the proposed system is robust for summary generation, with the Japanese version generally outperforming the English version on both the Gemini-2.5-pro and GPT-4.1 models. In contrast, for temporal expression extraction, the English version achieved higher accuracy, and Gemini-2.5-pro consistently outperformed GPT-4.1. Furthermore, in the human evaluation comparing the baseline and the proposed method, the evaluators almost always chose the proposed method. This preference was primarily attributed to its higher clinical relevance, fewer hallucinations, greater completeness, and improved readability, making it more suitable for supporting communication between patients and clinicians across languages.

In future work, we plan to implement **HealthTrajectory** for direct use in clinician-patient communication. This system will provide an interface that enables patients to upload information about their symptoms prior to pain onset. The system will generate a visual summary and timeline of the patient's journey. The system will also include features that support both patients and clinicians, allowing the outputs to be reviewed prior to use in clinical discussions.

8. Limitations

There are several limitations to this study. First, the experiments were conducted on a small dataset of COVID-19 patient narratives, requiring validation on a larger and more diverse dataset, including other diseases and narrative types. Second, the timeline annotation was performed by only one annotator, leading to a lack of agreement between annotators and also introducing the potential risk of subjective bias in the gold standard annotation. Third, for the generated patient journey images, each case was generated up to four times as a practical measure to reduce inconsistencies; however, some results still showed inconsistencies regarding severity. Furthermore, the current system is still in its early research phase, and these findings are intended to inform further development towards more practical use in the future.

9. Acknowledgements

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP), Project JPJ012425, and the Japan Society for the Promotion of Science (JSPS) Research Start-up Grant JP25K24412. We also sincerely thank the JICA Research Scholarship for their financial support and our research group members for their useful discussions and comments.

10. Bibliographical References

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. 2024. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*.

An Chu, Thong Huynh, Long Nguyen, and Dinh Dien. 2024. A comparative study of chart summarization. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 971–981.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Mohamed Elgaar, Hadi Amiri, Mitra Mohtarami, and Leo Anthony Celi. 2025. Meddecxtract: A clinician-support system for extracting, visualizing, and annotating medical decisions in clinical narratives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 481–489.

Matthias Ganzinger, Nicola Kunz, Pascal Fuchs, Cornelia K Lyu, Martin Loos, Martin Dugas, and Thomas M Pausch. 2025. Automated generation of discharge summaries: leveraging large language models with clinical data. *Scientific Reports*, 15(1):16466.

Georgi Grazhdanski, Vasil Vasilev, Sylvia Vassileva, Dimitar Taskov, Izabel Antova, Ivan Koychev, and Svetla Boytcheva. 2025. Synthmedic: utilizing large language models for synthetic discharge summary generation, correction and validation. *Journal of Biomedical Informatics*, page 104906.

Andres Ledesma, Niranjan Bidargaddi, Jörg Strobel, Geoffrey Schrader, Hannu Nieminen, Ilkka Korhonen, and Miikka Ermes. 2019. Health timeline: an insight-based study of a timeline visualization of clinical data. *BMC medical informatics and decision making*, 19(1):170.

Janek Lenz, Isabel Richter, and Sven Meister. 2025. Automated filtering and visualization of patient-centered data from electronic health records in emergency care: A scoping review. *Journal of Multidisciplinary Healthcare*, pages 6503–6517.

Hua Ma, Xiaoru Yuan, Xu Sun, Glyn Lawson, and Qingfeng Wang. 2024. Seeing your stories: visualization for narrative medicine. *Health Data Science*, 4:0103.

Dorothy Oluoch, Sassy Molyneux, Mwanamvua Boga, Justinah Maluni, Florence Murila, Caroline Jones, Sue Ziebland, Mike English, and Lisa Hinton. 2023. Not just surveys and indicators: narratives capture what really matters for health system strengthening. *The Lancet Global Health*, 11(9):e1459–e1463.

Geliang Ouyang, Jingyao Chen, Zhihe Nie, Yi Gui, Yao Wan, Hongyu Zhang, and Dongping Chen. 2025. nvagent: Automated data visualization from natural language via collaborative agent workflow. *arXiv preprint arXiv:2502.05036*.

Jan Scheer, Alisa Volkert, Nicolas Brich, Lina Weinert, Nandhini Santhanam, Michael Krone, Thomas Ganslandt, Martin Boeker, and Till

Nagel. 2022. Visualization techniques of time-oriented data for the comparison of single patients with multiple patients or cohorts: Scoping review. *Journal of medical Internet research*, 24(10):e38041.

Giovanni Spitale, Andrea Glässel, Mirriam Tyebally-Fang, Corine Mouton Dorey, and Nikola Biller-Andorno. 2023. Patient narratives–a still undervalued resource for healthcare improvement. *Swiss Medical Weekly*, 153:40022.

Masako Torishima, Michiko Urao, Takeo Nakayama, and Shinji Kosugi. 2020. Negative recollections regarding doctor–patient interactions among men receiving a prostate cancer diagnosis: a qualitative study of patient experiences in japan. *BMJ open*, 10(1):e032251.

Torbjørn Torsvik, Andreas Brun, Bjørn Ståle Benjaminsen, and Aslak Steinsbekk. 2025. Labvis: usability testing of a prototype tool for integrating timeline graphs and clinical notes. *BMC Medical Informatics and Decision Making*, 25(1):337.

Christopher YK Williams, Jaskaran Bains, Tianyu Tang, Kishan Patel, Alexa N Lucas, Fiona Chen, Brenda Y Miao, Atul J Butte, and Aaron E Kornblith. 2025. Evaluating large language models for drafting emergency department encounter summaries. *PLOS digital health*, 4(6):e0000899.

World Health Organization. 2021. *Global patient safety action plan 2021-2030: towards eliminating avoidable harm in health care*. World Health Organization.

Yang Wu, Yao Wan, Hongyu Zhang, Yulei Sui, Wucai Wei, Wei Zhao, Guandong Xu, and Hai Jin. 2024. Automated data visualization from natural language via large language models: An exploratory study. *Proceedings of the ACM on Management of Data*, 2(3):1–28.

Jiarui Yao, Eli Goldner, Harry Hochheiser, Sean Finan, John Levander, David Harris, Piet C de Groen, Elizabeth Buchbinder, Danielle Bitterman, Jeremy L Warner, et al. 2025. Systemic anticancer therapy timelines extraction from electronic medical records text: algorithm development and validation. *JMIR Bioinformatics and Biotechnology*, 6(1):e67801.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.