Interactive Visualization for Deep Research Prompt Refinement

Mauricio Villavicencio, Christopher Hardwick, Patricia Setiani, Anurag Purohit
Team Adamnn
villa669@umn.edu, hardw050@umn.edu, setia033@umn.edu, puroh029@umn.edu

Abstract

Researchers increasingly use ChatGPT's Deep Research to explore literature and form hypotheses, yet many struggle to craft prompts that balance breadth and depth - broad prompts yield shallow overviews, while narrow ones miss relevant perspectives. We propose an interactive visualization that scaffolds prompt refinement for Deep Research. The tool classifies user intent, flags missing constraints (e.g., evidence, scope, comparisons), and provides diagnostic signals about prompt scope and specificity while tracking how edits shift predicted response quality (e.g., coherence, relevance, and coverage). A small open source model powers critique and dual prompt rewrites; users then send only the finalized prompt to Deep Research. We evaluated the system in a within-subjects user study (N = 11), measuring perceived quality gains, prompt refinement efficiency, and user understanding of prompt effects. Our goal is to make prompt engineering transparent and to help researchers obtain deeper, more targeted results with fewer costly runs.

1 Introduction

Large Language Models (LLMs) like ChatGPT's Deep Research feature are increasingly being used by interdisciplinary researchers to explore scientific topics, synthesize literature, and generate new hypotheses. However, researchers struggle to craft prompts that elicit deep, coherent, and contextually relevant responses. Many struggle to specify scope, constraints, and intent when writing prompts. Broad prompts often return unfocused summaries, while overly specific prompts can unintentionally exclude relevant perspectives.

For this study, Deep Research refers to an advanced mode of generative AI that functions as an automated research assistant. It conducts multi-step exploration across online sources and synthesizes findings into structured, evidence-based reports. This functionality has recently been introduced in several LLM platforms, including ChatGPT, Gemini, Perplexity, and Copilot. Unlike regular LLM interactions, which focus on quick brainstorming or short question-answering responses, Deep Research supports comprehensive, fact-based investigation for complex research questions.

This gap is a serious challenge between model capability and user understanding. Even powerful LLMs depend heavily on prompt quality, yet users receive limited feedback on how their wording affects model reasoning. Current Deep Research interfaces provide text-based outputs but lack interactive guidance or visualization to help users iteratively refine their prompts.

To address this, our project proposes an interactive visualization tool that guides researchers through the prompt refinement process. The system visualizes how prompt edits relate to predicted response quality and missing requirements, helping users understand how prompt structure influences Deep Research outputs. By making the prompt engineering process transparent and interactive, we aim to enhance both the effectiveness and interpretability of Deep Research, ultimately empowering researchers to use LLMs more efficiently for scholarly exploration.

This work is intended for researchers, students, and practitioners who rely on LLM-based literature exploration tools and face time, cost, or iteration constraints when refining prompts for complex research questions.

In summary, our contributions are threefold: (1) an interactive, pre-run visualization system designed specifically for Deep Research prompt refinement; (2) a design that emphasizes prompt diagnosis, requirement awareness, and versioning rather than rapid prompt iteration; and (3) a mixed-methods user study demonstrating improved perceived success, reduced prompt-writing friction, and greater user understanding of how prompt structure affects Deep Research outputs.

2 Related work

Our project addresses related work both in understanding how researchers use the "deep research" functionality of LLMs, and in prompt engineering/fine tuning through visualization:

2.1 Human training / scaffolding for prompting

Breadth vs. Depth (Hayati, 2025): A paper that is currently under submission details how researchers use the deep research function on LLMs and some of its limitations. The key finding of their research is that the results of deep research are very prompt dependent. Their work focused on interdisciplinary research and highlighted the importance of prompt engineering to control the amount of breadth, depth, and connectedness of the different areas of research. They found that the deep research function excelled at providing broad responses related to their question but required specific prompts to achieve the depth the researchers were looking for. They also noted the fact that the deep research functionality often includes papers that are not relevant to the question at hand and requires special prompting to exclude these papers. They recommend that "the tool could go further by recommending relevant keywords, highlighting preferred vocabularies for prompting, or offering guidance on how to frame effective prompts" which directly motivates the design of our system.

What Should We Engineer in Prompts? (Ma et al., 2025) introduced Requirement-Oriented Prompt Engineering (ROPE), showing that effective prompting hinges on clearly articulated, checkable requirements (scope, evidence, format, constraints) before execution - rather than hacks like role-play or "think step-by-step." In a randomized study with 30 novices, ROPE's training/assessment suite with LLM feedback delivered ∼20% gains vs. ∼1% for conventional training, with requirement clarity correlating with output quality. This supports our pre-run prompt linting and constraint badges for Deep Research.

(Schulhoff et al., 2025) present The Prompt Report, a comprehensive survey that standardizes prompt-engineering terminology and techniques: a 33-term vocabulary, a taxonomy of 58 text LLM prompting techniques and 40 multimodal techniques, plus best-practice guidelines for SOTA models. By consolidating a fragmented literature via meta-analysis, it offers a clear map of the design space. For our project, this taxonomy grounds our pre-run checks (e.g., requirement setting, few-shot patterns) and helps justify which prompt strategies we visualize and recommend.

2.2 Visualization Systems for Prompt Engineering

ChainForge (Arawjo et al., 2024) allows users to rapidly evaluate different prompts with different wordings and on different LLMs. Their system is intended to be used with various well-defined prompting goals such as sentiment classification, text summarization, or classification. Their system allows users to create "flows" and allows them to insert tabular data into the system. The system evaluates the different elements of the table against different prompts against different models which allows for rapid output of different combinations in order for the user to find the best prompt.

PromptIDE (Strobelt et al., 2022) intends to simplify the workflow for benchmark LLM tasks by allowing the user to upload data and then choose from a variety of prompt options and then test those prompts on subsets of the data. This allows the user to rapidly iterate through prompts to find the best option for their task.

EvalLM (Kim et al., 2024) expands on the concepts of by introducing automatic evaluation of prompts. Like the previously mentioned systems, a user inputs data and an example prompt which the system then iteratively comes up with new prompts to test. However in this system, the success of a prompt is determined by a criteria that the user entered that is then tested with another LLM. They did a **comparative study** (N=12) that showed that EvalLM, when compared to manual evaluation, helped participants compose more diverse criteria, examine twice as many outputs, and reach satisfactory prompts with 59% fewer revisions. This inspires our post-run quality scoring.

These works evaluate prompt quality via rubric ratings, coherence or relevance metrics, and user studies - approaches we also employ. All these examples comprise an easy to use interface that allows the user to interact with the system even if they don't have a computer science background.

Unlike prior prompt-engineering systems, our tool is explicitly designed for long-running, high-cost LLM workflows such as Deep Research. Rather than enabling rapid prompt comparison, we emphasize pre-run diagnosis, requirement awareness, and prompt versioning to reduce unnecessary executions.

3 Visualization Development

3.1 Visualization Goals

Our visualization expands on prior work in prompt refinement by developing a system specifically designed for *Deep Research* workflows. Existing prompt-engineering systems typically rely on near-instantaneous LLM feedback, which is unsuitable for Deep Research settings where executions can take several minutes and are often limited by cost or usage constraints. As a result, users receive little guidance before committing to a long-running query.

Our goal is to support *pre-run* prompt refinement by helping users understand prompt structure, scope, and requirements before executing Deep Research. The system surfaces missing constraints, ambiguities, and scope issues, highlights opportunities to balance depth and breadth, and suggests concrete prompt edits that are likely to improve downstream research outputs. Rather than generating research content, the system focuses exclusively on diagnosing prompt quality and structure.

To accomplish this, we perform semantic analysis of the user's prompt using a lightweight, deterministic LLM (**LLaMA-3.3-70B-Instruct**) served via the Fireworks API. The model is constrained to return structured analyses, including inferred intent, detected and missing constraints, diagnostic quality scores, and prompt-level suggestions, which are then mapped directly to interactive visual encodings in the interface. These visualizations allow users to iteratively refine prompts while observing how edits shift predicted response quality.
Key components:

- **Prompt Tracking:** Maintain a chronological history of prompt versions to support iterative refinement and comparison.
- **Response Evaluation:** Compute diagnostic quality signals (Depth, Breadth, Coherence, Relevance) using lightweight LLM-based scoring, enabling users to reason about prompt trade-offs before execution.
- **Interactive Visualization:** Visually encode quality metrics, constraint presence, and version-to-version changes to make prompt effects transparent and interpretable.
- **Prompt Classification:** Infer high-level prompt intent (exploratory, analytical, comparative, brainstorming) to contextualize metrics and guide refinement.

3.2 Initial Versions

We originally proposed a graph model, where users prompts and LLM suggestions would be treated as nodes, and different prompt criteria would be treated as edges. The initial version of this graph was included in Figure 1. A graph visualization was chosen because we wanted to show the relationship between prompts, how a certain prompt evolves over time, and how the same suggestion can affect multiple prompts. We included a sidebar of the lefthand side of the visualization to allow the user to input different prompts, and then see how well those prompts performed based on different evaluation metrics such as depth, breadth, and coherence. We also displayed detected prompt constraints (e.g., evidence requirements, scope limits, exclusions) alongside written suggestions for improvement

After implementing this initial version we realized that it fell short in some ways. First, prompt refinement is usually a linear process, which didn't suit our graph interface which displays all the prompt iterations and nodes connecting them at once. Second, since our graph displayed all the prompts and all of their suggestions at once, the interface became overcrowded very quickly. Third, the metrics on the left were only available for the last inputted prompt, which doesn't allow the user to see how those metrics changed over time.

3.3 Final Visualization Design

Iterating on our initial graph-based prototype, we developed a refined interface centered on prompt quality signals, constraint awareness, and version-by-version comparison (Figure 2). The final design reflects the observation that prompt refinement for research is typically a *linear, incremental process*, rather than a branching exploration.

Prompt workspace and versioning. Users enter a prompt in a persistent left-hand workspace and trigger analysis via an *Analyze Prompt* action. Each submitted prompt becomes a new version in a chronological history. Versions can be navigated using both a timeline view and a compact linear prompt graph, enabling users to move backward and forward through iterations without losing context.

Metric visualization with deltas. For the selected prompt version, four quality metrics, Depth, Breadth, Coherence, and Relevance, are displayed as horizontal bar charts normalized to $[0, 100]$. To support comparative reasoning, each metric also displays a signed delta relative to the immediately preceding version, visually indicating whether a specific edit improved or degraded a quality dimension. This allows users to see, for example, when increasing scope improves breadth at the expense of coherence.

Constraint badges. Detected constraints are shown as discrete visual badges (e.g., Evidence, Scope, Comparisons), using binary encodings to indicate whether each requirement is present. These badges help users quickly identify missing structural elements that are known to affect Deep Research outcomes, such as lack of evidence specification or unclear comparison targets.

Actionable refinement suggestions. Below the metrics, the interface presents a ranked list of concrete, prompt-level suggestions generated by the analysis model. Suggestions are phrased as direct edits (e.g., "Specify a time range" or "Define comparison criteria") rather than abstract advice. These suggestions are also summarized in the prompt graph, allowing users to track how recommendations evolve across versions.

Prompt similarity and best-version cues. To further support reflective refinement, the system computes similarity between prompt versions based on their metric vectors and highlights the most similar prior version. In addition, an aggregate quality score (computed as the mean of the four metrics) is used to flag the highest-scoring version so far. These cues function as lightweight visual anchors rather than optimization targets, encouraging exploration while discouraging blind metric chasing.

Overall, the final visualization emphasizes transparency and interpretability: users can see *what* changed between prompt versions, *why* quality signals shifted, and *how* specific edits influence predicted Deep Research behavior—all before executing a costly run.

4 Evaluation

4.1 User Study Design

To evaluate the effectiveness of our study we conducted a user study with 11 university researchers, each who were familiar with LLM tools. Each participant completed three research tasks with the same goal:

Table 1: Participant demographics and familiarity. DR Fam. denotes familiarity with Deep Research tools, and Prompt Conf. denotes confidence in crafting effective prompts (both on a 1–10 Likert scale).

ID	Academic	Major	DR Fam.	Prompt Conf.
P1	UG	CS	3	4
P2	MS	Robotics	3	5
P3	MS	CS	7	10
P4	MS	CS	6	8
P5	MS	ECE	7	7
P6	UG	CS	3	7
P7	PhD	CS	4	4
P8	UG	Economics	5	5
P9	MS	DS	1	8
P10	UG	CS	1	7
P11	UG	CS	2	2

1. **Baseline:** Use Deep Research without assistance.
2. **Tool Assistance:** Refine prompts using our visualization tool before executing Deep Research.
3. **Base LLM comparison:** Prompt the standard ChatGPT interface without Deep Research.

Each session includes a pre-survey, a 30–40 min task interaction, and post-survey evaluating the tool's usefulness, interpretability, and perceived effectiveness. Metrics: output success (self-rated), rubric scores (depth, coherence, relevance), and qualitative feedback.

4.2 Data Collection and Analysis

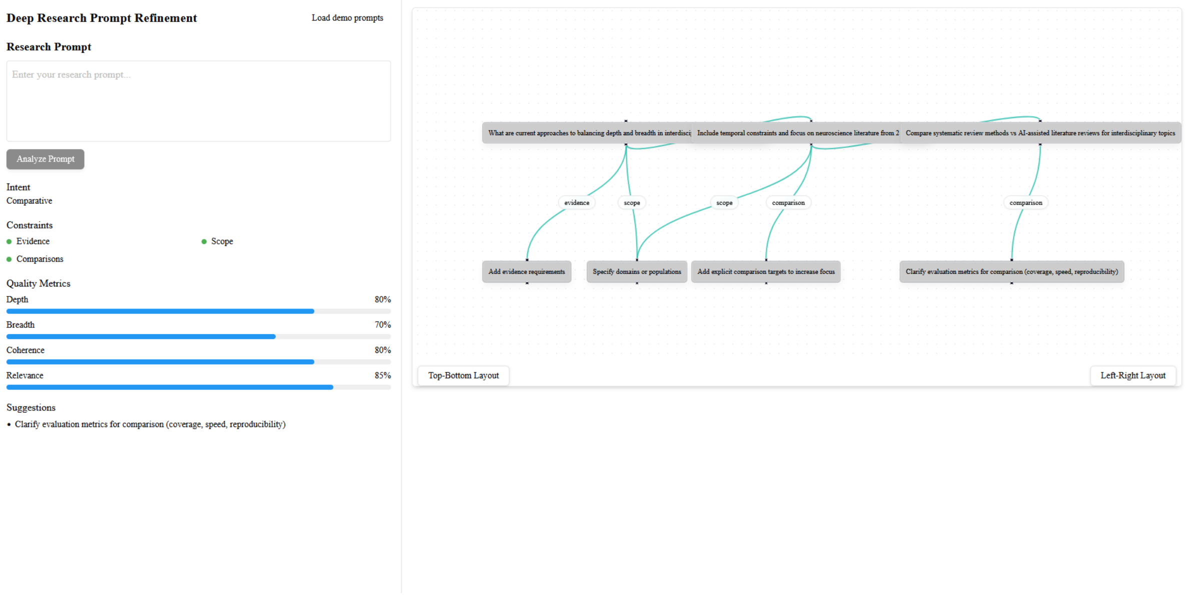
We collected:

- Prompt–response logs before and after refinement
- Quantitative ratings of success, satisfaction, and perceived improvement
- Qualitative feedback from post-interaction questions

We analyzed results to assess whether visualization-based feedback improved users' ability to refine prompts and reduced trial-and-error.

4.3 Post-Survey Measures

We focus on the following aspects in the post-survey questionnaire:



Deep Research Prompt Refinement

Research Prompt

Intent
Comparative

Constraints
- Evidence
- Scope
- Comparisons

Quality Metrics

Depth	80%
Breadth	70%
Coherence	80%
Relevance	85%

Suggestions
- Clarify evaluation metrics for comparison (coverage, speed, reproducibility)

What are current approaches to balancing depth and breadth in interdisci

Include temporal constraints and focus on neuroscience literature from 2

Compare systematic review methods vs AI-assisted literature reviews for interdisciplinary topics

Add evidence requirements

Specify domains or populations

Add explicit comparison targets to increase focus

Clarify evaluation metrics for comparison (coverage, speed, reproducibility)

Figure 1: Initial graph-based interface.

- Quantitative (scale 1-10): success of prompt results with and without our tool, satisfaction/usefulness of the tool, role of our tool in understanding prompt effects on model response, and the ease of using our tool for prompt-refinement
- Qualitative (textual): we asked users to expand on their thoughts about the usefulness of metrics, any other metrics they would like to see in the future, the aspects of the visual interface that were helpful beyond the metrics and suggestions for improvement.

5 Results

Across participants, quantitative ratings indicated higher perceived success when using our tool, along with high satisfaction and perceived usefulness. Qualitative feedback revealed consistent appreciation for the tool's suggestions and prompt-version visualization, and recurring requests for clearer metric explainability and improved editing/interaction flow.

5.1 Positive User Feedback

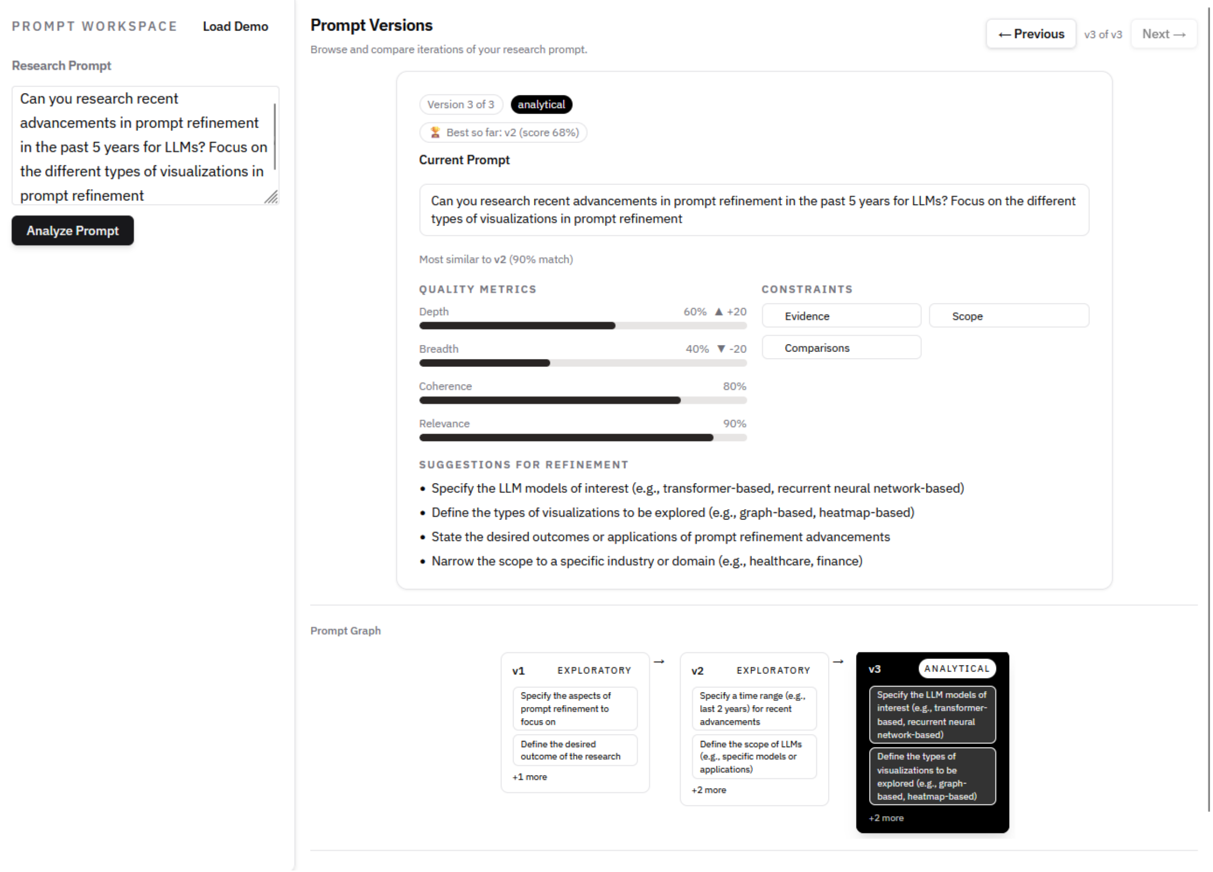
Higher perceived success with tool-assisted prompting. Participants reported greater success in achieving their research goals when using the tool (8.9 ± 1.1) compared to the baseline Deep Research condition (6.8 ± 1.5), reflecting an average improvement of +2.1 points on the 1–10 scale. Most participants also indicated strong satisfaction with the overall tool-assisted workflow (8.7 ± 1.4).

Perceived usefulness for both improvement and understanding. Participants rated the tool as useful for improving their Deep Research prompt (8.6 ± 1.2) and for understanding how prompt changes affect model responses (8.6 ± 1.3). In addition, 9/11 participants answered "Yes" when asked whether they would use a similar tool in future research tasks (the remaining 2 answered "Maybe").

Reduced friction in prompt writing. Participants reported that crafting prompts felt easier with the tool (ease: 7.6 ± 1.6) than without it (ease: 6.4 ± 1.6). Reported mental demand also decreased from 5.8 ± 2.0 (baseline) to 4.7 ± 2.3 (with tool), suggesting that the visualization and suggestions reduced cognitive load for some users.

Qualitative themes: suggestions and prompt versioning were especially valued. We systematically analyzed open-ended feedback to identify feature usage and preference patterns. The *suggestions* feature was mentioned as most helpful by 8/11 participants, with comments describing them as "straightforward" (P3), offering "valid suggestions" (P1), and providing "ways to change my prompt" (P10). The prompt versioning graph was explicitly valued by 3/11 participants for "version control" (P4) and comparing "different prompts between each other" (P10).

Regarding metrics, Depth and Breadth were most frequently cited as useful (7/11 and 6/11 par-



PROMPT WORKSPACE Load Demo

Research Prompt

Can you research recent advancements in prompt refinement in the past 5 years for LLMs? Focus on the different types of visualizations in prompt refinement

Analyze Prompt

Prompt Versions

Browse and compare iterations of your research prompt.

← Previous v3 of v3 Next →

Version 3 of 3 analytical

🏆 Best so far: v2 (score 68%)

Current Prompt

Can you research recent advancements in prompt refinement in the past 5 years for LLMs? Focus on the different types of visualizations in prompt refinement

Most similar to v2 (90% match)

QUALITY METRICS

Metric	Score	Change
Depth	60%	▲ +20
Breadth	40%	▼ -20
Coherence	80%	
Relevance	90%	

CONSTRAINTS

Evidence

Scope

Comparisons

SUGGESTIONS FOR REFINEMENT

- Specify the LLM models of interest (e.g., transformer-based, recurrent neural network-based)
- Define the types of visualizations to be explored (e.g., graph-based, heatmap-based)
- State the desired outcomes or applications of prompt refinement advancements
- Narrow the scope to a specific industry or domain (e.g., healthcare, finance)

Prompt Graph

v1 EXPLORATORY
- Specify the aspects of prompt refinement to focus on
- Define the desired outcome of the research
- +1 more

→

v2 EXPLORATORY
- Specify a time range (e.g., last 2 years) for recent advancements
- Define the scope of LLMs (e.g., specific models or applications)
- +2 more

→

v3 ANALYTICAL
- Specify the LLM models of interest (e.g., transformer-based, recurrent neural network-based)
- Define the types of visualizations to be explored (e.g., graph-based, heatmap-based)
- +2 more

Figure 2: Final visualization interface.

ticipants respectively), with one noting they "gave good indication of how specific a topic I want to research" (P1). Relevance was mentioned by 4/11 participants. However, 2/11 participants explicitly stated they did not rely on metrics, with one noting: "I didn't use the metrics explicitly but I used the suggestions" (P7). This indicates most participants (7/11) used suggestions as their primary refinement mechanism, with metrics serving as secondary validation.

5.2 Representative Refinement Patterns

To contextualize our quantitative findings, we describe representative patterns observed across participants.

Success pattern – Constraint addition. Multiple participants reported that suggestions helped them add missing specificity to initially broad prompts. One stated suggestions "changed my prompts in a way that made them more effective" (P6), with improved metrics described as "a result" of following actionable suggestions rather than the primary driver of refinement.

Success pattern – Version comparison and iteration tracking. Beyond individual suggestions, participants valued the systematic comparison capability. One stated: "I liked how easy it was to compare my different prompts between each other and seeing the difference in my descriptiveness" (P10), noting this helped track refinement trajectory and avoid unproductive iterations.

Failure pattern – Metric opacity without attribution. The most common complaint (4/11 participants) was inability to connect metric changes to specific prompt elements. Representative quotes include: "I don't know which part is corresponding to which part of my prompt (like how can I improve the depth part)" (P2), and "it would be interesting if the tool was able to determine what parts of the prompt affect the 4 different metrics" (P10). One participant expressed confusion when deleting examples caused large score changes: "I don't think examples matter so much" (P2), suggesting the model's reasoning was opaque to users. Future work should implement span-level attribution highlighting which prompt phrases drive each metric score.

Measure (1–10)	Mean ± SD
Success (baseline, no tool)	6.8±1.5
Success (with tool)	8.9±1.1
Satisfaction (with tool)	8.7±1.4
Usefulness for improving prompt (with tool)	8.6±1.2
Helpfulness for understanding prompt effects (with tool)	8.6±1.3
Ease of crafting (baseline)	6.4±1.6
Ease of crafting (with tool)	7.6±1.6
Mental demand (baseline)	5.8±2.0
Mental demand (with tool)	4.7±2.3

Table 2: Post-survey quantitative results (N=11). Higher is better for success/satisfaction/usefulness/ease; higher indicates more mental demand for mental-demand items.

Failure pattern – Intent misalignment toward specificity. Two participants reported the tool systematically pushed them toward narrow focus when they wanted breadth. Most clearly, P7 stated: "I felt that it was always pushing me to choose one or the other goals...even after specifying that I want a broader overview," and suggested investigating "whether there is any pattern that your tool is trying to steer the prompt towards." Another noted their research was "more open ended" and preferred breadth-oriented guidance (P5). This indicates the suggestion algorithm may have a systematic bias toward depth over breadth, requiring intent-aware generation with explicit breadth/depth parameters to respect user goals.

Failure pattern – Inconsistent suggestion quality. While suggestions were generally valued, some participants found them difficult to act on without additional context. One noted: "some are useful (they can help me to make my prompt more specific), but some are not (I cannot answer the question it gives)" (P2). The inconsistency in providing examples was also noted: "sometimes it gave examples (e.g.), but sometimes not" (P2). Two participants explicitly requested examples alongside suggestions to clarify what changes were needed (P2, P7). Providing template-based suggestions with concrete examples for each recommendation would improve actionability and consistency.

5.3 Areas for Improvement

Improve metric interpretability via attribution and clearer explanations. A recurring request was to make metric scores more explainable at the sentence/phrase level (e.g., indicating which part of the prompt caused Depth to decrease). Similarly, at least one participant reported that the metrics felt less useful without guidance on how to translate the scores into concrete edits. These comments suggest that adding lightweight attribution (highlighting spans that drive a score) and brief "what this score means" tooltips could improve interpretability.

Automatic prompt rewriting. Three participants requested the tool generate complete revised prompts rather than just suggestions. Requests included: "AI generate a whole new prompt based the suggestions" (P8) and "if the tool can automatically refine the prompt for me (or maybe provide more examples about a better prompt) would be great" (P11). This represents a desire for more direct assistance in the refinement process.

Onboarding and workflow clarity. Participants requested a short tutorial, popup, or walkthrough to explain the intended workflow. One specifically suggested adding "a tutorial or a popup to explain how it works and what to focus on" (P4). Additional clarity was needed for interface elements: one participant noted the constraint badges were confusing because "every time it is the same so I don't know what that mean" and wanted to "expand" the workflow visualization (P2). This aligns with the tool's goal of acting as a learning scaffold: users want clearer cues about how to use metrics and suggestions effectively.

UI/interaction improvements: editing and navigation. The most common UI complaint was the prompt editing experience: "Easier way to edit the text, cant see everything on the left small box" (P1). Additional requests included improving the organization and clarity of the prompt workflow display and making the visualization easier to manipulate.

Make suggestions more controllable and offer "one-click" rewrites. Participants suggested providing examples alongside suggestions, and some requested the ability for the tool to automatically generate a complete revised prompt based on selected suggestions. Another theme was controllability: users wanted the tool to better respect high-level intent (e.g., producing broader coverage when requested), motivating tighter alignment between intent detection (depth vs. breadth preference) and the final refined prompt generation.

5.4 Ethical considerations and responsible use

Because the system operates strictly at the prompt level and does not generate or modify research content, it reduces risks related to hallucination amplification, misattribution, or unintentional plagiarism. However, users may over-trust numeric quality signals without fully understanding their limitations, motivating future work on clearer uncertainty communication and metric transparency.

6 Conclusion

We presented an interactive visualization system for pre-run prompt refinement in Deep Research workflows. By surfacing missing constraints, diagnostic quality signals, and prompt versioning, our system helps users understand how prompt structure influences long-running, high-cost LLM outputs before execution.

In a within-subjects user study with university researchers, participants reported higher perceived success, reduced cognitive effort, and improved understanding of prompt effects when using our tool compared to unaided Deep Research. Qualitative feedback highlighted the value of actionable suggestions and prompt versioning, while also motivating future improvements in metric explainability, onboarding, and interaction design.

Overall, our findings suggest that pre-run prompt diagnosis and visualization can meaningfully reduce trial-and-error in Deep Research workflows. Future work includes improving attribution for quality signals, supporting more controllable prompt rewrites, and evaluating the system with broader researcher populations and objective outcome measures. All prompts, survey instruments, and evaluation procedures used in this study are available in the accompanying GitHub repository.

References

Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. Chainforge: A visual toolkit for prompt engineering and llm hypothesis testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–18. ACM.

Shirley Anugrah Hayati. 2025. Breadth vs. depth: Orchestrating deep research in interdisciplinary knowledge discovery (under submision).

Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–21. ACM.

Qianou Ma, Weirui Peng, Chenyang Yang, Hua Shen, Ken Koedinger, and Tongshuang Wu. 2025. What should we engineer in prompts? training humans in requirement-driven llm use. *ACM Transactions on Computer-Human Interaction*, 32(4):1–27.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2025. The prompt report: A systematic survey of prompt engineering techniques.

Hendrik Strobelt, Albert Webson, Victor Sanh, Benjamin Hoover, Johanna Beyer, Hanspeter Pfister, and Alexander M. Rush. 2022. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models.

A Pre-Study Survey

This is the pre-study survey for our Deep Research prompt visualization study. Participants were asked to answer based on their background and prior experience with LLM tools and Deep Research.

Demographics

1. What is your name?
2. What is your gender?
   - Female
   - Male
   - Non-binary
3. What is your age range?
   - 18–25
   - 26–35
   - 36–45
   - 45+
4. What is your academic level?
   - Undergraduate student

- Master's student
- Junior PhD student (year 1–3)
- Senior PhD student (year 4+)
- Other

5. What is your major? (e.g., computer science, data science, educational psychology, etc.)

6. What is your current research topic?

Experience with LLM Tools

Participants were asked to rate the following on a 1–10 scale:

- How familiar are you with using LLM tools? (LLM tools include ChatGPT, Google Gemini, Microsoft Copilot, etc.; 1 = Not at all familiar, 10 = Very familiar)

- How often do you use LLM tools for research? (1 = Never, 10 = Very often)

- How familiar are you with Deep Research? (If never used Deep Research, select 1; 1 = Not at all familiar, 10 = Very familiar)

- How confident are you in crafting effective prompts? (1 = Not confident, 10 = Very confident)

B Post-Study Survey

This is the post-study survey for our Deep Research prompt visualization study. Participants were asked to answer based on their experience with and without the visualization tool during the session.

General Questions

1. Name

2. How successful were you in achieving your research goal without the visualization tool?

3. How successful were you in achieving your research goal with the visualization tool?

Usefulness & Perceived Impact

Participants were asked to rate the following on a 1–10 scale:

- How satisfied were you with the overall process when using the visualization tool? (1 = Not satisfied, 10 = Very satisfied)

- How useful was the visualization tool for improving your Deep Research prompt? (1 = Not useful, 10 = Very useful)

- How helpful was the visualization tool in understanding how your prompt affects the model's response? (1 = Not helpful, 10 = Very helpful)

- Would you use a similar tool in your future research tasks? (Yes / No / Maybe)

Cognitive Load & Usability

Participants rated the following on a 1–10 scale:

- How easy or difficult was it to craft your Deep Research prompt without the visualization tool? (1 = Very difficult, 10 = Very easy)

- How mentally demanding was the prompt-writing process without the visualization tool? (1 = Not demanding, 10 = Very demanding)

- How easy or difficult was it to craft your Deep Research prompt with the visualization tool? (1 = Very difficult, 10 = Very easy)

- How mentally demanding was the prompt-writing process with the visualization tool? (1 = Not demanding, 10 = Very demanding)

Open-Ended Reflection

1. Which quality metric(s) in the visualization did you find most useful for your research task? Why?

2. Are there any other metrics you wish had been included for evaluating your prompt?

3. What aspects of the visualization tool did you find most helpful or intuitive (beyond the metrics)?

4. What suggestions do you have for improving the tool?