# Room of Thoughts: Exploring the Role of AI Assistants in 3D Space

Joseph Rupertus
joerup@princeton.edu
Princeton University
Princeton, New Jersey, USA

Parastoo Abtahi
abtahi@cs.princeton.edu
Princeton University
Princeton, New Jersey, USA

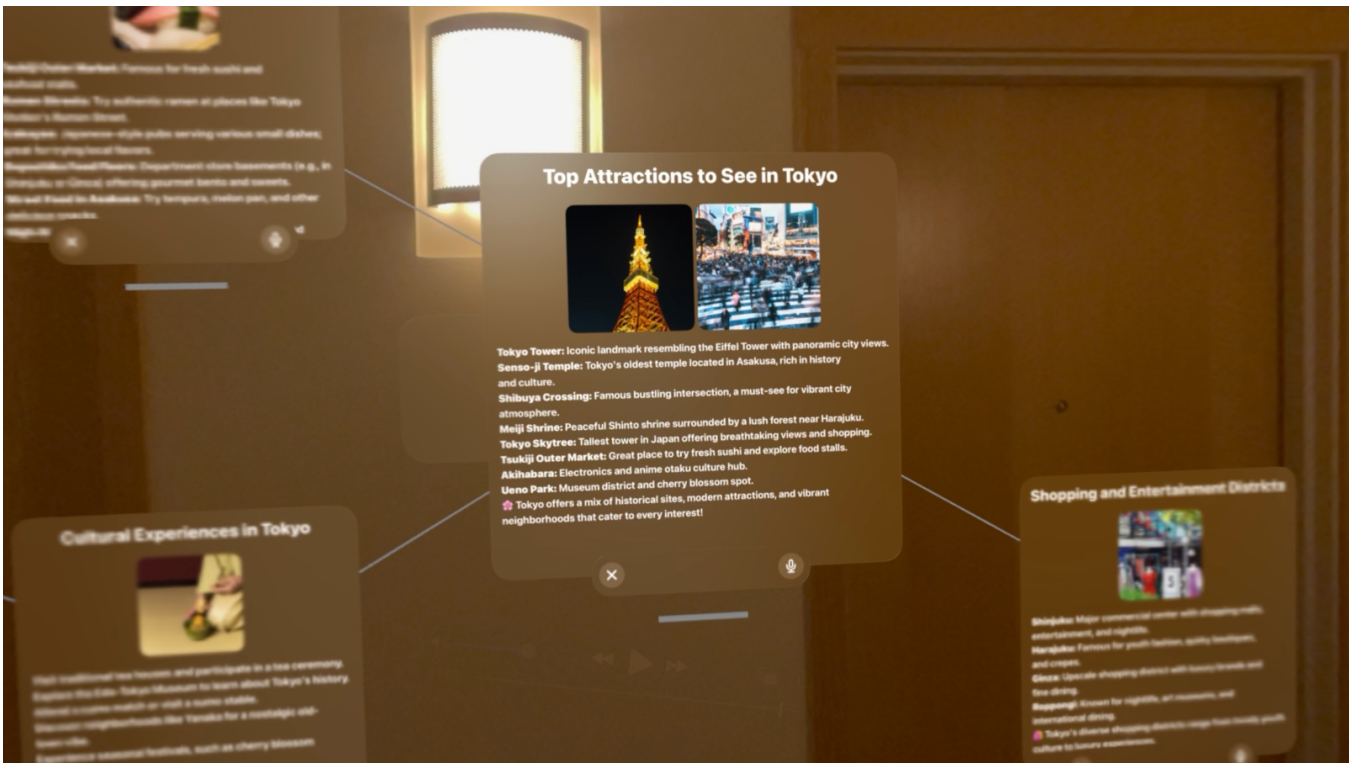


**Figure 1: Room of Thoughts displays AI-generated content as a graphical "mind map" anchored to the user's physical room.**

## ABSTRACT

We present Room of Thoughts, a novel way to interact with Large Language Models (LLMs) in a visual, graphical interface within augmented reality on Apple Vision Pro. Room of Thoughts combines visualized organization with spatial anchoring by using virtual windows embedded within the user's physical room that are separated by topic and logically connected by visual links. Room of Thoughts aims to make interaction with LLMs more visual and more engaging, and enable higher levels of exploration and learning.

## 1 INTRODUCTION

Since their inception and popularization in recent years, users have interacted with Large Language Models (LLMs) primarily through chat interfaces, similar to what one might find in a messaging app. While this conversational format feels intuitive, it represents just one way of interacting with these models. To date, most LLM applications have focused on conversational exchanges in linear text, leaving relatively unexplored how alternative formats like visual or spatial arrangements might change the experience of interacting with AI-generated content. Adherence to a linear chat format somewhat constrains how the content is presented and revisited, as it forces users to engage with complex responses as continuous streams of text. There may be more effective interaction media that improve information absorption and make it easier to revisit prior ideas.

This project explores a visual and spatial alternative to the linear chat interface: presenting LLM outputs as a network of panels arranged graphically in augmented reality. The goal is to go beyond the medium of text on a screen and into 3D space in augmented reality, creating a "bulletin board" or "mind map" of thoughts displayed in the user's physical room. Since content is retrievable in a physical location rather than in a text chain, users can simply return to that physical location rather than scrolling back to find the information from before. The spatial arrangement also has other benefits, in that it allows users to mentally associate content with

physical space, a technique known as the "memory palace" that has been shown to improve memory retention [7].

The evaluation of the system aims to answer three questions:

- Does the system improve memory retention and enhance comprehension of the LLM response over traditional 2D chat interfaces?
- How does the system affect cognitive load compared to traditional 2D chat interfaces?
- To what extent do users prefer to interact with LLMs in a spatial environment over a chat interface? (i.e. what types of tasks are best suited for the system?)

## 2 RELATED WORK

### Graphical LLM Output

Prior work in Human-Computer Interaction has explored various approaches to enhance information processing and interaction using LLMs. Sensecape [8] and Graphologue [4] explore novel ways of interacting with LLM outputs through visual interfaces that allow for diagrams rather than text and the ability to view different levels of abstraction. MoGraphGPT [10] extends this direction by combining LLMs with a graphical interface, enabling non-programmers to create 2D interactive scenes through natural language and direct manipulation. This project will build on this research to expand visual LLM output into a 3D spatial arrangement.

### LLM Output in Physical Space

Existing research has demonstrated examples of how LLMs can be used to generate content for a user's physical space. LLMR [9] presents a framework where large language models generate and modify interactive mixed reality environments by producing Unity code, enabling dynamic scene creation and object manipulation directly in 3D space through natural language prompts. Similarly, LLMER [2] introduces a system that uses LLMs to generate structured JSON data to control XR environments, leading to simpler interaction methods for object creation and animation, and allowing for real-time interaction with virtual elements in the user's physical surroundings.

Other research has shown how virtual content can be best integrated into a user's physical space to take full advantage of mixed reality. SituationAdapt demonstrates how virtual elements and screens can adapt intelligently within augmented environments to minimize user disruption [5]. However, existing apps to interact with LLMs for Apple Vision Pro and other Mixed Reality headsets use traditional 2D chat interfaces embedded in 3D space, as seen in the ChatGPT app on Apple Vision Pro [6]. These interfaces do not fully make use of the physical space but merely overlay the same chat interface that is found on desktop and mobile devices into an AR environment. This project aims to go a step further by integrating LLM output into various elements within the user's physical environment in a graphical way.

## 3 APPROACH

Room of Thoughts combines the ideas of graphical organization and spatial anchoring for LLM output. Rather than presenting LLM interaction as a linear text conversation, or as a graphical structure on a 2D screen, the system instead arranges responses as a network of windows spread out in 3D space, with each window representing a distinct facet of the user's prompt. Windows are connected by visual links to help users understand the relationships between topics at a glance.

Each response is structured hierarchically, with an overview that branches into subtopics, allowing users to explore information at different levels of detail. This layout encourages exploration where users can move between threads of information non-linearly and spatially rather than scrolling through a chat history. The graphical structure mirrors the way ideas naturally branch and connect. By anchoring content directly in physical space, the system also provides the opportunity for users to employ spatial memory techniques and directly associate the content with the space.

## 4 IMPLEMENTATION

Room of Thoughts is a visionOS app built using SwiftUI and RealityKit. The app makes use of an immersive space that displays AI-generated content all around the user's physical room.

*Dictated User Input.* The user is first presented with a small window containing a microphone button to dictate a prompt as input. They simply tap the button and speak their prompt out loud. The dictated input is processed using Apple's native Speech framework in Swift and converted to text.

*Content Generation.* The system calls OpenAI's API to generate content to respond to the user's prompt. It prompts the GPT-4.1 model to respond using a structured output of 'window' objects which contain titles, content, and image descriptions along with a short 'overview' description of the entire output. Windows are organized into a tree structure after generation.

The AI model is instructed to generate between 1 and 5 windows depending on the complexity of the topic asked about. The first window acts as an overview of the topic while any and all subsequent windows explore specific, distinct facets or topics related to it. The AI is also instructed to be brief and concise when possible but still output enough information to be useful. Any additional "follow-up" prompts later use a separate context that instructs the model to generate more windows that are then added to the existing tree structure.

*Image Generation.* The image descriptions are first generated within windows as noted above. The AI is instructed to generate these descriptions describing between 0 and 2 hypothetical relevant images that do not contain diagrams or text. Immediately following window generation, the system makes a separate call to OpenAI's image generation API to actually create the images themselves using the descriptions that were returned in the first call. This returns an image URL which the system associates with the window objects.

*RealityKit Scene Creation.* RealityKit entities are created for each response and added to an immersive space inside the visionOS environment. Using the specific windows that were outputted by the API, the system makes custom window panel entities in RealityKit and attaches SwiftUI views that contain the relevant text

and images for each window. The entities are given additional RealityKit components that enable gesture interactions via a bottom drag indicator to mimic standard visionOS system windows.[1] The system also creates connection meshes that visually link windows together by a line, following the tree structure that the windows are stored in.

The root window entity is attached to an `AnchorEntity` which anchors the content to a physical wall in the user's space that is selected internally by visionOS. However, the user is free to move the windows around using drag gestures to any position in 3D. When the root window is repositioned, it reorients to directly face the user. When other windows are repositioned, they orient in the same direction as the root window.[2]

*Final Output.* Once the API has returned the response and the RealityKit content has been created, the scene appears in front of the user. They then see a full graphical interface of organized windows arranged in a network with text and images. At this point, the system uses Apple's AVFAudio framework to give a brief spoken overview of what the user is seeing in the visualization.

*Follow-Up Prompts.* Each window contains its own dictation input button where the user can prompt the AI for further context, clarification, or additional questions. The additional generated content is simply added to the tree structure as a child of the window that was selected, and additional RealityKit entities are added to the scene.

## 5 METHODS

This evaluation aimed to test whether the graphical mind map interface improves memory retention and user experience compared to a traditional linear layout. We evaluated the system with $n = 12$ users, all undergraduate students from Princeton University, with varied areas of study including engineering, natural sciences, and the humanities. Most participants reported some prior experience with LLMs, but limited exposure to AR apps. All participants were of similar age (18–22) and received $10 compensation for their participation.

### Part 1: System Comparison

The first phase compared the experimental graphical interface used in Room of Thoughts with a control condition. The control condition presented the same content as the graphical mind map but formatted as a single, linear block of text in one window. Note that both conditions still used augmented reality.

Participants were assigned to one of four groups and instructed to dictate the same travel-related prompt for two cities: Nairobi, Kenya, and Bergen, Norway. Specifically:

```
I am planning a trip to [city] and I would like some
recommendations on what to see.
```

However, the order of the cities differed between groups, and the city assigned to the control condition also differed between groups. Each of the 4 combinations was assigned to one group.

| Group | First Task | Second Task |
|---|---|---|
| 1 | Nairobi (Experimental) | Bergen (Control) |
| 2 | Bergen (Experimental) | Nairobi (Control) |
| 3 | Nairobi (Control) | Bergen (Experimental) |
| 4 | Bergen (Control) | Nairobi (Experimental) |

**Table 1: Order of conditions and cities for each group.**

The system was slightly modified during this phase to return a specific hard-coded response to each of these prompts, allowing for standardization among participants. Pairing specific cities with either the experimental or control condition in a counterbalanced design helped eliminate variability in both the content of the responses and the order in which information was presented.

*NASA TLX.* After each of the two prompts, participants were instructed to fill out the NASA TLX questionnaire [3] to assess their perceived cognitive workload.

*Memory Recall Task.* Finally, participants were instructed to answer 10 multiple-choice memory recall questions related to the generated content they observed: 5 questions involving information about Nairobi, and 5 questions involving information about Bergen. The accuracy of these responses was then computed and averaged among each city and group.

We computed the difference in mean accuracy between the Nairobi questions and the Bergen questions for each group, and organized them based on which of the two cities was observed in the experimental vs. control condition for that group. This aimed to determine whether there is any difference in memory retention by observing the graphical interface versus the control condition.

### Part 2: User Experience

The second phase of the study allowed participants to explore the system freely, and specifically instructed them to make use of the "follow up" feature to further add to their generated content.

*System Usability.* Participants completed ten standardized questions from the System Usability Scale (SUS) [1] on a scale of 1 to 7, meant to evaluate the usability and effectiveness of the application.

*Qualitative Feedback.* Participants also answered open-ended questions to provide qualitative feedback. Specifically, they were asked about their preferences between the graphical interface and the control interface, their thoughts on the "follow up" feature, how augmented reality contributes to the usefulness of this type of system, and any suggestions for improvements they may have.

## 6 RESULTS

### Quantitative Results

*Memory Recall Task.* Participants demonstrated higher accuracy when recalling information from the graphical interface compared to the control interface. Accuracy scores were analyzed at the individual participant level, comparing each participant's performance across the two conditions. On average, participants recalled 16.7

[1]We did not use system visionOS windows here because there are restrictions on programatically arranging and sizing them, so we opted for creating custom windows instead.

[2]This is a simplification that was used to avoid extra complexity in programming rotation gestures, but it would be nice to have more rotation options later.

percentage points more with the graphical interface (SD = 23.9). A paired t-test confirmed that this difference was statistically significant, $t(11) = 2.20$, $p = 0.050$, with a 95% confidence interval ranging from 1.5% to 31.8%. Additionally, all four counterbalanced groups showed higher mean accuracy in the graphical condition, suggesting the effect was consistent across different content and presentation orders.

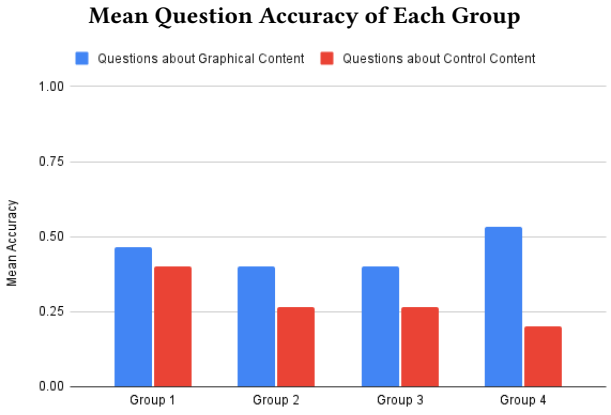


**Figure 2: The mean accuracy for each group for questions related to (a) the city which was presented in the graphical interface and (b) the city which was presented in the control interface.**

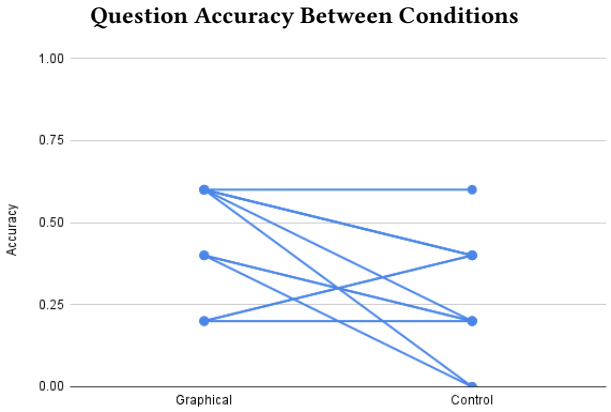


**Figure 3: The accuracy for each individual participant in the graphical (experimental) condition and the control condition.**

*NASA TLX.* The NASA Task Load Index (TLX) revealed that participants perceived the graphical interface as more mentally demanding (+2.78 points on a 100 point scale) and more physically demanding (+5.56) than the control interface. Temporal demand (hurriedness) was rated similarly between conditions (+1.39). Notably, participants reported feeling less successful (-5.56) and that the graphical condition required more effort (+6.94). However, frustration and stress levels were rated the same across both interfaces (0.00 difference).

*System Usability.* Our system was given a mean score of 81.25 (SD = 12.41), which falls within the "Excellent" usability range, well above the standard benchmark score of 68. This indicates that participants found the system highly usable, despite the higher mental and physical workload reported in the NASA TLX.

## Qualitative Feedback

*Graphical Interface.* Most participants preferred the graphical interface over a traditional linear format, especially for the purposes of learning and exploring new ideas. Dividing topics visually made the experience feel less cumbersome, and made it easier to draw connections between them. Users felt this format was more visually appealing and helped in navigating complex information. P12 appreciated the visual reminders of how information was linked, especially when diving into deeper topics. P5 noted that while a single window worked well to go in depth, the graphical interface encouraged discovery and connections that might not have emerged otherwise. P11 reflected that while reading feels more natural in a linear format, graphical layouts excel at facilitating exploration and conversation. However, some noted the visualization could be overwhelming at times, especially with large amounts of content, and suggested instead the system use shorter, simpler phrasing.

*Continued Context and Interaction.* Participants valued the ability to tap on a window and directly dictate follow-up questions without the need for excessive scrolling through a list of written messages. The graphical approach made it easy to revisit earlier topics and branch into new areas, an improvement over a traditional chat interface. P10 noted that this approach encouraged non-linear exploration, which is especially applicable to taks like coding, where users may want to go in many different directions simultaneously. P5 reflected that the system is well-suited for exploratory learning scenarios where users may not know the order in which they want to engage with the material. The tree structure of the graphical layout and the ability to continue prompting to expand the network in a given direction is ideal for this back-and-forth mental movement.

*Spatial and AR Experience.* The spatial aspect of the interface in AR received mixed responses. Some found that embedding the information into physical space made the experience more engaging and digestible, helping them to focus and position things conveniently. For others, the spatial design felt more like a novel backdrop rather than an enhancement, with some preferring the comfort and familiarity of a screen-based environment, and with some concerns over the physical discomfort of wearing the headset. Few participants explored the room extensively, but many interacted with the floating windows, moving them closer or repositioning them for better visibility. Several noted that the 3D mind map in AR felt more immersive than a 2D screen would allow. P12 suggested leveraging this spatial setup to create customizable “desktops” or memory palace arrangements that could aid information retention.

*Suggestions for Improvement.* Participants suggested several areas for refinement. One popular idea was improving rotation controls by allowing connected windows to rotate around the user in a spherical arrangement. Participants also recommended adding indicators, animations, or sounds to signal when the system had finished generating responses. Better dictation controls, such as a hands-free, always-listening mode, were also a frequent request. Incorporating more multimedia elements and offering freeform control over text, such as detaching or rearranging content, could further improve the experience. Some users reported hardware-specific issues with the Vision Pro, such as eye strain, unclear visuals, and environmental distractions, which, if mitigated, could further improve comfort and usability.

## 7 DISCUSSION

From user testing, we found supporting evidence for the use of AR graphical visualizations when interacting with LLMs. Participants demonstrated higher memory retention with the graphical interface compared to a traditional linear format, suggesting that organizing LLM outputs into networks of windows can support learning and comprehension.

A key takeaway from participant feedback was that the graphical interface was especially well-suited for exploration and learning. Several participants noted that visually dividing topics made it easier to absorb information and allowed them to draw connections between ideas more easily. This structure encouraged a non-linear flow of interaction, making it feel natural to branch into subtopics or revisit earlier windows without the limitations of scrolling through a linear chat history.

At the same time, the benefits of spatially anchoring the generated content in AR remain less clear. While some participants appreciated the ability to arrange information in physical space, others felt the AR aspect was more of a novelty than a useful feature, citing the discomfort of the headset and a preference for familiar screen-based environments. This suggests that even though the graphical organization itself is valuable, the spatial aspect may need to be further developed to determine how to best provide additional benefits to users.

Despite reporting higher cognitive and physical workload when using the graphical interface, participants rated the system as highly usable overall, with a System Usability Scale score of 81.25, well above the standard benchmark. This indicates that while the interface may require more mental effort, it still remains effective and intuitive for users.

## 8 FUTURE WORK

This project contributes to the exploration of how graphical and spatial interfaces can improve human interactions with LLMs, and highlights areas for future research in combining LLMs with mixed reality environments. In future versions of this system, we plan to expand the system to support 3D models, diagrams, and charts, adding more visual richness to the environment, as well as more control over window positioning and layout. Future work should further assess this type of system in terms of learning, engagement, and usability, with larger studies across a wider variety of use cases and prompts, to confirm and expand on the results of this study. There should also be more research done to determine whether spatial anchoring in AR can improve memory retention and whether these kinds of interfaces can support different types of tasks beyond learning, such as creative work or problem-solving.

## REFERENCES


[1] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (11 1995).

[2] Jiangong Chen, Xiaoyi Wu, Tian Lan, and Bin Li. 2025. LLMER: Crafting Interactive Extended Reality Worlds with JSON Data Generated by Large Language Models. *IEEE Transactions on Visualization and Computer Graphics* 31, 5 (May 2025), 2715–2724. https://doi.org/10.1109/TVCG.2025.3549549 arXiv:2502.02441 [cs].

[3] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, Vol. 52. Elsevier, 139–183.

[4] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3586183.3606737

[5] Zhipeng Li, Christoph Gebhardt, Yves Inglin, Nicolas Steck, Paul Streli, and Christian Holz. 2024. SituationAdapt: Contextual UI Optimization in Mixed Reality with Situation Awareness via LLM Reasoning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3654777.3676470

[6] Michael Nuñez. 2024. OpenAI launches ChatGPT app for Apple Vision Pro. https://venturebeat.com/ai/openai-launches-chatgpt-app-for-apple-vision-pro/

[7] Ayisha Qureshi, Farwa Rizvi, Anjum Syed, Aqueel Shahid, and Hana Manzoor. 2014. The method of loci as a mnemonic device to facilitate learning in endocrinology leads to improvement in student performance as measured by assessments. *Advances in Physiology Education* 38, 2 (June 2014), 140–144. https://doi.org/10.1152/advan.00092.2013

[8] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3586183.3606756

[9] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. https://doi.org/10.48550/arXiv.2309.12276 arXiv:2309.12276 [cs].

[10] Hui Ye, Chufeng Xiao, Jiaye Leng, Pengfei Xu, and Hongbo Fu. 2025. MoGraphGPT: Creating Interactive Scenes Using Modular LLM and Graphical Control. https://doi.org/10.48550/arXiv.2502.04983 arXiv:2502.04983 [cs].

## A LLM PROMPT

This prompt was submitted to GPT-4.1 via OpenAI's API for initial context:

```
You are a helpful AI assistant designed to provide a visualization of some topic that the user asks about. Please respond to the user's prompt which is contained in the next message. **This is very important.** The user's next message will contain the topic which you will respond to.

Your task is to return structured JSON using the following schema:

- "windows":
   You will generate an array of window objects as a response to the user's prompt. Each window object must include:
   - "title": The title of the window. Do not use "Window 1: ", "Topic 1: ", etc., simply use a title.
   - "content": The window's main content, in markdown format. Be concise and keep the text relevant to what the user asked.
   - "images": An optional array of strings. Each string should be a concise phrase that describes what the generated image should show, including the topic and key visual details. Do not include diagrams or text, only actual photos. You may include 0 to 2 image prompts per window. Only use 2 images for the first window. All other windows must have maximum 1 image. Only output an image if it is fully related to the content. If the image is only tangentially related, simply output nothing.
   You can generate between 1 and 5 windows. Use more windows if the topic is broad and requires multiple aspects, or fewer if the topic is specific and direct.
   For complex topics requiring multiple windows:
      1. Each window should cover a distinct aspect of the topic.
      2. The first window should provide an overview with a title matching the overall topic (do not include the word "Overview") and contain ample information.
      3. All subsequent windows should focus on specific, separate facets of the topic.
      4. If the topic does not allow a clear breakdown into aspects, simply return a single, concise window.
   Aim for clarity by being brief without omitting necessary details, and use bulleted lists and emojis where appropriate.

- "overview":
   Provide a very brief overview in natural, casual language (no more than 2 sentences) that summarizes what the user is looking at. Avoid any extra language.
```

This prompt was submitted to GPT-4.1 via OpenAI's API for additional context during a follow-up:

```
Now, your task is to return additional structured JSON to *add* to the windows you have already generated. Try to avoid repeating information from other windows you have already generated.
Simply return:
- "windows": Additional window object(s), in the same format as before, to address the user's additional prompt. Aim to be concise. You should not include more information than what the user asked for. If the prompt is very broad and your response will be long, use multiple windows. If the prompt is very specific and a short, direct response will suffice, use one window.
```

Received 27 April 2025