

Generative AI in Knowledge Work: Design Implications for Data Navigation and Decision-Making

Bhada Yun*
University of California, Berkeley
Berkeley, California, USA
bhadayun@berkeley.edu

Dana Feng*
University of California, Berkeley
Berkeley, California, USA
danafeng@berkeley.edu

Ace S. Chen
University of California, Berkeley
Berkeley, California, USA
12aschen@berkeley.edu

Afshin Nikzad
University of Southern California
Los Angeles, California, USA
afshinni@usc.edu

Niloufar Salehi
University of California, Berkeley
Berkeley, California, USA
nsalehi@berkeley.edu

Abstract

Our study of 20 knowledge workers revealed a common challenge: the difficulty of synthesizing unstructured information scattered across multiple platforms to make informed decisions. Drawing on their vision of an ideal knowledge synthesis tool, we developed Yodeai, an AI-enabled system, to explore both the opportunities and limitations of AI in knowledge work. Through a user study with 16 product managers, we identified three key requirements for Generative AI in knowledge work: adaptable user control, transparent collaboration mechanisms, and the ability to integrate background knowledge with external information. However, we also found significant limitations, including overreliance on AI, user isolation, and contextual factors outside the AI's reach. As AI tools become increasingly prevalent in professional settings, we propose design principles that emphasize adaptability to diverse workflows, accountability in personal and collaborative contexts, and context-aware interoperability to guide the development of human-centered AI systems for product managers and knowledge workers.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in interaction design*.

Keywords

Knowledge Synthesis, Information Visualization, Human-AI Interaction, Large Language Models, Interaction Design

ACM Reference Format:

Bhada Yun, Dana Feng, Ace S. Chen, Afshin Nikzad, and Niloufar Salehi. 2025. Generative AI in Knowledge Work: Design Implications for Data Navigation and Decision-Making. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3706598.3713337>

*Yun and Feng contributed equally to this work (order was decided by coin toss).



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '25, Yokohama, Japan*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1394-1/25/04

<https://doi.org/10.1145/3706598.3713337>

1 Introduction

“Managers are not confronted with problems that are independent of each other, but with dynamic situations that consist of complex systems of changing problems that interact with each other. I call such situations, messes.”

- Russell Ackoff, 1979

According to some estimates, 80% of the data in organizations is messy, unstructured, and rarely used [76]. Knowledge workers face challenges in handling the scale and complexity of unstructured data, which leads to difficulties in extracting reliable information and making informed decisions [3, 32, 54, 55, 88, 89]. AI-based tools have the potential to address these challenges by automatically structuring and synthesizing information [62, 101]. However, AI models are not well equipped to take over the entire data analysis process, as it is often an iterative and collaborative process that relies on the domain knowledge of the decision maker [78]. Furthermore, the potential for hallucinations in the model [42] and biases [82] raise concerns about the validity of purely AI-generated outputs. As companies increasingly integrate AI into their employees' workflows, it is important to understand what opportunities and limitations generative AI poses for data navigation and decision-making in knowledge work.

In this paper, we build an AI-enabled system for product managers (PMs) as a proxy for exploring the experience and opinions of knowledge workers with generative AI. Prior literature has defined the use of proxy systems as utilizing existing technologies to ideate for future technologies [66]; we follow a similar pattern, developing an AI system to glean principles for future knowledge work tools. Our formative study with 20 knowledge workers, including PMs, consultants, journalists, and graduate students, revealed three critical challenges in their work: searching and synthesizing vast amounts of unstructured data, extracting reliable insights, and maintaining effective collaboration. Knowledge workers struggle to synthesize information from a variety of sources, ranging from user interviews to research articles [18, 26]. This challenge is compounded by the tacit nature of domain expertise required to effectively analyze and compare data [12, 39]. We found that successful AI tools for human AI collaborative knowledge synthesis must integrate work context, personal domain knowledge, and the iterative nature of decision-making [64].

We designed Yodeai to integrate into existing knowledge workflows by enabling knowledge workers to explore and synthesize unstructured data (e.g., product reviews and customer interview transcripts) through interactive widgets, customizable workflow templates that transform unstructured data into structured insights. To empirically examine knowledge workers' experiences with AI-enabled tools, we conducted an 80-minute study with a group of 16 product managers. Participants role-played as PMs at Notion, a software company, analyzing two data collections (12 YouTube interview transcripts and 132 App Store reviews) using three interactive widgets (User Insights, Pain Point Tracker, and Q&A). The study involved a divergent phase where participants brainstormed six new feature or bug fix proposals, followed by a convergent phase where they prioritized their top three proposals. Over 20 minutes of system onboarding, 25 minutes of task work, and pre/post task interviews and surveys, we gathered rich insights about knowledge workers' interactions with and perspectives on AI-enabled knowledge synthesis.

Our results show a need for AI knowledge work tools to support diverse data exploration and decision-making workflows, enabling participants to leverage tool components in unique ways that suit their individual needs and preferences. With Yodeai, we show that the flexibility of the system allowed users to integrate their own domain expertise with the AI's outputs, as well as verify their own and others' work. However, our findings also highlighted potential challenges, such as overreliance on automation, navigating stale data, and privacy concerns.

Based on these findings, we propose three key design implications for developing effective and adaptable AI interactions for complex knowledge work: adaptability, accountability, and interoperability. Our work reveals a deeper understanding of the use of AI in knowledge work, as well as the challenges and opportunities that AI presents. To summarize, our contributions are the following:

- Specific challenges knowledge workers face in navigating data through organization and structure, collaboration in terms of context and integrations, and decision making with prioritization and search.
- Empirical evaluation of the opportunities and limitations of generative AI for complex knowledge work with Yodeai, an AI tool that supports data exploration, verification, prioritization, and collaboration.
- Design implications for AI in knowledge work that depends on understanding and reasoning over data, such as product management (Table 3), emphasizing adaptability to diverse workflows, personal and collaborative accountability, and context-aware interoperability.

2 Background And Related Work

2.1 AI use in Knowledge Work

The integration of AI, such as generative models, has led to productivity gains in multiple domains of knowledge work [10, 20], including management consulting [17], software development [67, 85], and writing [57]. AI can assist writers with tasks such as planning, translation, and proofreading [22], while UX designers benefit from AI-generated design briefs and filler text for mockups [45]. In the medical field, AI has been applied to rehabilitation assessment

[41] and sepsis treatment [71]. A 25-year survey revealed that AI can assist with general human resource management tasks, such as recruitment, interviews, and employee training [65], while in academia, interviews demonstrated the use of LLM in HCI research [36].

In addition to investigating new ways of using generative AI for specific knowledge work tasks, existing literature has also focused on interviews with knowledge workers about their experiences with and perceptions of AI. Research workshops conducted by Woodruff et al. [91] in industries of knowledge work such as advertising, law, and software development revealed that many workers believed that generative AI can automate menial work, but that a human in the loop is needed to ensure a certain standard is met. The workshops also revealed that AI could cause issues of dehumanization, as AI lacks the ability to perform interpersonal work, disconnection, where the work produced lacks social awareness and originality, and disinformation, where hallucinations occur. A large-scale survey conducted by Brachman et al. [8] revealed how knowledge workers utilize LLM in their work, as well as how they would like to use LLMs for creation, information, advice, and automation. Finally, a two-week diary study by Kobiella et al. on young professionals under the age of 30 further emphasizes the issue of disconnect, where workers feel a lack of challenge, ownership and quality in LLM outputs, as well as deskilling, where barriers to information synthesis have been reduced [38].

We build on this work and develop an interactive system, Yodeai, to observe how knowledge workers interact with AI tools to navigate information and make decisions in practice. With this system, we are able to identify practical insights supported by direct hands-on experience, extending beyond studies that only use surveys or visual demonstrations [8, 46, 91]. Through this approach, we contribute design principles for future AI systems in knowledge work.

2.2 AI-Assisted Data Exploration and Collaboration

HCI research has long been invested in studying and developing new models and interactions for information seeking and synthesis [4, 7, 21, 25, 31, 32, 44, 86, 87, 90]. Knowledge workers often rely on external representations such as file systems or information 'boxes' to navigate the vast scale of information at different levels of abstraction [14]. To help understand external and personal data, various interfaces have been developed using LLMs. These interfaces employ a wide range of output formats, from whiteboards [61, 80], hierarchies [35, 80], summarization [15, 48], graphical structures [30, 43], to box structures [14].

Researchers have developed tools to support the navigation of exploratory data, both external (e.g., publicly available information from the Internet) and personal data (e.g., transcribed interviews). Synergi generates LLM-based summaries for papers cited in academic articles [35], while Selenite combines web and LLM-generated content to offer an overview of options and criteria for activities such as shopping [48]. In the realm of personal/collaborative data exploration, numerous LLM tools have been created across disciplines such as data science and writing [15, 24]. Recent developments like GPT plugins [58] and personalized GPTs [59]

have highlighted the challenge of integrating personal data and specialized actions into AI systems. Building on insights from personal LLM agents that handle data analysis, developer APIs, and web integrations [93], our system combines web-sourced information (e.g. Apple App Store reviews) with user's personal files to support comprehensive knowledge synthesis.

Prior research has demonstrated various approaches to help users explore LLM outputs more effectively. For example, Sensecape [80] facilitates non-linear tasks like trip planning through text extraction and semantic exploration features, while ExploreLLM [51] enables structured thinking and exploration of options for tasks such as restaurant discovery. Building on these advances, Yodeai introduces interactive widgets that allow users to engage with LLM outputs through multiple modalities—conversational, visual, and quantitative—to deepen their understanding and exploration of AI-generated insights.

The design of human-AI interactions presents unique challenges due to AI's uncertain capabilities and variable outputs [95]. Researchers have addressed these challenges through specialized LLM frameworks for data exploration, such as decomposing LLM generation into input units, model instances, and output spaces [37], and using graphical models to break down tasks for improved LLM prompting [5]. Recent work has proposed comprehensive frameworks for human-AI collaboration, including multimodal interaction approaches [74], design principles for generative AI applications [83], and guidelines for AI interactions across different stages, from before use, during use, when something goes wrong, and over time [1]. While these frameworks provide valuable theoretical foundations, our work extends beyond theory by implementing a prototype AI system, allowing us to derive concrete design implications from an empirical study with a focused group of knowledge workers that inform the design of more human-centric future knowledge work tools.

2.3 AI Supported Decision-Making and Collaboration

HCI research has produced diverse decision support tools for both personal and professional contexts [68, 69, 98]. Personal tools such as Mesh [11] and ProactiveAgent [52] facilitate faster problem solving through techniques such as comparison tables and task decomposition [47, 97]. In the professional sphere, specialized tools have emerged to address domain-specific needs: ParaLib for library comparison [94], Threddy for research analysis [34], Luminate for creative decision-making [79], and SepsisLab for medical diagnosis [100]. Building on these academic advances and industry solutions like Dovetail's automatic thematic clustering and Productboard's feature request prioritization, we tailored our system specifically for product management workflows.

The effectiveness of AI-assisted decision-making depends on a complex interplay of factors related to the human decision maker, the task at hand, and AI itself [19]. Literature surveys [40] and user behavioral studies [56, 96] have examined human-AI decision-making tasks, evaluation metrics, and the impact of AI partners on human decision-making. This literature emphasizes the nuances of joint decision-making, as user background knowledge can affect the problem-solving process and results [28], with trust and

explainability playing key roles in the adoption of generative AI applications [1]. Our study extends this understanding by examining these dynamics specifically within product management—a domain where decisions often require synthesis of diverse user needs, technical constraints, and business priorities. In doing so, we uncover patterns that apply to knowledge work context where workers must synthesize unstructured information, balance competing priorities, and make evidence-based decisions.

3 Formative Study

We conducted semi-structured remote interviews with 20 participants (E1-E20) located in the United States. To ensure diverse perspectives on knowledge work and information management, we recruited participants across various professional roles: 2 journalists (E1, E18), 5 startup founders (E2, E6, E7, E9, E16), 3 graduate students (E3-E5), 2 consultants (E8, E11), 7 product managers (E10, E12-E15, E19, E20), and 1 associate director (E17). Each interview lasted 45-60 minutes and participants were compensated \$20.

3.1 Interview Protocol Development

Our semi-structured interviews used an exploratory approach to uncover knowledge workers' information management practices, challenges, and needs. The flexible format allowed participants to elaborate on topics most significant to them. For each section outlined below, we prepared a list of potential open-ended questions to select from as the conversation unfolded. The interviews were structured into three main stages:

3.1.1 Information Management. We started each interview by exploring how participants personally manage their information, seeking specific anecdotes and insights into the types of information they worked with as well as the tools they have used. Example questions include:

- Is there a specific type of information (e.g. medical records) that you find particularly challenging to organize?
- Can you tell me about how you organize your information? What tools/apps do you use?
- Have you tried other solutions or methods to organize your information in the past? What worked and what didn't?

3.1.2 Collaboration Dynamics. Next, we explored how their information management practices influenced their work with others, given the collaborative nature of knowledge work. Example questions include:

- How do you share information with others (e.g. colleagues)? Are there challenges with your current method?
- Do you collaborate with others on documents or files? If so, what challenges do you face in this process?
- Can you tell me about times you needed to capture new information quickly? Where do you encounter times like this the most—work/personal/team?

3.1.3 Future Tool Perspectives. The interviews concluded by exploring participants' tool needs as well as their attitudes toward AI assistance. Example questions include:

Table 1: Participant Overview and Background

ID	Professional Experience	Interview Summary
E1	Journalist, 19 years of experience	Working to extract usable data from very large sets of unstructured info. Stuck using Google Drive.
E2	Startup founder. 10 YOE	Loves Notion. Hard to find info from different sources and types, needs a better way to cluster info.
E3	Grad student. 7 YOE	Organizing is effortful. Hard to structure info when working/white-boarding. Likes when a tool has visual elements.
E4	Grad student. 8 YOE	“Wish I could do everything in one place,” but all collaborators have their own methods.
E5	Grad student. Has 7 YOE	Hard to remember where info is stored: “searching for info is like half of my day.” Values interoperability.
E6	Startup founder with 9 YOE	Uses Google Drive because it’s cheap and works. Hard to find things when not aware of what is being searched for.
E7	Startup founder with 10 YOE	Existing tools are unorganized and lack structure. A dream tool would automatically organize and help with search.
E8	Finance, undisclosed YOE	Uses AI for search and brain-dumping, and Heptabase for nonlinear thinking. Wants “little boosts” in workflow.
E9	Data scientist and PM with 7 YOE	Google Drive is great for sharing and collaboration, but lacks a streamlined organizational structure.
E10	PM with 11 YOE	PMs spend a lot of time organizing, but it’s wasted if team members don’t follow through.
E11	HR consultant with 26 YOE	Wants a smarter way with AI to distill and organize information into themes, priorities, insights.
E12	Senior Growth PM with 7 YOE	There are tradeoffs between spending time organizing things and spending time moving things forward.
E13	Platform PM with 5 YOE	Hard to organize information across multiple platforms, tools, and sources and know what’s up to date and useful.
E14	Software dev and PM with 11 YOE	Main challenges are gathering the right kinds of information, and effectively communicating it to stakeholders.
E15	Product manager with 27 YOE	Tools are just a means to an end; it’s important for info to be internalized, not just stored.
E16	Startup founder with 12 YOE	Challenge is to keep track of many different sources and make sure nothing falls through the cracks.
E17	Associate director with 18 YOE	Lack of standard organization makes it difficult to communicate, but things work, so there isn’t urgency to improve.
E18	Investigative journalist with 17 YOE	Need a way to structure information that is 1) at a very large scale and 2) often difficult to access/read (noisy).
E19	Senior AI PM with 17 YOE	Wants help with concrete PM needs (sprint planning, navigating many document types, persona building).
E20	PM and research lead with 12 YOE	Can see many use cases for tools, especially interested in ability for AI to format information into templates.

- Are there certain features or capabilities you wish existed in your current tools that would make organizing information easier for you?
- What would an ideal solution look like for you?
- How do you feel about using AI or automation in helping you organize or find information?

3.2 Data Analysis

To analyze the data, we employed line-by-line coding of each interview transcript, with a primarily top-down approach [6] using predefined categories of “Organization”, “Search”, “Collaboration”, and “Tool Needs and Wants” that was developed after conducting the interviews. Next, we identified common subthemes across the codes, including “information relationships” or “getting context”

using affinity mapping [23]. In the next section, we describe our findings from this formative study that led to the core design thesis of Yodeai and our follow up user study: enabling efficient and collaborative data navigation and decision-making.

3.3 Data Navigation: Organization & Structure

General remarks: Our analysis revealed that the challenges of information organization varied significantly depending on the use of the tool and the context of work. Google Drive users, particularly in startup environments that “move fast” (E7, E9), reported specific difficulties with their information organization systems. These challenges included frequent disorganization due to inconsistent folder structures, unclear file naming conventions, and the tendency to

create individual bloated documents for all meetings, leading to priorities being missed or overlooked (E2, E4, E9). Meanwhile, Notion users, while appreciating the tool's flexibility, noted challenges with feature overwhelm and too many templates (E3). Participants also emphasized that they work with both quantitative (sales metrics) and qualitative data (transcripts) (E13-14), which require different methods of analysis.

Tool wants and needs. Participants voiced the need for a centralized platform that could automatically infer an organizational structure and provide intuitive search functionality (E7, E10, E11, E13). For instance, E2 wanted a tool that could help “organize things while I’m on the go” while E10 shared that they would want a tool that gives snapshots of information and intelligently indexes the information. P20 also noted the importance of standardized templates, and the wish to have AI produce things in template forms as they already use Google suite templates.

3.4 Collaboration: Context & Integrations

General remarks: Corporate knowledge workers (17 out of 20 participants) emphasized frequent information sharing within their companies. While E12 stressed the importance of organized information for team accessibility, E17 highlighted that there is no standard operating procedure on how to share and manage information, and that people do whatever works for them in terms of tool choice. E6 further emphasized the time sink of manual sharing, stating “I’m stuck, I literally need to spend an hour—no—a day and a half sifting through all these docs, tagging people, adding them.” Additionally, E12 noted that a lack of context makes it hard to identify the right stakeholders, while E14 highlighted challenges in communicating results to stakeholders who “don’t speak the same language.”

Tool wants and needs: Participants wanted tools that would allow them to understand the knowledge bases of their colleagues without requiring real-time communication (“offline context-gaining”) (E2). E13 noted that a tool that could pull in notes and meeting transcripts would be very important. Integration with existing platforms (e.g. Salesforce for customer data) was deemed crucial for cross-team collaboration (E3, E5, E7-E9, E13, E20). E6 emphasized the need for a tool that could handle collaboration “when it’s other people’s content that is interlaced with my workflow”.

3.5 Decision Making: Prioritization & Search

General remarks: Participants expressed the need for tools that support the analysis, ideation, and prioritization of data, revealing the various divergent and convergent aspects of decision making (E11-E13, E17, E18). They reported difficulties in finding previously known information (E2, E9, E16). E5 described information search as “half of my day” while E16 explained that “finding info is hard, opening ten documents to find the one thing you were looking for.” Many found keyword search limited, only useful when sufficient information context is known, which can be troublesome when sharing data repositories with others who lack familiarity with the data (E2, E6, E7, E12).

Tool wants and needs: Participants wanted a visualization that helps them understand content relationships (E12) and recognized

the potential in an AI assistant that could help “prioritize ideas, and give hotspots or find themes for me” (E11). They emphasized the need for semantic search, thematic clustering, and information visualization (E2, E9, E12). E12 and E14 expressed the desire for an AI assistant to take notes and generate action items and meeting summaries, while E13 preferred a system capable of independently identifying insights and patterns without relying on user-prompted questions. Importantly, many participants (E6, E9, E10, E15) emphasized showcasing sources, as “people want to be able to know what the authenticity of the document is” (E1).

3.6 Additional Findings

While this paper focuses on the formative study’s interview findings that are most relevant to our system design, our study generated additional insights about domain-specific challenges outside of product management, such as in journalism or startup customer discovery. These findings have informed our broader research agenda and will be explored in future work focusing on enterprise knowledge management practices and AI-assisted collaboration.

4 Yodeai

To investigate generative AI’s potential in collaborative data navigation and decision-making, we developed Yodeai, an AI system focusing on PMs as a representative case of knowledge workers. We chose PMs because they exemplify the complexity of modern knowledge work—they regularly synthesize diverse data sources, collaborate across teams, and make data-driven decisions. Furthermore, our formative study included a large number of PM or PM adjacent experiences.

4.1 Design Process

The design of human-AI interactions presents challenges due to uncertainty in the capabilities and the complexity of the output [95]. Previous studies by Amershi et al. [1] and Weisz et al. [83] used modified heuristic evaluations to create design principles for AI tools, asking participants to choose AI interfaces/tools and evaluate them for applications and violations of their own design guidelines. Building on this approach, we created Yodeai to provide a hands-on experience, allowing us to observe and investigate the interactions of knowledge workers with AI features in real time [46].

Our design process began with a basic chatbot interface, building on PMs’ familiarity with AI tools. To validate and expand our formative study findings, we conducted a public poll on All Our Ideas with 111 PMs found on LinkedIn (Figure 2), asking them to rank and propose features that could improve their PM workflow. The poll surfaced additional use cases like generating Jira tickets and product requirement documents (Figure 3), complementing our formative study insights. While poll respondents showed interest in product document generation and competitive analysis tools, our formative study participants emphasized a more fundamental need for systemizing user research processes. In particular, the poll highlighted both “compile B2B customer feedback” and “extract and summarize user pain points from feedback and reviews”, while participants like E12 wanted tools for theme identification, and E20 expressed interest in a tool for organizing customer research and interrogating data through chat. E10’s observation about the

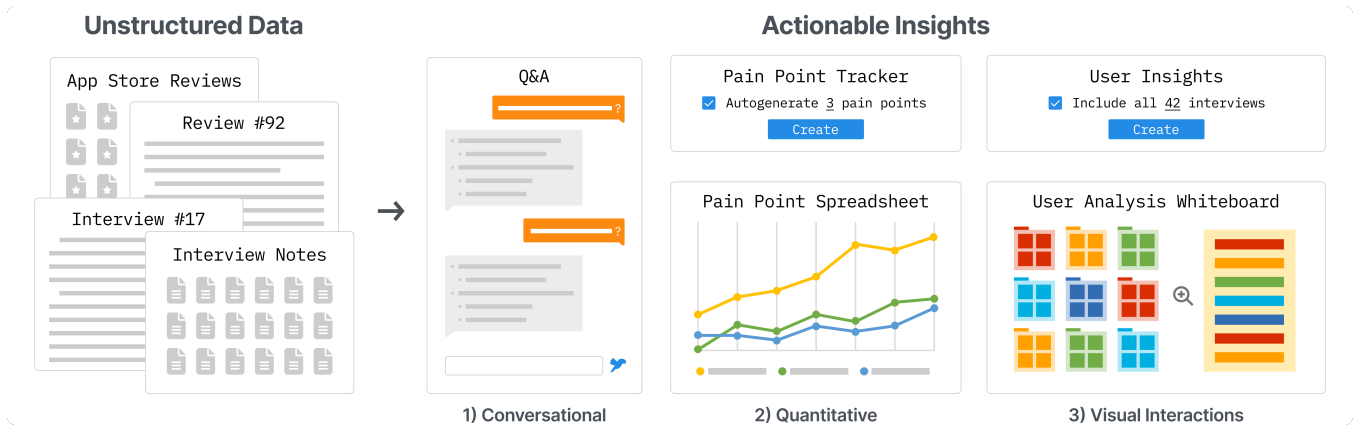


Figure 1: Yodeai hosts three AI-powered workflow templates (widgets) that transform unstructured data into actionable insights. 1) Q&A: ask questions and retrieve relevant information with citations (conversational). 2) Pain Point Tracker: cluster and quantify key issues over time (quantitative). 3) User Insights: view multi-granular summaries on a whiteboard (visual). With these widgets, users can explore data, identify key points, and visualize metrics to support exploration and decision-making.

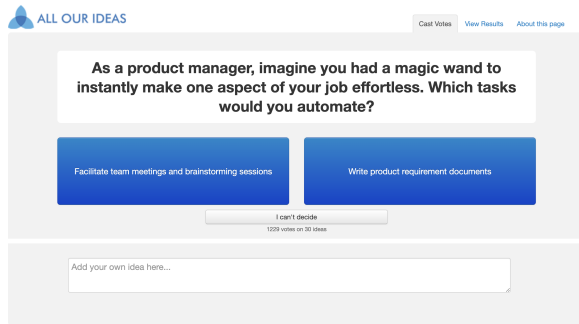


Figure 2: A voting interface for PMs to either submit their own ideas in response to the prompt or vote on existing ideas for PM tasks that AI could potentially assist with.

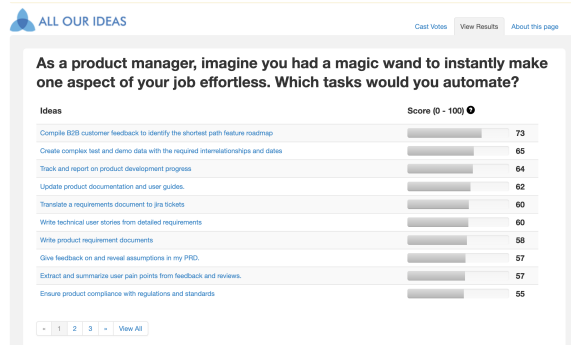


Figure 3: The top 10 results from the public poll, with each idea's score representing its estimated likelihood of winning against a randomly selected idea (100 indicating a guaranteed win, 0 indicating a guaranteed loss).

lack of a unified platform for information reuse and revisit led to our development of reusable, shareable, and customizable widgets.

We focused on interview and review analysis tasks as they mirror common PM activities [70] and parallel similar tasks in other knowledge work domains like consulting and academia, making our findings more broadly applicable.

Thus, we decided to focus on three specific widgets (i.e., Q&A, User Insights, and Pain Point) inspired by the poll and the formative study.

4.2 Organization & Structure, Context & Integrations: Overall System

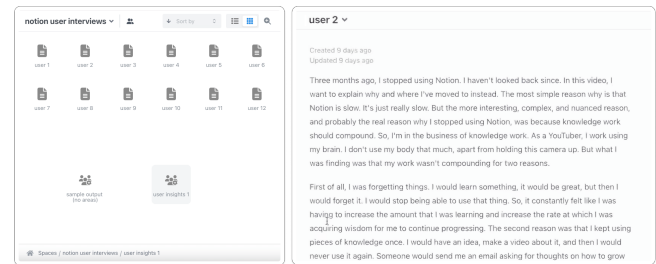


Figure 4: Yodeai's space and page structure. Left: A space containing multiple pages including raw data (user interviews) and widget outputs (user insights). Right: Individual page view showing an interview transcript. This organization allows users to maintain connections between source data and derived analyses.

Based on the formative study's emphasis on having a centralized platform that carries context, Yodeai offers a structured system of pages and spaces (Figure 4). A single space can contain multiple pages, including various data types such as text files, PDFs, spreadsheets, and whiteboards. Pages can be grouped flexibly across spaces and each of the pages have a generated AI summary of the content, giving users quick overviews with options for detailed

exploration. Finally, the system facilitates collaboration through shared spaces and widget outputs.

4.3 Prioritization & Search: The Q&A Widget

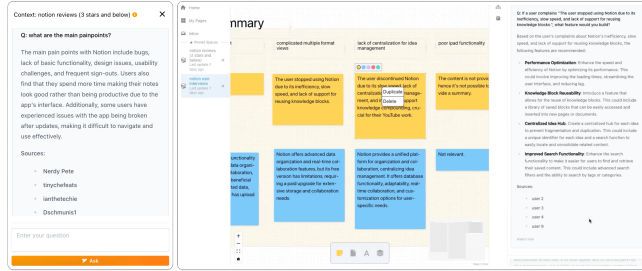


Figure 5: The Q&A widget provides conversational access to data across different contexts. Left: Direct querying of notion reviews about pain points, with sources linked. Right: Analysis of User Insights widget output, where users can ask follow-up questions about patterns and themes identified in the sticky notes.

The Q&A widget (Figure 5) operates on the data in the current active space, allowing team members to “talk to the data” and make sense of new information repositories, as needed by the formative study participants. This promotes efficient collaboration and reduces the time spent on manual data organization and communication that often comes with sharing data [13, 77]. This feature promotes increased engagement and control [53] and enables intuitive navigation. For example, if a user is working in a space containing multiple user interviews, they can ask the Q&A widget to identify users who mentioned a specific topic. The widget will then retrieve the relevant information along with sources for easy cross-referencing, as requested by the formative study participants.

4.4 Organization & Structure: The User Insights Widget

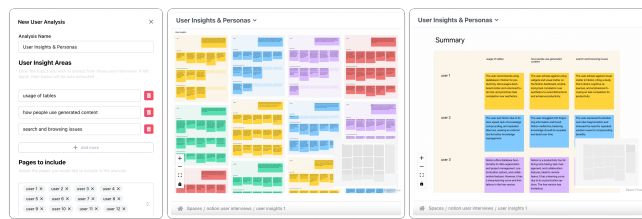


Figure 6: The User Insights Widget allows users to input insight areas and select pages for analysis. The widget generates a whiteboard view with sticky notes presenting user-specific insights and a summary, grouped into relevant areas of analysis.

The User Insights widget (Figure 6) transforms unstructured user data into an interactive whiteboard visualization—a design choice that emerged from our formative study findings about PMs’

need to synthesize overall patterns. The widget automatically clusters user data in a space (e.g., interview transcripts) and presents dual-granularity summaries [75]: lower-level sticky notes showing each user’s specific perspectives, and a higher-level synthesis of patterns across all documents. Users can specify multiple areas of analysis to focus on or leave it blank for full automation, allowing them to understand themes and sentiments without reading lengthy text, while still maintaining some control [37]. This flexibility emerged from our observations of how some PMs want a system to automatically identify themes for them.

4.5 Organization & Structure, Context & Integrations: The Pain Point Tracker

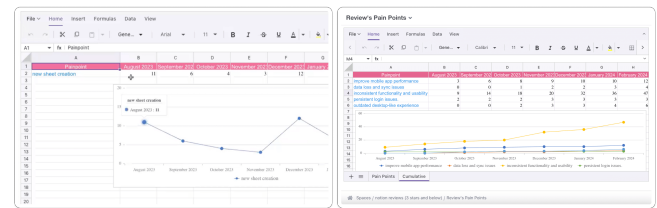


Figure 7: The Pain Point Tracker’s output varies based on user input. Left: Auto-generated analysis identifying and tracking general pain points over time. Right: A more focused analysis based on user-specified areas of interest, showing how providing additional context to the LLM yields different insights. Both views include spreadsheet formats and cumulative graphs to support temporal analysis.

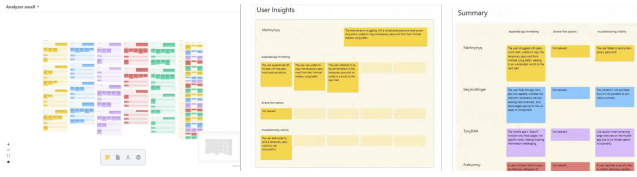
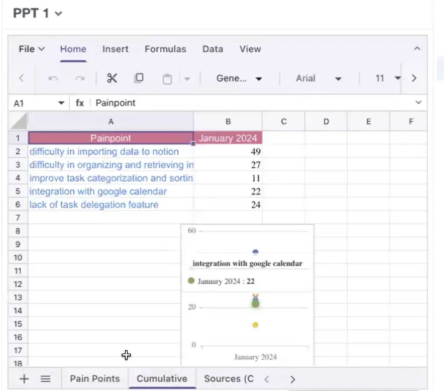
The Pain Point Tracker (Figure 7) enables iterative quantitative analysis of user feedback through varying levels of user guidance, as requested by participants in the formative study who worked with quantitative data. When run in auto-generation mode, it independently clusters data to identify main pain points, while in guided mode, it incorporates user-specified categories, reducing the need for prompt engineering [99]. This flexibility allows PMs to explore data from different angles: starting with automated clustering to discover unexpected patterns, then iteratively refining the analysis by providing more specific guidance based on their domain knowledge or hypotheses. The widget consistently provides both spreadsheet views of occurrence counts and cumulative graphs over time, supporting prioritization decisions based on observed patterns and perceived impact [33]. Finally, since the formative study revealed the need for integrations with outside platforms, the Pain Point Tracker has an option to first pull data from Apple store reviews, considering the rating and date metadata, to populate the space, before running its analysis.

4.6 Widget Flexibility and Adaptability

The flexibility to create multiple instances of each widget, with varying levels of user control from fully automated to user-directed, encourages experimentation and iteration [37, 83, 84]. For example, PMs can create multiple User Insights widgets to analyze the same interviews through different thematic lenses, or several Pain Point

Table 2: Overview of Widget Modalities, Data Types, and Use Cases

Widget	Modality	Example Data Types	Example Use Cases
Q&A	Conversational Exploration	Text files, PDFs, interview transcripts	General Q&A, prioritization, feature brainstorming, general theme assessment, competitive analysis
User Insights	Spatial Organization	Text files, PDFs, interview transcripts	Auto or pointed clustering of themes and feedback, various granularities (e.g., high-level topic or low-level topic summarization)
Pain Points	Quantitative Analysis	Text files, PDFs, transcripts (with dates)	Auto or pointed thematic plotting over time, keyword plotting, product tracking over time

**Figure 8: The User Insights Widget used on Apple Store reviews.****Figure 9: The Pain Point Tracker used User Interviews with no date.**

Trackers to compare issue trends across different user reviews, supporting both divergent thinking (exploring multiple perspectives) and convergent thinking (synthesizing findings).

Furthermore, the widgets have minimal requirements regarding the types of data they can process (Table 2). The Q&A widget breaks all the pages in a space into chunks and uses both the pages' content and previous history as context when answering questions. The User Analysis widget assumes that each page is a user interview or an entity that the PM wants to separately analyze. Thus, the User Analysis widget can even be used on Apple store reviews, although the output would not be nearly as detailed as for interviews (Figure 8). Similarly, the Pain Point Tracker assumes the data to have timestamps to be able to graph across time; if used on user interviews

without timestamps, the pain points it identifies will still be useful, though the graph would not (Figure 9).

4.7 Implementation Details

Yodeai is a web application built using Next.js for the frontend, Supabase for data storage and vector search, and Python for the backend. The widgets are powered by GPT-4, a large language model developed by OpenAI. For the Q&A widget, we use retrieval augmented generation (RAG) and obtain the top 5 chunks related to the user question with vector search, feeding it as context to GPT-4. For the User Insights widget auto-generated topics, we use vector averaging to get the top 5 chunks that are related to a single page, and generate three main topics, iteratively updating them with each page. In terms of generating the content for each user and topic, we also use RAG to retrieve relevant portions of the interviews, summarizing them with GPT-4. Finally, the Pain Point Tracker uses k-means clustering to cluster evenly segmented chunks of all the reviews, and then retrieves the 5 most representative chunks of each cluster to serve as context for GPT-4 to determine an overarching pain point. We use additional hyperparameters such as cosine similarity thresholds or relevance scores generated by GPT-4 to exclude chunks that are too far from the center, and then proceed to total the count of reviews for each pain point. For embedding words into vectors, we use the bge-large-en model hosted on Hugging Face. Finally, we use React Flow for the whiteboard integration, SyncFusion for the spreadsheet, and Mantine for UI components. All code is open source: <https://github.com/yodeai/yodeai>.

5 User Study

With Yodeai, our aim was to answer the following question: *What opportunities and limitations does generative AI pose for data navigation and decision-making in product management?*

5.1 Participants

We recruited 16 PMs in the United States (age: $M = 31.25$, $SD = 9.46$; gender: 5F, 11M) through a survey shared on LinkedIn. 15 of them identified as PMs and 1 as a developer advocate. All of them had several years of experience ($M = 6.31$, $SD = 6.97$), currently work in companies ranging from startups to large corporations, and explained that their day-to-day work involves a significant amount of PM tasks. Every participant mentioned that synthesizing user

insights and tracking and evaluating pain points are within their primary responsibilities. 14 participants were familiar with AI tools, stating that they use an LLM (ChatGPT, Gemini, or Claude) in their work, ranging from daily to every two weeks. Every participant was compensated with a \$50 Amazon gift card.

5.2 Procedure

Each semi-structured interview lasted approximately 80 minutes and was divided into five main sections. The interviews were conducted remotely over Zoom, and participants were asked to share their screens during the system onboarding and main task sections. Each interview was conducted by two interviewers (both authors), who took turns conducting the pre-/post-interview questions and the task itself while the other person took notes. All procedures were approved by the IRB of our institution and informed consent was obtained from each participant.

5.2.1 Pre-Task Interview (15 minutes). The pre-task interview aimed to gather information about the participants' professional background and their current practices related to product management, user/pain point analysis, and AI usage.

5.2.2 System Onboarding (20 minutes). During the onboarding process, we shared with participants Yodeai spaces containing publicly available data about Notion, an app for note taking and information organization. Specifically, we had two spaces: an Interview space, which contained 12 long and unstructured YouTube transcripts, and a Review space, which contained 132 reviews from the Apple App Store that were 3 stars and below. We started by explaining to participants how to use the Q&A widget, prompting them to ask a question related to the data in the spaces, such as "give me a pain point users have when using Notion," to showcase how the system pulls information from the spaces containing interview transcripts and scrapes relevant data to provide a response. Next, we showed them how to use the Pain Point Tracker by auto-generating a couple of pain points from all the reviews. Finally, we introduced the user insights widget by having participants generate a whiteboard view of a subset of the interviews.

5.2.3 Task (25 minutes). The main task was a role-play scenario in which participants assumed the role of a PM at Notion. Using the two spaces containing all relevant information, participants were instructed to use Yodeai to propose 6 new feature or bug fixes (divergent phase), and then prioritize the top 3 proposals (convergent phase). During the task, participants were encouraged to articulate their thought process and concerns as they interacted with Yodeai.

5.2.4 Post-Task Interview (20 minutes). Following the main task, we conducted a semi-structured post-task interview to gather participants' feedback on their experience using Yodeai, as well as general comparisons with existing tools. We asked open-ended questions to encourage participants to share their thoughts and reflections on the tool and the use of AI in their work.

5.2.5 Post-Interview Survey. After the interview, we asked participants to complete a survey to collect quantitative data on various

aspects of the study, including the task design, participants' experience with Yodeai, and their data exploration and decision-making processes.

5.2.6 Data Analysis. We analyzed the interviews using thematic analysis [9], with at least two authors coding each interview to ensure reliability. We employed line-by-line coding for both pre/post-task and task sections, creating inductive codes (e.g., "Trust and Explainability", "Summarization"). This bottom-up approach [6] was more suitable for our semi-structured interviews than a top-down method, as the high level research questions were too broad to effectively group the findings, and reading through the transcripts and notes from the interviews line by line allowed for detailed grouping and labeling that could eventually be clustered together, to ensure that as much insight was gleaned as possible. We then used affinity diagramming [23] in FigJam to cluster quotes and actions, identifying themes and sub-themes. Finally, we selected relevant sub-themes for data navigation and decision-making, ensuring a comprehensive analysis aligned with best practices in HCI qualitative studies. We use prior work to help interpret and generalize our findings, which are placed on the backdrop of how PMs utilize AI tools.

6 Results

6.1 What opportunities and limitations does generative AI provide for data navigation?

6.1.1 Opportunity #1: Providing a starting point at differing levels of abstraction.

Yodeai Specific Findings: During the exploration and research stage of the user study task, participants (P3-P6, P16) appreciated how Yodeai helped them navigate large amounts of data and condense raw information into different levels of abstraction. Participants (P1, P3-P4, P6, P10, P13) highlighted the efficiency gained in navigating and making sense of new information using the proxy system Yodeai; they saw the widgets as starting points for their research process. P3 mentioned that Yodeai offers a "first pass of what the data looks like," while P15 noted that the outputs help them see what people are mostly mentioning to get a "general sense of what people like about the product." P5 also noted that Yodeai is particularly helpful when "dealing with a high volume of external data that you are not directly familiar with."

Participants explained that creating whiteboards with sticky notes or spreadsheets of pain points are familiar tasks in their day-to-day jobs, and without the generative AI widgets in Yodeai, the entire process would have been manual. Each participant had a different preference for the outputs of the widgets, as shown by Figure 10. Some participants, such as P13, found the granularity of the Pain Point Tracker too general and used the Q&A widget to narrow down the pain points, while others, such as P16, appreciated the high-level categorizations. Participants like P6 liked the in-depth breakdown for each participant in the User Insights widget whiteboard, while others, such as P8, gravitated towards the high-level summary, finding the sticky notes overwhelming. The 16 participants stated that the Q&A widget was a crucial part of their workflow, the sources allowing them to navigate and find specific pages without reading each one. For example, P4 was able to find

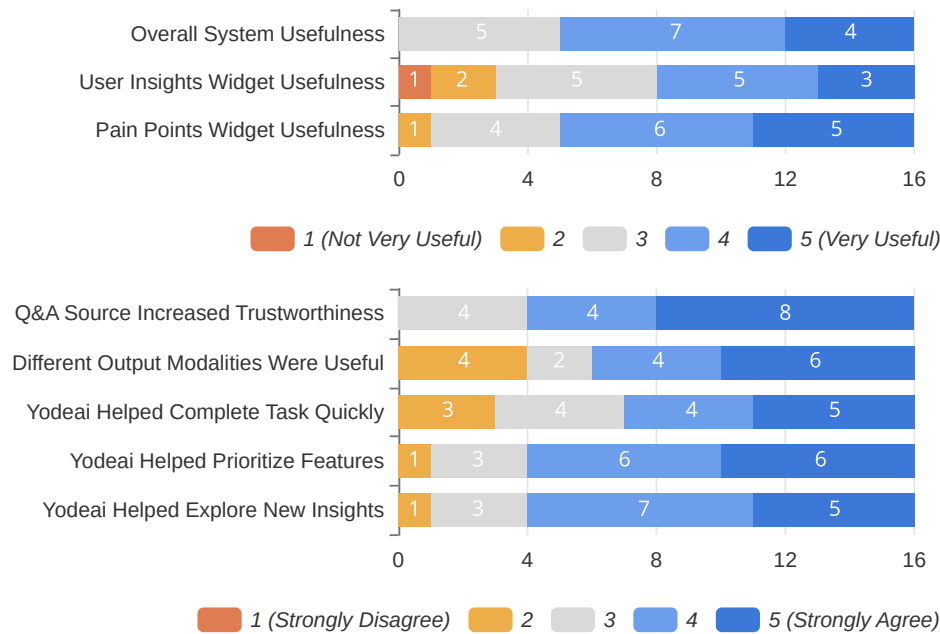


Figure 10: Evaluation of Yodeai: most participants found Yodeai’s widgets useful for exploring and prioritizing features.

two user interviews that specifically mentioned “overwhelm of complexity” by asking for references.

General Findings: Many PMs echoed the notion that AI helps them combat “blank page syndrome” and gives them starting point content and ideas (P1, P7, P13). For example, P13 stated that if they wanted to learn more about enterprise collaboration, they would use ChatGPT to ask questions such as “What are the different ways of collaborating via commenting or message threads?” Additionally, PMs deal with high variability in the type, quantity, and granularity of data. P16 noted the difference between building version 1 and version 10 of a product; with version 10, there is a lot of user data and logs to sift through, while for version 1, the data would primarily come from high level competitive analysis, as used by P5, P8 and P12. P15 also clarified the different types of data they have to work with: quantitative usage analytics vs qualitative user interviews, while P8 disclosed that they use surveys, user interviews, and comparative analysis. In addition, when presenting findings to various stakeholders, the presentation format differs. P1 noted that they would use a whiteboard-like output to present to the CEO, while P9 writes detailed PRDs (product requirements document) with a list of interviews and competitors for their engineering team. Finally, the level of seniority also influences the granularity of work that a PM does; P13, a senior PM, stated “I’m more involved in the high level strategy and the high level prioritization of features, whereas a more junior PM [...] would still do a cycle of going even deeper into the customer needs.”

6.1.2 Opportunity #2: Organized and transparent collaboration through an audit trail.

Yodeai Specific Findings: Most participants (12/16) appreciated Yodeai’s shareable spaces and pages, agreeing that its structure facilitated transparent collaboration. Nearly all (15/16) found sharing widget outputs and Q&A results beneficial for teamwork. P8, while acknowledging Yodeai’s helpfulness for shared information, felt the outputs were “too raw” for direct team presentation. P16 specified they would share within the product team but not with marketing or sales, highlighting the need for polished results for certain team members. Yodeai’s raw data collection allows for exploration by team members involved in data navigation. P12 noted its particular benefit for larger product teams, enabling independent exploration of interviews and data. P7 and P9 emphasized the importance of visibility and transparency regarding feature status and data for stakeholders. P9 and P10 stressed the need for engineers to develop empathy, with P9 stating, “if I share [Yodeai] with engineers, it’s gonna be useful in that they would get a firsthand view of what users are actually talking about [...] Engineers never talk with users. And it’s the PM, and the salespeople who always talk with the users, so it’s our duty to share those insights with the engineers, so that they are also empowered by the problem they’re solving.”

General Findings: Participants noted collaboration within and outside their immediate teams. P10 mentioned working with marketing, designers, engineers, customers, and partners, while P2 emphasized collaborating with VP, CTOs, and developers. Each cross-functional partner requires different outputs and data synthesis. P7 explained, “depending on the use case, customer success managers want to know the specifics of when a feature will come out so they can tell the customer. And for engineers and designers, it would be more about the creativity of creating different ideas as

a collaborative effort.” In addition, without an audit trail, P7 mentioned that sometimes people give a problem without context, so they have to spend two weeks to “chase context”.

6.1.3 **Limitation #1: Overreliance and reduction of insight depth.**

Yodeai Specific Findings: Some participants expressed concerns about the quality of automated outputs. P2 felt a lack of connection with auto-generated sticky notes, stating, “even though the data is tagged nicely on sticky notes, because these are auto-generated, I don’t feel that connection with those sticky notes.” In terms of the User Insights widget, P6 noted that the sticky notes were overly verbose when summarizing key points, preferring shorter, more focused content, while P7 felt “blocked” after seeing the summarizations and big themes, unsure of how to dig deeper.

General Findings: Participants emphasized the importance of engaging directly with source data for deeper understanding. P4 explained that PMs “cannot fully understand the issue without actually reading into the original text.” While AI can improve the insights gathering process through summarizations and overviews, human intuition and contextual understanding generate more nuanced insights. P5 highlighted the value of direct customer feedback: “Reading customers’ direct words and getting the gist of it from AI are two different things. You understand more about the feelings and emotions behind the feedback when you look at the actual comments and words used by the customers.”

6.1.4 **Limitation #2: Outdated data and priorities.**

Yodeai Specific Findings: While participants appreciated Yodeai’s centralization of information (P2, P10-P11, P16), some (P3, P8, P11) suggested improvements in navigation, such as the ability to overlay and combine information. P11 emphasized the importance of tracking pain points over time to identify outdated feedback. Many participants (P2, P11-P13, P15, P16) expressed interest in integrations from other sources like Reddit or LinkedIn, especially if these could auto-update with the latest information.

General Findings: Participants highlighted the challenge of maintaining up-to-date information when using LLMs. P5 noted that information can quickly become outdated, particularly with news updates. In terms of pain point identification, P11 stressed the need for “constant revision” in prioritizing pain points, emphasizing the importance of shifting focus as solutions are implemented. While LLMs help process large amounts of data, they struggle to identify stale or irrelevant insights over time; P2 stated that using AI would require constant extra context, like “whether this pain point is already being worked on”. Users must continuously update the data fed into LLMs by adding new material and removing outdated information. P15 confirmed this, stating that his company either interfaces directly with customers or uses forums like Reddit to gather current feedback, while P7 uses customer success conversations or quarterly surveys.

6.1.5 **Limitation #3: Privacy and confidentiality.**

Yodeai Specific Findings: Participants expressed hesitancy about using Yodeai professionally without company approval. P12 emphasized the need to ensure “confidentiality of my company’s source

content.” Participants were more likely to trust tools with integrations from well-established companies like Slack or OpenAI (P2, P9), with P13 expressing delight that Yodeai had a Figma-like interface with the User Insights widget.

General Findings: Companies are cautious about adopting external AI tools due to confidentiality risks. Participants (P9, P11-P13) mentioned using only “company approved” tools. The rapidly evolving AI landscape complicates this, as approval processes may become outdated. As a result, many PMs use general-purpose tools like ChatGPT with redacted information (P3, P13) or rely on their company’s enterprise chatbots (P6, P11-P12, P15).

6.2 **What opportunities and limitations does generative AI provide for decision-making?**

6.2.1 **Opportunity #1: Customization to diverse workflows.**

Yodeai Specific Findings: Each of the 16 participants followed a unique workflow, demonstrating Yodeai’s adaptability in assisting diverse decision-making processes. Figure 11 illustrates the variety in participants’ approaches, including different starting points, ideation methods, cross-validation techniques, and prioritization strategies. Participants either combined content from both spaces into one feature list (P1, P2, P8, P12, P14, P16) or focused on one space (P6-P7). P5 highlighted Yodeai’s flexibility: “[Yodeai] is also open in a way because I can influence the results by providing specific keywords. So it’s a mix of both. If I want to have control, I can, and Yodeai provides options for that. But if I don’t want to, I can also just let the platform generate results based on its own analysis.” All participants agreed on the usefulness of generating multiple widget outputs and editing focus areas. Some participants used widgets in unexpected ways, such as applying the Pain Point Tracker to user interviews or the User Insights widget to reviews (P2, P6, P8, P10, P12). In addition, some participants expected the different widgets to know about each others’ outputs; for example, P4 thought that the Q&A widget would have context about the graph generated to ask questions about it. Others wanted the Q&A widget to know about the structure of Yodeai and meta information such as how many pages a space has (P9, P14).

General Findings: PMs highlighted the diversity of real-life workflows. For instance, P6 prefers manual note-taking during interviews, while P9 records interviews for later analysis. PMs often juggle multiple roles, as noted by P3 (UX researcher, data analyst) and P8 (design, engineering, legal issues). The level of data analysis also varies significantly between B2B and B2C companies. P16 explained: “[B2B] customers will tell you what they need and you don’t rely a lot on user research, but for B2C you have customers that ask for features and you need to see the data and see how many people are using it and you need to use other metrics to measure and decide.” P12 and P15 confirmed that B2B PMs focus primarily on accounts rather than customer reviews and interviews.

6.2.2 **Opportunity #2: Intermixing of background knowledge.**

Yodeai Specific Findings: Participants combined personal heuristics with LLM reliance during exploration. P9 used product knowledge to guide their exploration, inputting specific features into the

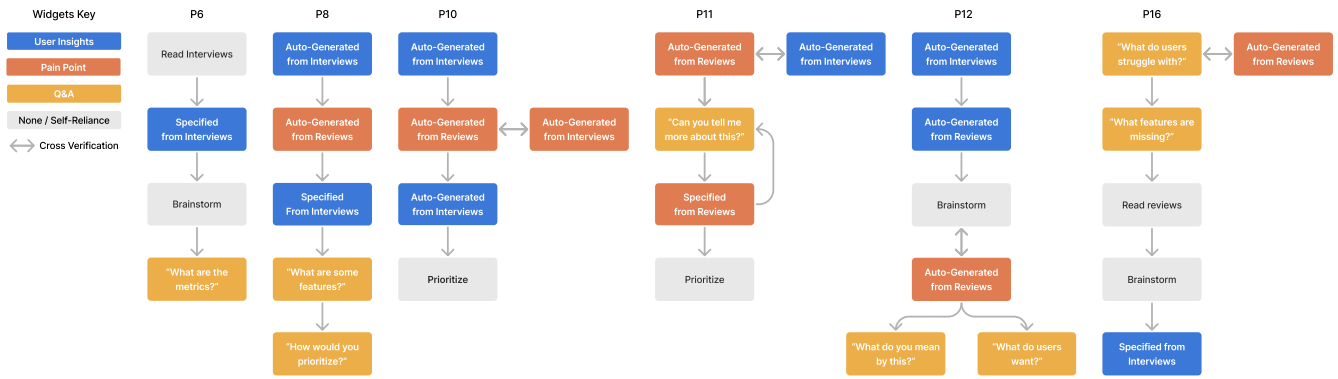


Figure 11: Visualization of 6 participants' decision-making workflows using Yodeai. The use of each widget at each step is indicated with a color: blue for User Insights widget, red for Pain Point Tracker, yellow for Q&A widget with the question asked, and gray for no widget use. Double arrows represent cross-verification between steps. “Interviews” and “Reviews” refer to the respective Notion spaces in Yodeai, while “Auto-Generated” and “Specified” describe widget creation without or with pre-specified fields. This figure shows that each participant used the widgets in their own unique way, emphasizing the importance of tool flexibility.

Pain Point Tracker. P7 leveraged the Q&A widget and User Insight widgets for brainstorming. P6 mixed approaches, considering their “own background and experience, and previous customer contacts or previous contacts with other products” while reviewing Yodeai’s summaries. P8 noted, “even though I have some personal opinions on the user pain points based on a very brief review, once I throw that into the user insight widget, it could pretty effectively relate to whatever pain points users describe. That helps me make informed decisions regarding what features I’m trying to build.” Some participants, such as P1, focused on self-specified topics and used personal heuristics to prioritize features like “encryption” and “gamification of on-boarding.” Others such as P11, P12, and P15 categorized bugs and features separately, emphasizing the importance of fixing bugs before implementing new features. This demonstrates how Yodeai’s widgets help PMs validate and expand upon initial assumptions, leading to more informed decision-making.

General Findings: PMs develop instincts and heuristics over time. P7 stated that senior PMs have experience to know “I have tried this before in this context, and it didn’t work out” to quickly validate ideas. P15 noted that they are tightly intertwined with their product and have instincts on prioritization.

6.2.3 Opportunity #3: Iterative prioritization.

Yodeai Specific Findings: Participants employed a combination of Q&A, personal experiences, and intuitions for prioritization. While some (P8, P14) utilized Q&A during the convergent phase, others (P1, P5, P11, P15) relied solely on their experiences. The Pain Point Tracker provided a high-level overview, but participants often prioritized based on personal judgment. For instance, P12 categorized pain points into features or bugs, giving higher priority to bugs. Participants increased information granularity through various methods: they asked Q&A for details (P1-P6, P11, P12, P14), referenced surrounding reviews and interviews (P2, P6, P16), and broke down points using new widget generations (P3-P4, P8, P11).

However, some participants wanted even more direction from Yodeai; P13 mentioned that they would want a list of next steps out of the insights, while P4 wanted to have effort estimation next to each of the next steps.

General Findings: P5 stated that in their previous PM experience, their manual process was to identify 5 most important pain points from customers, identify how the product would solve these points and what customers would gain, and then narrow it down to 2 pain points and build solutions. Thus, participants stressed the importance of quantitative data for decision-making (P6, P11, P14). P12 noted, “as a PM, numerical data is hard to come by, any numbers is a big plus.” Time-based information was valued for identifying stale feedback and tracking changing sentiments (P3-P5, P11-P12, P16). Some preferred displaying sentiment for each pain point (P9) or using visualizations like word clouds or pie charts (P9, P3).

6.2.4 Opportunity #4: Facilitating data validation and collaborative trust.

Yodeai Specific Findings: Participants emphasized the importance of cross-comparing and cross-verifying data for validating outputs. They appreciated the inclusion of sources in Q&A outputs, which increased trust in LLM output (7 Strongly Agree, 5 Agree, 4 Neutral). Cross-validation methods varied among participants: some compared information between spaces (P3-P5, P10-P11, P16), others overlaid Pain Point Tracker graphs (P11), while some manually read source data (P2-P3, P6, P16) or used the Q&A widget (P1-P3). Specifically, participants like P2 and P16 employed a mix of manual review reading, the Q&A widget, and the Pain Point Tracker, leading to features such as “create templates for task management based on domain” (P2) and “improve search functionality with highlighting” (P16). Some participants, like P4, created both auto-generated and specified User Insight and Pain Point widgets to glean further details on what features to build. Participants valued Yodeai’s audit trail feature for team collaboration and decision-making transparency. P9 stated they would share spaces with managers and mentors,

noting that “managers are very much interested in how you are identifying the problem because the problem is important.”

General Findings: For personal work with generative AI, PMs adopted various approaches to mitigate misinformation. Some avoided ChatGPT for research due to hallucination fears (P9), while others used AI only for summarization or revisions (P3, P9). Cross-verification methods included checking with Google (P1, P7), requesting sources from AI (P2, P12), and comparing against company documentation (P5, P8). P11 noted that “a lot of PMs are not willing to risk their job on the output of the model,” emphasizing the need for constant data efficacy checks. Many of the PMs echoed that their process for pain point and user interview analysis is manual due to the need for validation. They stated that their first step to user interview analysis was to read over their notes or the transcripts, as many of them wanted to be familiar with the raw data before starting out (P1, P5-P6, P8, P12, P14, P16). In terms of the pain point identification process, P9 and P16 stated that they would read through the actual reviews word for word. When they did use AI, many of the PMs used it as a direct Q&A format to ask questions about the interviews (P10, P12), emphasizing that they would be able to “check” the AI since they were already familiar with the raw data (P4). However, they noted that when they did use chatbots such as ChatGPT, they would have to keep reprompting and asking questions to structure the answers better (P1, P3, P7).

Due to the collaborative nature of PM work, PMs echoed that it is also difficult to verify the work of others; P11 shared that one of their tasks as a PM was to blindly trust an Excel sheet of pain points from users without knowing the data’s origin or how it was analyzed.

6.2.5 **Opportunity #5: Trade-off between creativity and objectivity.**

Yodeai Specific Findings: During the identification of new bug fixes and features, PMs utilized both objective information from the widgets and data, and was creative either by themselves or with Yodeai. For example, P6 ideated features like “30 days trial version of paid premium subscription” and “nudges on old tasks/notes” based on the auto-generated User Insights widget output and their past experience, while P5 prioritized high-level bugs like “login and access issues” and “data loss” using the auto-generated Pain Point Tracker and their personal intuitions. Others used AI for creativity, such as P8, who heavily used the Q&A widget to brainstorm and prioritize features, resulting in a final list that included features such as “support for Apple pencil for handwritten notes and sketches.”

General Findings: Participants emphasized the value of leveraging LLMs’ creative aspects in decision-making. PMs such as P7 would prompt engineer by giving ChatGPT a pain point, provide context as a PM and their specific company, and ask for suggestions. P1 also highlighted ChatGPT’s potential in the ideation phase, and the fact that regenerating is useful to avoid fixation: “The answers [ChatGPT] gives can spark new ideas in the ideation part.” P10 also discussed how LLMs can uncover insights by combining data from disparate sources and handling edge cases and knowledge gaps. In such scenarios, prompting may not be very useful, as users need to recognize that “new insights could be something that we were not thinking about” and may not be present in the existing dataset.

However, PMs such as P9 and P11 echoed the notion that they used AI to summarize or edit information rather than spout new insights. For instance, P5 leverages LLMs for background research and writing assistance due to English not being their first language, whereas P14 uses ChatGPT to refine information they are already familiar with.

6.2.6 **Limitation #1: Contextual factors outside of the model’s reach.**

Yodeai Specific Findings: Participants emphasized the importance of considering factors beyond Yodeai’s outputs when making critical decisions. These factors include business context, collaborator opinions, monetization benefits, and product strategy (P4-P6, P11, P12, P15, P16). P6 noted, “at the end of the day you still need a person who actually knows the business context to look at and review the feedback given by the tools.” Additionally, participants (P11, P13, P16) stressed the need for an export feature in Yodeai to work on widget outputs externally, such as exporting to Figma for design teams or overlaying Pain Point Tracker outputs with their product roadmap to see what happened when they shipped.

General Findings: Participants highlighted the high-stakes nature of their work environment, and noted that they did not really use AI for prioritization, due to fear of hallucinations and other factors. P11 mentioned frequent data verification due to potential job loss from errors, while P2 noted the value of manual effort in ensuring correct decision-making. P11 emphasized the complexity of professional decision-making, including factors like budget, release schedules, marketing, and sales. P11 demonstrated prioritizing a pain point not listed as “high priority” by the Pain Point Tracker, emphasizing user retention: “if an app crashes, a user is gone [...] that’s why app crashing would be my first feature.” P16 stressed the need to incorporate priorities from various stakeholders (sales, marketing, support teams, CEO) and continuously refine feature lists based on recurring issues, available resources, and company goals. P2, P6, and P15 stated that the job requires so much contextual knowledge and that it is not “good enough” to make decisions at work. P4 gave a bit of insight into the collaborative nature of the prioritization process, stating “In my normal workflow [I] port [the list] to my engineers and see like the effort that’s gonna require, and see if the effort corresponds to the impact that’s gonna create again.”

6.2.7 **Limitation #2: Isolation and Confirmation Bias.**

Yodeai Specific Findings: P6 highlighted a potential limitation of Yodeai and similar AI tools, stating, “it’s good to be able to say you don’t know the answer to something, or to literally be proven wrong, because that’s how you learn. So if Yodeai is constantly telling me, ‘Oh yeah, this is the right thing,’ it could really lead me down the wrong path.”

General Findings: Participants reported using at least 24 different tools with AI integrations, highlighting the proliferation of AI in knowledge work contexts. While these tools offer potential benefits, they also raise ethical concerns. P7 expressed concerns about AI undermining collaboration: “I don’t know if AI actually enables collaboration because it’s just like firing people and doing more with less. A single designer can come up with the ideas and not have anyone else involved, and collaboration kind of dies there.” They

also noted potential risks to creativity: “I feel like AI is putting the whole creativity theme at risk [...] now you only need one person to create 30 ads instead of bringing people together to generate these ideas.” P1 pointed out the tendency of AI, particularly ChatGPT, to be overly agreeable: “I feel like often, to conform to my biases, it’s just trying to be agreeable, I think, in general.”

7 Discussion and Design Implications

Our empirical study reveals three critical design implications for AI tools across knowledge-intensive and decision-making domains, initially examined through the lens of product management but broadly applicable to contexts requiring complex information synthesis and strategic decision-making. These insights emerge from studying how knowledge workers navigate data-rich environments where decisions carry significant organizational impact.

First, AI tools must support flexible workflows that adapt to different work styles and analytical approaches. Second, they must enable robust verification and collaborative accountability given the high-stakes nature of professional decision-making. Third, they must balance the integration of diverse data sources with security concerns and evolving organizational priorities. We detail these implications in Table 3, connecting each to PM specific tasks while highlighting their broader relevance to knowledge work.

7.1 Design for Workflow Flexibility and User Autonomy

PM specific task: User Interview/Review Analysis

Rather than imposing new workflows, AI tools must adapt to how an individual PM goes through their tasks. Our study revealed that PMs have personal established and effective practices for analyzing user feedback. The role of AI should be to enhance these practices while letting PMs maintain control over how and when to use AI assistance.

7.1.1 Control over AI-Influence. AI tools must let PMs dictate their preferred level of AI involvement rather than enforcing a fixed interaction model. Our study revealed stark differences in how PMs chose to work with AI: some relied heavily on automated features for efficiency (P8 completed tasks in 23 minutes using the Q&A widget), while others deliberately limited AI use to preserve learning opportunities, preferring to read reviews and interviews manually to extract emotional context and engage in deeper reflection [91]. Even within individual workflows, PMs varied their AI usage—some used the User Insights widget solely for initial summarization before switching to manual analysis, while others leveraged the Pain Point widget throughout their process from identification to prioritization. This suggests AI tools should provide clear affordances for adjusting AI involvement at any point in the workflow, letting PMs fluidly shift between automated and manual approaches based on their current needs and preferences.

7.1.2 Encourage Versatile Use. AI tools should be designed with the expectation of creative re-purposing rather than assuming fixed use cases. Prior work shows users tend to appropriate workflows for unintended tasks [49], and our findings confirm this. For instance, P15 suggested using Yodeai to analyze competitor products when entering new markets, which was far beyond the tool’s original

purpose, but showcases the versatility of the widget. In addition, different product lifecycle stages require different forms of data analysis, from early-stage competitive analysis to detailed feature refinement in mature products [2]. Varying experience levels lead to diverse data representation preferences as well. The tool’s combination of conversational, visual, and graphical formats proved valuable not because each format had a predetermined use, but because it helped PMs discover multiple entry points into their data; PMs often face large volumes of data and can easily become fixated on a single AI output, but diverse representations help reduce missed nuances, as shown by the PMs who tried to use different widgets on the same space. This adaptability was particularly crucial given the collaborative nature of PM work—the same analysis often needed to shift between detailed, unpolished data for product teams and high-level overviews for stakeholders. This also highlights the need for diverse data representations and options to export outputs; for example, Yodeai’s whiteboard results that emulate Figma enables collaboration with team members in other roles, like designers. Tools should therefore provide modular components (e.g. widgets) that can be repurposed rather than rigid, task-specific features. Additionally, AI tools should enable consideration of both input data and previous outputs, similar to how PMs wanted the Q&A widget to incorporate insights from other widgets, though some, like P16, were concerned about the risk of self-pollution if global context was enabled.

7.1.3 Objectivity and Creativity Modes. AI tools need to explicitly support and differentiate between objective analysis and creative exploration rather than conflating these distinct thinking modes. Our study revealed a clear tension in how PMs approached decision-making. Some preferred using AI exclusively for objective data consolidation while relying on personal judgment for creative insights, (such as P6 in Figure 11) whereas others actively sought AI-generated suggestions for feature ideation (such as P8 in Figure 11). This distinction reflects the dual nature of PM work: needing both rigorous analysis of user feedback and creative exploration of possible solutions. Tools should provide clear affordances for switching between these modes, letting PMs deliberately choose when to focus on objective data synthesis versus creative exploration. This separation helps PMs maintain clarity about when they’re analyzing existing feedback versus exploring new possibilities, while still supporting natural transitions between these modes as they refine features and align them with business goals and engineering constraints.

7.2 Ensuring Accountability and Validation in High-Stakes Decisions

PM specific task: Feature and bug fix brainstorming

Product decisions have far-reaching implications for users, engineering resources, and business objectives. Our study revealed that PMs need comprehensive accountability mechanisms that go beyond simple AI verification—they need tools that help build and demonstrate decision confidence across multiple stakeholders.

7.2.1 Many Ways of Cross Verification. AI tools must support multiple complementary verification strategies that PMs can combine based on their context and confidence needs. Unlike personal AI

Table 3: Summary of Design Implications, Interactions, and Suggestions for Enhancing AI Tool Usability in Product Management and Knowledge Work.

Design Implications	Proposed Interactions	Example Suggestions for Product Management
1. Adaptability Enable users to dictate their unique data analysis process	Control over AI-Influence Provide a starting point with adjustable controls to set the level of use and influence over the AI	• Task specific AI integrations (e.g. Jira ticket generator) that adhere to one step in a PM workflow, with starting templates • Allow users to adjust prompts to tailor the tool to their needs (e.g. a slider for the level of technical detail in output) • Expose AI input data for users to extract emotional insights and nuances
	Encourage Versatile Use Offer many ways to organize and interpret the same input data	• Develop a versatile tool for general goals, but adaptable to specific tasks (e.g. a tool for product performance analysis for both company and competitor tools) • Options to view data in multiple representations and granularities (e.g., violin plot for demographics, sticky notes for high-level themes) • Tailor outputs for different team members (e.g., high-level reports for stakeholders, technical details for engineers)
	Objectivity and Creativity Modes Different modes for objective, data-driven tasks vs. tasks that require more novelty	• Mode that focuses on summarization and information retrieval (objective) versus discussing multiple POVs and edge cases (creative) • Be transparent about its limitations to certain prompts (objective) vs being receptive and responsive to a wide range of prompts (creative)
2. Accountability Ensure transparency and verifiability of AI outputs	Many Ways of Cross Verification Allow users to verify a tool's outputs, ensuring it holds itself accountable	• Quantitative justification: The AI should provide numerical evidence • Attribution of ideas: The AI correctly identifies who said a statement • Implement a search functionality that enables precise identification of documents, specific lines, and timestamps for attributed statements
	Bias Mitigation Explicitly communicate confidence levels and potential biases rather than providing misleading certainty	• Detect and warn about leading questions (e.g. "What users hate X?" into "What do users think about X?") • Enable comparative analysis by showing multiple perspectives side-by-side (e.g. different interpretations of same user feedback)
	AI-Use Audit Trail All responses should be matched with the user prompts and data sources when sharing for reputability	• Showcase the path a user took to utilizing the AI output, from their revisions to their final conclusions (such as a timestamped log) • Allow other users to clone the context and existing environment surrounding the AI tool use, for their own exploration
3. Interoperability Facilitate safe integration of internal and external data	Security-Aware Integration Implement tiered data protection from local to cloud environments, with clear boundaries for proprietary information	• Clearly list the tool's features and data usage plans for company compliance teams to speed up the tool approval process • Only accept data from trusted platforms, have backups of past data, and sanitize and filter new data to prevent poisoning
	Business Context and Priority Integration Keep the model up to date with the latest priorities and company status	• Allow for ways to 'zoom in' on AI outputs in order to glean additional information that the initial output might not contain • Allow for regeneration of outputs as priorities and information change • Provide workflow status tracking and task management
	Multi-Pipeline Data Flow Enable the ingestion and prioritization of multiple data sources whilst ensuring updated and relevant data	• Build data ingestion pipelines that filter for content relevancy, date, and specific users in public and private user platforms • Provide an option to rank and prioritize certain data sources over others or to include / exclude specific data sources (e.g. to prioritize stakeholder interviews over regular customer interviews)

use, where errors like hallucinations and misrepresentations may have less impact, corporate AI use—especially in brainstorming product enhancements and fixes—directly affects company success. Our findings reveal the need for robust cross-validation methods, such as source citation functionality that enables users to trace AI-generated insights to original data [27]. This was evidenced by how participants built confidence by combining multiple approaches: comparing findings across different data spaces, generating multiple widget outputs to compare with each other, and frequently returning to the source data. Thus, AI tools should allow direct comparison of outputs at various levels, such as comparing regenerations of the same or different AI tools, comparing outputs with the same tool and parameters but different data, or comparing outputs from different tools with the same data. This not only allows PMs to be able to see common themes brought up by several different tool outputs, but also allows them to see the outlier outputs that they might have to investigate further.

7.2.2 Bias Mitigation. Tools must help PMs identify and counteract potential biases in both AI suggestions and their own decision-making processes. Our study highlighted challenges in AI accountability, including inherent biases in LLM prompts that can skew results [72] and the risk of confirmation bias in AI tool usage. We observed that participants without AI expertise struggled with effective prompting, often overgeneralizing or incorrectly equating AI interaction with human communication [99]. This was reflected in participant behavior—P6 emphasized the importance of being “proven wrong” for learning, while P1 noted AI’s tendency to be overly agreeable. Some participants (P9, P11) deliberately limited AI use to basic summarization to avoid potential bias in more complex analyses. This suggests tools need explicit features for bias identification and mitigation, helping PMs distinguish between objective data synthesis and potentially biased interpretations.

7.2.3 AI Use Audit Trail. Trust in today’s workplace extends beyond human colleagues to encompass AI tools. In their high stakes work environment, many PMs recognized the importance of delivering accurate and valuable work, both for the sake of themselves and their peers. They valued knowing exactly where the data came from, which can be increasingly difficult when handed materials and conclusions from other team members without additional context. P7’s observation that LLM use often becomes a solitary activity also highlighted the risk of creating information silos [91], particularly when AI tools become overly accommodating to individual viewpoints. Although research has examined transparency in AI algorithms for HR purposes [63], our findings emphasize the equal importance of transparency in AI interactions for PM tasks such as feature brainstorming. Thus, AI tools should incorporate lasting audit trails that allow for replicability. Audit trails must also go beyond tracking AI usage to actively support team alignment and stakeholder buy-in, both in a practical and motivational sense. Large organizations with multiple PMs benefit from knowledge sharing that provides context for AI outputs, allowing team members to reassess feature lists or discover new AI interaction techniques. This is particularly important given that LLMs can perform poorly with certain types of prompts, such as negated prompts [29]. Beyond verification, audit trails build motivation and empathy among team members. Engineers better understand the importance of features

through user feedback, while stakeholders gain insight into the rigorous brainstorming process that informs proposals.

7.3 Enable Secure System Interoperability

PM specific task: Feature and bug fix prioritization

PMs must connect multiple systems and data sources to make informed decisions. Thus, when considering which features or bug fixes to prioritize first, PMs need AI tools that balance the nuance of gathering and adapting to external information while safeguarding internal information that is sensitive to the company. Our study reveals that effective PM work requires seamless yet secure integration between AI tools and existing company infrastructure, data pipelines, and business systems.

7.3.1 Security-Aware Integration. AI tools must integrate securely with existing company systems while protecting sensitive information. Privacy emerged as a critical concern, with AI systems facing vulnerabilities including data poisoning and personal information leakage [16]. Our participants developed various coping strategies—some restricting themselves to company-internal chatbots, while others manually redacted sensitive information before using external tools like ChatGPT, which learns from user inputs [60]. Notably, PMs showed greater trust in AI tools integrated with familiar platforms like Slack, supporting research that known company logos increase perceived trustworthiness on unfamiliar sites [50]. This suggests tools need built-in integration capabilities with approved company systems while maintaining clear boundaries for sensitive data handling. The current lengthy process of approving generative AI tools for corporate use creates significant adoption barriers. Tools need to provide clear security guarantees and integration paths to speed this process.

7.3.2 Business Context and Priority Integration. Tools must connect with existing business systems to support dynamic priority management and decision-making. As business priorities shift, AI tools should allow PMs to iteratively refine and dig deeper on specific ideas or themes, enabling them to explore different avenues. For instance, P3 asked the Q&A widget for more details about pain points identified by the Pain Point Tracker, and P11 broke down existing high level points in one Pain Point Tracker output by reusing them as topics of focus for a new output of the Tracker. AI tools must support an interactive feedback loop for prioritization [81], enabling PMs to delve deeper into feature or bug details to understand how they align with broader business goals or evaluate trade-offs, such as when an engineer notes the high implementation cost of a large feature. Yodeai’s regenerable widgets demonstrated the value of allowing regenerability and maintaining consistent connections to underlying systems while supporting evolution in priorities over time.

7.3.3 Multi-Pipeline Data Flow. Tools must support intelligent integration of multiple data pipelines while maintaining data relevance and freshness. Our study revealed consistent desire for multiplatform data integration, particularly for B2C products where direct customer interviews may be impractical. Unlike general-purpose AI tools, PM-focused systems must enable explicit filtering and source differentiation across pipelines, as insight quality depends heavily on input data relevance. PMs expressed concerns about

overwhelming AI systems with excessive or irrelevant data, which could hinder the extraction of valuable insights. This concern is supported by existing research showing LLMs' sensitivity to irrelevant information in problem solving [73] and RAG applications' susceptibility to distraction by semantically similar but irrelevant data [92]. The timeliness of data is also crucial, as PMs need up-to-date information to assess product performance and evaluate the effectiveness of their actions. Several PMs noted their satisfaction with the time-dependent graphs in the Pain Point Tracker. Therefore, AI tools that incorporate external data should include auto-updating integrations or customizable weighting systems to prioritize certain data sets over others.

7.4 Study Limitations

Our study's depth-first approach, focusing primarily on PMs in software companies and specific tasks within Yodeai, provides valuable insights but may limit broader generalization. While PMs are representative of knowledge work domains and Yodeai's functionalities are applicable to various information synthesis tasks, replicating the study across diverse fields could validate and expand our findings. Participants noted that Yodeai's perceived usefulness depended on data scale and familiarity, suggesting that experiences might differ in production environments with more complex datasets. The short-term nature of our study, while offering in-depth initial insights, may not fully capture long-term benefits and challenges of generative AI in real-world settings. Despite these limitations, our research provides a crucial foundation for understanding immediate reactions to AI-assisted tools in knowledge work, informing both future research directions and the design of more comprehensive, longitudinal studies in this rapidly evolving field.

7.5 Future Work

Future research should expand upon our findings to further optimize AI-powered tools in knowledge work contexts. Longitudinal studies are crucial to investigate the long-term impact of tools like Yodeai on productivity, collaboration, and decision-making. Simultaneously, developing and evaluating strategies to mitigate potential risks and biases associated with AI use in knowledge work settings is essential, focusing on reducing confirmation bias, preventing overreliance on automation, and ensuring ethical deployment. Testing the effectiveness of our proposed design implications—adaptability, accountability, and interoperability—in real-world settings across diverse knowledge work domains will help refine and adapt these principles to better suit various professionals and industries. Additionally, exploring the potential applications of AI-powered tools beyond product management, such as in research, academia, journalism, and consulting, will provide valuable insights into the specific needs, challenges, and opportunities in these domains, leading to the development of more targeted and effective AI solutions for a broader range of knowledge work contexts.

8 Conclusion

Our research investigates the challenges and opportunities of integrating AI-powered tools in knowledge work, focusing on data structuring and decision-making processes. Through the development of Yodeai and user studies with knowledge workers and PMs,

we uncovered the need for AI tools to allow for user control, verification methods, and collaborative approaches. Our findings highlight both the potential benefits of AI assistance in product management and critical limitations. We propose design implications prioritizing adaptability for diverse workflows, personal and collaborative accountability, and context-aware interoperability to address these challenges. As AI continues to evolve and integrate into professional settings, our work provides a foundation for developing user-centric AI tools that augment human capabilities rather than replace them. Moving forward, addressing the identified challenges and opportunities will be crucial in fostering effective human-AI collaboration, ultimately leading to more informed, efficient, and ethical decision-making processes across various knowledge work domains.

References

- [1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fournery, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [2] Marc Annacchino. 2003. *New product development: from initial idea to product management*. Butterworth-Heinemann.
- [3] Ofer Bergman, Ruth Beyth-Marom, and Rafi Nachmias. 2003. The user-subjective approach to personal information management systems. *Journal of the American Society for Information Science and Technology* 54, 9 (2003), 872–878.
- [4] Michael Bernstein, Max Van Kleek, David Karger, and MC Schraefel. 2008. Information scraps: How and why information eludes our personal information management tools. *ACM Transactions on Information Systems (TOIS)* 26, 4 (2008), 1–46.
- [5] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687* (2023).
- [6] Andrea J Bingham and Patricia Witkowsky. 2021. Deductive and inductive approaches to qualitative data analysis. *Analyzing and interpreting qualitative data: After the interview* 1 (2021), 133–146.
- [7] Richard Boardman and M Angela Sasse. 2004. " Stuff goes into the computer and doesn't come out" a cross-tool study of personal information management. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 583–590.
- [8] Michelle Brachman, Amina El-Ashry, Casey Dugan, and Werner Geyer. 2024. How Knowledge Workers Use and Want to Use LLMs in an Enterprise Context. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
- [9] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.
- [10] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).
- [11] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2020. Mesh: Scaffolding comparison tables for online decision making. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 391–405.
- [12] Mei-Ling Chen and Hao-Chuan Wang. 2018. How Personal Experience and Technical Knowledge Affect Using Conversational Agents. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion* (Tokyo, Japan) (*IUI '18 Companion*). Association for Computing Machinery, New York, NY, USA, Article 53, 2 pages. doi:10.1145/3180308.3180362
- [13] Anita Crescenzi, Yuan Li, Yinglong Zhang, and Rob Capra. 2019. Towards Better Support for Exploratory Search through an Investigation of Notes-to-self and Notes-to-share. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 1093–1096.
- [14] Anita Crescenzi, Austin R Ward, Yuan Li, and Rob Capra. 2021. Supporting metacognition during exploratory search with the OrgBox. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 1197–1207.
- [15] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond text generation: Supporting writers with continuous automatic text summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.
- [16] Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2024. Security and privacy challenges of large language models: A survey. *arXiv preprint*

- arXiv:2402.00888 (2024).
- [17] Fabrizio Dell'Acqua, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Kraye, François Candelon, and Karim R Lakhani. 2023. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper* 24-013 (2023).
 - [18] Peter J. Denning. 2006. Infoglut. *Commun. ACM* 49, 7 (jul 2006), 15–19. doi:10.1145/1139922.1139936
 - [19] Eva Eigner and Thorsten Händler. 2024. Determinants of LLM-assisted Decision-Making. *arXiv preprint arXiv:2402.17385* (2024).
 - [20] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130* (2023).
 - [21] Daniel A Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, et al. 2020. Mapping and taking stock of the personal informatics literature. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–38.
 - [22] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social dynamics of AI support in creative writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.
 - [23] Gunnar Harboe and Elaine M Huang. 2015. Real-world affinity diagramming practices: Bridging the paper-digital gap. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 95–104.
 - [24] Md Mahadi Hassan, Alex Knipper, and Shubhra Kanti Karmaker Santu. 2023. Chatgpt as your personal data scientist. *arXiv preprint arXiv:2305.13657* (2023).
 - [25] Marti Hearst. 2009. *Search user interfaces*. Cambridge university press.
 - [26] Omar Hniche and Rachid Saadane. 2023. Challenges and Opportunities of Integrating AI with IFRS in Accounting Systems Insights from Morocco. In *Proceedings of the 6th International Conference on Networking, Intelligent Systems & Security* (, Larache, Morocco,.) (NISS '23). Association for Computing Machinery, New York, NY, USA, Article 61, 9 pages. doi:10.1145/3607720.3607785
 - [27] Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185* (2023).
 - [28] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, LibuSe Hannah VepRek, and Gabrielle Quinn. 2023. Advancing Human-AI complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–29.
 - [29] Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer learning for natural language processing workshop*. PMLR, 52–62.
 - [30] 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*. 1–20.
 - [31] William Jones. 2010. *Keeping found things found: The study and practice of personal information management*. Morgan Kaufmann.
 - [32] William P Jones and Jaime Teevan. 2007. *Personal information management*. University of Washington Press.
 - [33] Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 99–127.
 - [34] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An interactive system for personalized thread-based exploration and organization of scientific literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–15.
 - [35] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–19.
 - [36] Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2024. "I'm categorizing LLM as a productivity tool": Examining ethics of LLM use in HCI research practices. *arXiv preprint arXiv:2403.19876* (2024).
 - [37] Tae Soo Kim, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. Cells, generators, and lenses: Design framework for object-oriented interaction with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–18.
 - [38] Charlotte Kobiella, Yarhy Said Flores López, Franz Waltenberger, Fiona Draxler, and Albrecht Schmidt. 2024. "If the Machine Is As Good As Me, Then What Use Am I?"—How the Use of ChatGPT Changes Young Professionals' Perception of Productivity and Accomplishment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.
 - [39] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. doi:10.1109/VLHCC.2013.6645235
 - [40] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).
 - [41] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez Bermúdez i Badia. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.
 - [42] Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2023. From chatgpt to factgpt: A participatory design study to mitigate the effects of large language model hallucinations on users. In *Proceedings of Mensch und Computer 2023*. 81–90.
 - [43] Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023. MedDM: LLM-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441* (2023).
 - [44] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 557–566.
 - [45] Jie Li, Hancheng Cao, Laura Lin, Youyang Hou, Ruihao Zhu, and Abdallah El Ali. 2023. User Experience Design Professionals' Perceptions of Generative Artificial Intelligence. *arXiv preprint arXiv:2309.15237* (2023).
 - [46] Jie Li, Hancheng Cao, Laura Lin, Youyang Hou, Ruihao Zhu, and Abdallah El Ali. 2024. User experience design professionals' perceptions of generative artificial intelligence. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.
 - [47] Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A Myers. 2019. Unakite: Scaffolding developers' decision-making using the web. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 67–80.
 - [48] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2023. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. *arXiv preprint arXiv:2310.02161* (2023).
 - [49] Tao Long, Katy Ilonka Gero, and Lydia B Chilton. 2024. Not Just Novelty: A Longitudinal Study on Utility and Customization of AI Workflows. *arXiv preprint arXiv:2402.09894* (2024).
 - [50] Paul Benjamin Lowry, Tom Roberts, and Bryce Eryn Caine. 2005. The effect of familiar logos on trust of websites designed for small user interfaces. In *11th International Conference on Human-Computer Interaction (HCI 2005)*, Las Vegas, Nevada, July, Citeseer. 22–27.
 - [51] Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc Le, and Ed Chi. 2023. Beyond ChatBots: ExploreLLM for Structured Thoughts and Personalized Model Responses. *arXiv preprint arXiv:2312.00763* (2023).
 - [52] Yumeng Ma and Jiahao Ren. 2023. ProactiveAgent: Personalized Context-Aware Reminder System. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–3.
 - [53] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.
 - [54] Catherine C Marshall. 2008. Rethinking personal digital archiving, Part 1: Four challenges from the field. *D-Lib Magazine* 14, 3 (2008), 2.
 - [55] Catherine C Marshall. 2008. Rethinking personal digital archiving, part 2: implications for services, applications, and institutions. *D-Lib Magazine* 14, 3 (2008), 3.
 - [56] Imani Munyaka, Zahra Ashktorab, Casey Dugan, James Johnson, and Qian Pan. 2023. Decision Making Strategies and Team Efficacy in Human-AI Teams. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–24.
 - [57] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192.
 - [58] OpenAI. 2023. ChatGPT Plugins. <https://openai.com/blog/chatgpt-plugins>.
 - [59] OpenAI. 2023. Introducing GPTs. <https://openai.com/blog/introducing-gpts>.
 - [60] OpenAI. 2024. Privacy Policy. <https://openai.com/en-GB/policies/privacy-policy/>.
 - [61] Srishti Palani, Yingyi Zhou, Sheldon Zhu, and Steven P Dow. 2022. InterWeave: Presenting Search Suggestions in Context Scaffolds Information Search and Synthesis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
 - [62] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering* (2024).
 - [63] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2022. Designing fair AI in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.

- [64] Witold Pedrycz and Shyi-Ming Chen. 2015. *Granular Computing and Decision-Making: Interactive and Iterative Approaches*. Springer Publishing Company, Incorporated.
- [65] Vijay Pereira, Elias Hadjielias, Michael Christofi, and Demetris Vrontis. 2023. A systematic literature review on the impact of artificial intelligence on workplace outcomes: A multi-process perspective. *Human Resource Management Review* 33, 1 (2023), 100857.
- [66] Jos Pierson, An Jacobs, Katrien Dreessen, B Lievens, I Van den Broeck, and Wendy Van den Broeck. 2006. Walking the interface: uncovering practices through proxy technology assessment. *Proc. of EPIC 2006* (2006), 40–54.
- [67] Crystal Qian and James Wexler. 2024. Take It, Leave It, or Fix It: Measuring Productivity and Trust in Human-AI Collaboration. *arXiv preprint arXiv:2402.18498* (2024).
- [68] Sukumar Rathnam and Michael V Mannino. 1995. Tools for building the human-computer interface of a decision support system. *Decision Support Systems* 13, 1 (1995), 35–59.
- [69] Leonardo Rundo, Roberto Pirrone, Salvatore Vitabile, Evis Sala, and Orazio Gambino. 2020. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. *Journal of biomedical informatics* 108 (2020), 103479.
- [70] Gary R Schirr. 2013. User research for product innovation: Qualitative methods. *The PDMA handbook of new product development* (2013), 231.
- [71] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 99–109.
- [72] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking. *arXiv preprint arXiv:2402.05880* (2024).
- [73] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*. PMLR, 31210–31227.
- [74] Yang Shi, Tian Gao, Xiaohan Jiao, and Nan Cao. 2023. Understanding design collaboration between designers and artificial intelligence: A systematic literature review. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–35.
- [75] Herbert A Simon. 1990. Bounded rationality. *Utility and probability* (1990), 15–18.
- [76] Rolf Sint, Sebastian Schaffert, Stephanie Stroka, and Roland Ferstl. 2009. Combining unstructured, fully structured and semi-structured information in semantic wikis. In *CEUR workshop proceedings*, Vol. 464. Citeseer, 73–87.
- [77] Alexander Smirnov, Andrew Ponomarev, and Tatiana Levashova. 2023. Towards a Methodology for Developing Human-AI Collaborative Decision Support Systems. In *International Conference on Computer-Human Interaction Research and Applications*. Springer, 69–88.
- [78] Lucille Alice Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.
- [79] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2023. Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. *arXiv preprint arXiv:2310.12953* (2023).
- [80] 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*. 1–18.
- [81] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2023. The Metacognitive Demands and Opportunities of Generative AI. *arXiv preprint arXiv:2312.10893* (2023).
- [82] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.
- [83] Justin D Weisz, Jessica He, Michael Muller, Gabriela Hofer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. *arXiv preprint arXiv:2401.14484* (2024).
- [84] Justin D Weisz, Michael Muller, Jessica He, and Stephanie Houde. 2023. Toward general design principles for generative AI applications. *arXiv preprint arXiv:2301.05578* (2023).
- [85] Matt Welsh. 2022. The end of programming. *Commun. ACM* 66, 1 (2022), 34–35.
- [86] Steve Whittaker. 2011. Personal information management: from information consumption to curation. *Annual review of information science and technology* 45, 1 (2011), 1.
- [87] Steve Whittaker, Victoria Bellotti, and Jacek Gwizdzka. 2006. Email in personal information management. *Commun. ACM* 49, 1 (2006), 68–73.
- [88] Steve Whittaker and Julia Hirschberg. 2001. The character, value, and management of personal paper archives. *ACM Transactions on Computer-Human Interaction (TOCHI)* 8, 2 (2001), 150–170.
- [89] Steve Whittaker and Candace Sidner. 1996. Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 276–283.
- [90] Tom D Wilson. 1999. Models in information behaviour research. *Journal of documentation* 55, 3 (1999), 249–270.
- [91] Allison Woodruff, Renee Shelby, Patrick Gage Kelley, Steven Rousso-Schindler, Jamila Smith-Loud, and Lauren Wilcox. 2023. How Knowledge Workers Think Generative AI Will (Not) Transform Their Industries. *arXiv preprint arXiv:2310.06778* (2023).
- [92] Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How Easily do Irrelevant Inputs Skew the Responses of Large Language Models? *arXiv preprint arXiv:2404.03302* (2024).
- [93] Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junjing Zhao, Qian Liu, Che Liu, et al. 2023. Openagents: An open platform for language agents in the wild. *arXiv preprint arXiv:2310.10634* (2023).
- [94] Litao Yan, Miryung Kim, Bjoern Hartmann, Tianyi Zhang, and Elena I Glassman. 2022. Concept-annotated examples for library comparison. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- [95] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.
- [96] Ryan Yen, Nicole Sultanum, and Jian Zhao. 2024. To Search or To Gen? Exploring the Synergy between Generative AI and Web Search in Programming. *arXiv preprint arXiv:2402.00764* (2024).
- [97] Ryan Yen, Jiawen Zhu, Sangho Suh, Haijun Xia, and Jian Zhao. 2023. CoLadder: Supporting Programmers with Hierarchical Code Generation in Multi-Level Abstraction. *arXiv preprint arXiv:2310.08699* (2023).
- [98] Yuliang Yun, Dexin Ma, and Meihong Yang. 2021. Human-computer interaction-based decision support system with applications in data mining. *Future Generation Computer Systems* 114 (2021), 285–289.
- [99] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [100] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M Padilla, Jeffrey Caterino, Ping Zhang, et al. 2023. Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis. *arXiv preprint arXiv:2309.12368* (2023).
- [101] Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. 2023. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984* (2023).