DOI:10.1145/3655615

AI agents are extending the capabilities of traditional search engines to help users tackle complex tasks.

BY RYEN W. WHITE

Advancing the Search Frontier with AI Agents

AS MANY OF US in the information retrieval (IR) research community know and appreciate, search is far from being a solved problem. Millions of people struggle with tasks on search engines every day. Often, their struggles relate to the intrinsic complexity of their task and the failure of search systems to fully understand it and serve relevant results.[38] The task motivates the search, creating the problem that searchers attempt to solve, and drives search behavior as they work through different task facets. Complex search tasks require more than support for rudimentary fact finding or re-finding. Research on methods to support complex tasks includes work on generating query and website suggestions,[15] personalizing and contextualizing search,[4] and developing new search experiences, including those that span time and space.[1,40] The recent emergence of generative artificial intelligence and the arrival of assistive *agents* based on this technology have the potential to offer further assistance to searchers, especially those engaged in complex tasks. These advances have profound implications for the design of intelligent systems and for the future of search itself. This article, based on a keynote given by the author at the 2023 ACM SIGIR Conference, explores these issues and how AI agents are advancing the frontier of search-system capabilities, with a special focus on information interaction and complex task completion.

Taking Search to Task

Tasks are a critical part of people's daily lives. The market for dedicated task applications that help people with their "to do" lists is likely to grow significantly—effectively tripling in size—over the next few years.[a] There are many examples of such applications that can help both individuals (for example, Microsoft To Do, Google Tasks, Todoist) and teams (for example, Asana, Trello, Monday.com) tackle their tasks more effectively. Over

a https://www.verifiedmarketresearch.com/product/task-management-software-market/

» key insights

- **Search is an unsolved problem; millions of people struggle with complex tasks on Web search engines every day.**
- **The emergence of generative AI and assistive agents based on it promise to revolutionize task completion, aiding users in navigating and resolving complex search tasks.**
- **These advancements present both challenges and extraordinary opportunities to redefine the landscape of information access and use, propelling search toward new horizons.**

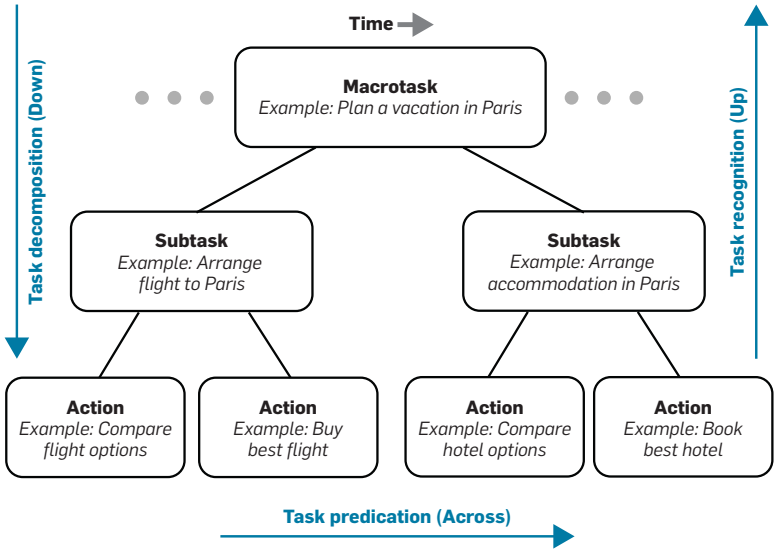
IMAGE BY LANA PO

time, these systems will increasingly integrate AI to better help their users capture, manage, and complete their tasks. In information access scenarios such as search, tasks play an important role in motivating searching via people's gaps in knowledge and problem-solving needs.[3,11] AI can be central in these search scenarios too, especially in assisting with complex search tasks.

Tasks in search. Tasks drive the search process. The IR and information science communities have long studied tasks in search[26] and many information-seeking models consider the role of tasks directly.[3,11] Prior research has explored the different stages of task execution (for example, pre-focus, focus formation, post-focus), task levels, task facets, tasks defined on intents (for example, informational, transactional, and navigational; well-defined or ill-defined; and lookup, learn, or investigate), the hierarchical structure of tasks, the characteristics of tasks, the attributes of task-searcher interaction (for example, task difficulty), and, a focus of this article, task complexity.[8]

As a useful framing device to help conceptualize tasks and develop system support for them, tasks can be represented as *trees* comprising macrotasks (high-level goals), subtasks (specific components of those goals), and actions (specific steps taken by searchers toward the completion of those components).[26] Figure 1 presents an example of a "task tree" for a task involving an upcoming vacation to Paris. Included are examples of macrotasks, subtasks, and actions. Moves around this tree correspond to different task applications such as task recognition (up), task decomposition (down), and task prediction (across). Only actions (for example, queries, clicks, and so on) are directly observable to traditional search engines. However, with recent advances in AI agents—primarily more support for natural language interactions to improve alignment between searchers and AI agents, but also an increase in system awareness of short- and long-term contexts—more aspects of macrotasks and subtasks are becoming visible to and more fully understood by search systems. Challenges in working with tasks include how to represent them within search systems, how to observe more task-relevant activity and content to develop richer task models, and how to develop task-oriented interfaces that place tasks and their completion at the forefront of user engagement. Task complexity deserves a special focus in this article given the challenges that searchers can still face with complex tasks and the significant potential of AI to help searchers resolve them.

Figure 1. Task-tree representation for a complex task involving planning a vacation to Paris, France. The tree depicts different task granularities (macrotask, subtask, action) and different task applications (decomposition, prediction, recognition) as moves around the tree. Time progresses from left to right via a sequence of searcher actions (queries, result clicks, pagination, and so on). Only actions are observable in traditional search engines. Aspects of subtasks and macrotasks may be observable to AI agents when searchers provide higher-level descriptions of their goals in natural language.



Complex search tasks. Recent estimates suggest that half of all Web searches are not answered.[b] Many of those searches are connected to complex search tasks. These tasks are ill-defined and/or multi-step; span multiple queries, sessions, and/or devices; and require deep engagement with search engines (involving many queries, backtracking, branching, and so on) to complete them.[15] Complex tasks also often have many facets and cognitive dimensions, and are closely connected to searcher characteristics such as domain expertise and task familiarity.[38]

To date, there have been significant attempts to support complex search tasks via humans, such as librarians and subject matter experts, and search systems, including both general Web search engines and those tailored to specific industry verticals or domains. The main technological progress so far has been in areas such as query suggestion and contextual search, with new experiences being developed that utilize multiple devices, provide cross-session support, and enable conversational search. We are also now seeing an emerging wave of search-related technologies in the area of generative AI.[23]

Before proceeding, let us dive into the different types of existing and emerging search support for complex tasks in more detail:

▸ *Suggestions, personalization, and contextualization:* Researchers and practitioners have long developed and deployed support such as query suggestion and trail suggestion (see, for example, Hassan et al.[15] and Singla et al.[28]), including providing guided tours and suggesting popular trail destinations as ways to find relevant resources. This coincides with work on contextual search and personalized

b https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

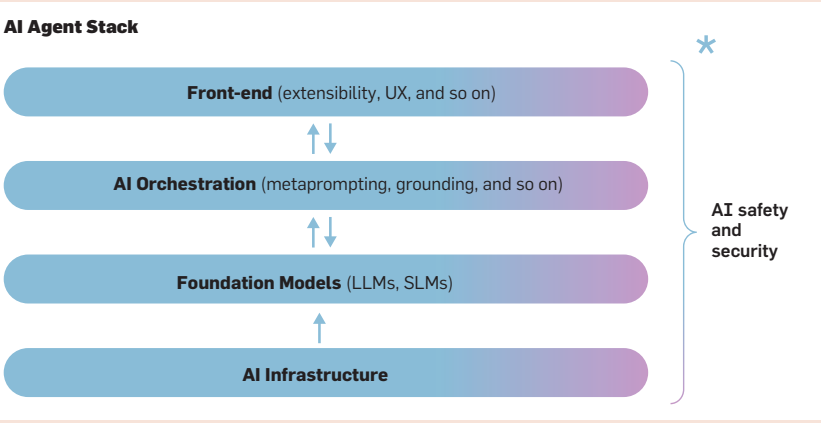
search (for example, Bennett et al.[4] and Teevan et al.[30]), where search systems can use data from the current searcher such as session activity, location, and reading level, as well as the searcher's long-term activity history to provide more-relevant results. Search engines may also use cohort activities to help with cold-start problems for new users and augment personal profiles for more-established searchers.[31]

▸ *Multi-device, cross-device, and cross-session:* Devices have different capabilities and can be used in different settings. Multi-device experiences (for example, those discussed in White et al.[40]) utilize multiple devices *simultaneously* to better support complex tasks such as recipe preparation, auto repair, and home improvement that have been decomposed into steps manually or automatically.[44] Cross-device and cross-session support[1,37] can help with ongoing or background searches for complex tasks that persist over space and time. For example, being able to predict task continuation can help with "slow search" applications that focus more on result quality than on the near instantaneous retrieval of search results.[29]

▸ *Conversational experiences and generative AI:* Natural language is an expressive and powerful means of communicating intentions and preferences with search systems. The introduction of clarification questions on search engine result pages (SERPs),[42] progress on conversational search,[13] and even "conversations" with documents, where searchers can inquire about document content via natural language dialogue,[32] enable these systems to more fully engage with searchers to better understand their tasks and goals. There are now many emerging opportunities to improve the alignment between search systems and their users, and support more tasks, via large-scale foundation models such as OpenAI's GPT, Google's Gemini, and Meta's Llama, including offering conversational task assistance via chatbots such as ChatGPT.

All of these advances, and others, have paved the way for the emergence of a new class of generative-AI-powered assistive agents that can help people make progress in their complex search tasks.

Figure 2. AI agent stack depicting the various layers and the important role of AI safety and security across the stack. Foundation models can be either large language models (LLMs), with trillions of parameters, or small language models (SLMs), with just a few billion parameters. The star (*) symbolizes that there can be fleets of cooperating agents (discussed in the "Multi-agent" section).



AI Agents

Agents are applications of modern AI (that is, AI based on foundation models and similar technologies) to help people with complex cognitive tasks. At Microsoft, we refer to these as *copilots*, which work alongside humans to empower them and amplify their cognitive capabilities.[c] Copilots have conversational user interfaces that allow their users to engage with them via natural language; are powered by foundation models such as GPT-4; are extensible with skills, tools, and plugins; and are scoped to specialized domains or applications (including search). Copilots are designed to keep humans at the center of the task-completion process and augment human capabilities to help them complete a broader range of tasks in less time and with less effort.

The general AI agent stack (Figure 2) contains four layers. The *frontend* covers the user experience and extensibility with plug-ins, enabling developers to provide additional visible tools to the agent. The *AI orchestration* layer handles the internal information flows, prompting, grounding, and executing any tools or plugins and processing their responses, among other things. Agents leverage the power of large *foundation models* that can be provided to the developer as is or specialized to specific tasks, domains, or applications; developers can also bring their own models to use to power agent functionality. This all runs on top of massive-scale *AI infrastructure* hosted in the cloud on platforms such as Microsoft Azure, Google Cloud, and Amazon Web Services. Underpinning all of this is a need for a strong commitment to responsible AI, which ensures that agents are safe, secure, and transparent. We can do this via an iterative, layered approach with mitigations spanning the model, prompts, grounding, and user experience.

AI agents can, among other things, help users attain goals, maximize utility, and perform automated functions. Examples of these agents include the Apple Siri and Amazon Alexa personal digital assistants that can answer questions and assist with task management; GitHub Copilot,[d] an AI pair programmer that can reduce developer effort, enable more task success, and significantly expedite task completion; and Auto-GPT,[e] a fully autonomous agent that can decompose tasks into sub-tasks and execute them independently on a user's behalf to support goal attainment.

AI agents are also emerging in search systems. Popular Web search engines such as Bing and Google are adding agent functionalities in the form of conversational assistance: Bing has Copilot, mentioned earlier, and Google has Gemini, a similar service. In

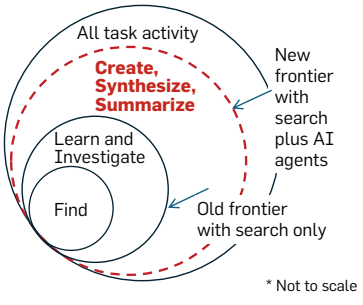
c https://copilot.microsoft.com

d https://github.com/features/copilot
e https://auto-gpt.ai/

search, agents can help searchers tackle a broader range of tasks than information finding and go deeper than surface (that is, SERP-level) interactions with content by synthesizing answers on the searcher's behalf. They also enable searchers to communicate their intents and goals more directly. Returning to the task tree (Figure 1), the focus on engaging agents via natural language interactions allows both searchers and systems to consider higher-level task representations (macrotasks, subtasks) in addition to the more granular actions (queries, result clicks, pagination, and so on) that searchers already perform when engaging with traditional search engines.

Agents in search. Agents and chat experiences are a complement to, not a replacement for, traditional search engines. Search engines have existed for decades and serve a valuable purpose: providing near-instantaneous access to answers and resources for a broad range of search requests. These existing and emerging modalities can and should work well together to help searchers tackle a wider range of tasks.

Figure 3. Advancing the search frontier. Visualizing the set of possible tasks that can be tackled with search only today (that is, finding, learning, and investigating) plus the expansion on the frontier into support for higher-order task activities with the addition of AI agents (for example, adding AI support for creative inspiration, synthesis, and summarization).



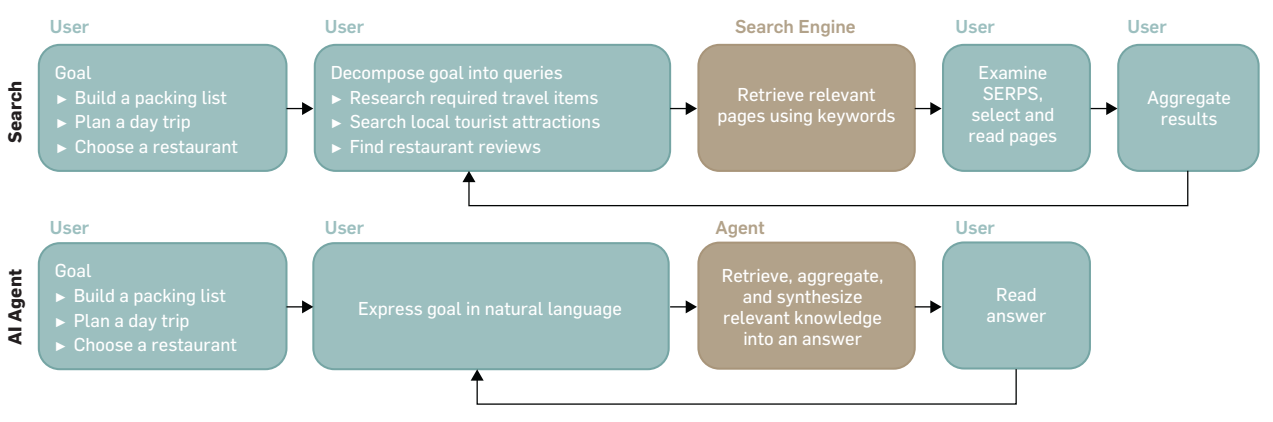
The ability of agents to better understand intentions and provide assistance beyond fact finding and basic learning or investigation will advance the *search frontier* (that is, what search systems are capable of and what types of tasks they can support), broadening the range of tasks that searchers can complete. These might include, for example, direct support for tasks requiring creative inspiration, summarizing existing perspectives, or synthesizing those perspectives to generate new insights (Figure 3). This moves us toward more-intelligent search systems that can help with all task completion, covering the full universe of tasks for which people might need search support, including actuation capabilities to act on tasks in the digital and physical worlds.

One way to define the range of tasks that agents can support is through Bloom's taxonomy of learning objectives.[17] At the pinnacle of that taxonomy is creation, which we have only scratched the surface in supporting with next-word prediction using transformer models. We are already seeing expansions into content types beyond text (images, video, audio, and so on) and could consider support for other creative tasks including planning, analysis, and invention. There are also many other layers in Bloom's taxonomy (for example, evaluation-help searchers make judgments and decisions; application-help searchers complete new tasks, understanding and explaining ideas and concepts to accelerate learning) that could form the basis for future search frontiers.

Beyond offering greater capabilities, the introduction of AI agents into search will also change how people engage with search systems. In agents, the primary mode of interaction is natural language, with some recent support for other input and output modes via the introduction of image- and video-generation models such as Stable Diffusion, DALL-E, and Sora. Agents can generate direct answers, with source attribution for provenance, to build trust with users and drive traffic back to content creators, which is important to incentivize further content creation that will fuel future foundation models.

The overall search interaction flow is also different between search engines and AI agents. When using agents, searchers do not need to decompose their goal into sub-goals or sub-queries, examine SERPs and landing pages, and aggregate or synthesize relevant knowledge from retrieved information. Continuing our running example macrotask of vacation planning from earlier, Figure 4 compares information interaction in the two modalities for some task-related goals. In AI agents, the responsibility for generating answers is delegated by the searcher to the system, which poses

Figure 4. Information interactions in a traditional search engine versus an AI agent.



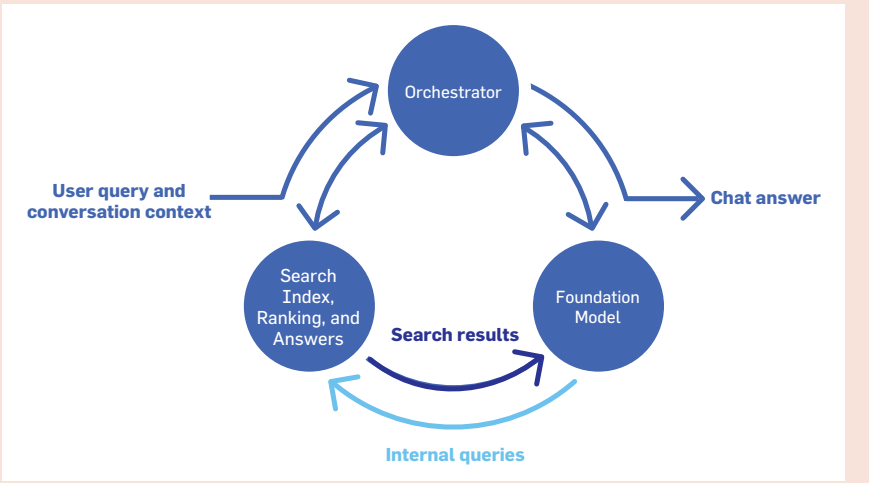
challenges in terms of human control and human learning, discussed later in this article.

Adding agents to search engines. It is neither practical nor necessary to deploy AI agents for all search tasks. Foundation model inference is expensive at massive scale and search engine algorithms have been honed over decades to provide relevant results for a broad range of tasks such as navigation and fact-finding. Conversational interfaces are less familiar for searchers, so it will take time for searchers to adapt to this way of searching. Traditional search engines are sufficient when searchers know exactly what they want. Agents are helpful for more-complex search tasks or in situations where searchers may be struggling to find relevant information. Task complexity can be estimated using aggregate metric, such as the amount of engagement with the search engine (for example, number of query reformulations) for similar tasks historically. As generative AI appears in more applications and searchers better understand agent capabilities, the tasks that searchers bring to agents deployed in search settings will likely evolve and expand, and may well increase in complexity.

We will also see a growth in search experiences that unify traditional search and agents. In a step toward this, search engines such as Bing and Google are already integrating dynamic answers from foundation models into their SERPs for some queries (for example, the AI Overviews in Google's so-called Search Generative Experience[f]). In these experiences, search results and other answers can be shown together on the SERP with answers from generative AI, allowing searchers to easily engage with them as desired, including asking follow-up questions inline. There are also ways for searchers to move between modalities based on their task and personal preferences. AI agents can also provide searchers with control over other aspects; for example, Bing offers an ability to adjust conversation style and tone, although it is not clear that searchers are sufficiently familiar with agents at this time to use these more nuanced controls effectively.

f https://www.google.com/sge

Figure 5. High-level overview of the typical generative AI search process in search engines. The query and the context are passed to the orchestrator, which coordinates with the foundation model to create internal queries and generate answers. The orchestrator may also integrate content (for example, search results and direct answers) from the search engine.



Search agents such as Microsoft Copilot and Google Gemini use retrieval-augmented generation (RAG)[18] to ground agent responses via timely and relevant results. Using RAG has many advantages, including: There is no need to retrain massive foundation models over time; search results provide relevant and fresh information to foundation models; and it provides a provenance signal connecting generated content with online sources. In response to a searcher prompt, the foundation model generates internal queries iteratively that are used to retrieve the results that form context for the agent answers created using generative AI. Displaying these queries to searchers inline in dialogue, as has been the case in Copilot, creates greater transparency and helps build trust with searchers that the system is understanding their tasks and goals. The orchestrator can also pull in relevant instant answers from the search engine, such as weather, stocks, and sports, and display those in agent responses instead of or in addition to the answers generated by the foundation model. Figure 5 shows the high-level search process from query (plus conversation context) to answer, and the role of various key system components.

AI agents also enable search engines to support more complex search tasks. Using search alone would require more searcher effort to examine search results and manually generate answers or insights (see recent work on the Delphic costs and benefits of search[6]). Of course, there are different perspectives on task complexity, such as the agent perspective, denoting the amount of computation, requests, and so on required for the system to complete the task, and the searcher perspective, denoting the amount of manual effort required for the human searcher to generate an answer and complete a task. The accompanying table considers the task complexity from these two different perspectives and (again, drawing from Bloom's taxonomy) provides some current, anecdotal examples of the types of tasks that both searcher and systems may find to be more or less complex. Assuming that foundation model costs will drop and sophistication will increase, we focus here on the task complexity for searchers.

Challenges

Despite the promise of AI agents to dramatically improve information literacy, there are significant challenges in deploying them in search systems at scale that we must find ways to overcome. These include issues with the agent output shown in response to searcher requests, the impacts that the agents can have on searchers, and shifts in the degree of control that humans have in the search process that come from introducing agents:

Table. Anecdotal examples of high-level tasks from Bloom's taxonomy of varying complexities from searcher/AI agent perspectives. Tasks such as "Find" and "Analyze" have similar complexities for both humans and machines. It is easier for machines to create content than for humans, but more difficult for machines to verify the correctness of information.

		Searcher	
		Low	***High***
AI Agent	*Low*	Find Recognize List Define	Create Evaluate Compare Predict
	High	Verify Decide Teach Plan	Analyze Investigate Solve Invent

▸ *Hallucinations:* Searchers rely a lot on the answers from agents, but those answers can be erroneous or nonsensical. So-called hallucination is a well-studied problem in foundation models. Agents can hallucinate for many reasons, a main one being gaps in the training data. RAG, discussed earlier, is a way to help address this by ensuring that the agent has access to up-to-date, relevant information at inference time to help ground its responses. Injecting knowledge from other external sources, such as knowledge graphs and Wikipedia, can also help improve the accuracy of agent responses. An issue related to agents surfacing misinformation is toxicity (offensive or harmful content), which can also be present in the agent output and must be mitigated before answers are shown to searchers.

▸ *Biases:* Biases in the training data, such as social biases and stereotypes,[20] affect the output of foundation models and hence the answers provided by agents. Synthesizing content from different sources can amplify biases in this data. Agents are also subject to biases from learning from their own or other AI-generated content via feedback loops: Biased historical sequences lead to biased downstream models. Agents may also amplify existing cognitive biases, such as confirmation bias, by favoring responses that align with searchers' existing beliefs and values and by providing responses that are optimized to keep searchers engaged with the agent, regardless of the consequences for the searcher.

▸ *Human learning:* Learning may be affected by the use of AI agents, since they remove the need for searchers to engage as fully with the search system and the information retrieved. Learning is already a core part of the search process.[21,35] Both exploratory search and search as learning involve considerable time and effort in finding and examining relevant content. While this could be viewed as a cost, this deep exposure to content also helps people learn. As mentioned earlier, agent users can ask richer questions, allowing them to specify their tasks and goals more fully, but they then receive synthesized answers generated by the agent, creating fewer new—or simply different—learning opportunities for humans that must be understood.

▸ *Human control:* Supporting search requires considering the degree of searcher involvement in the search process, which varies depending on the search task.[2] Agents enable more-strategic, higher-order actions (that is, higher up the "task tree" in Figure 1) than typical search systems. It is clear that searchers want control over the search process. They want to know what information is and is not being included and why. This helps them understand and trust the system output. As things stand, searchers delegate full control of answer generation to the AI, but the rest is mixed, with searchers having less control of search mechanics (for example, queries) but more control of task specifications (via natural language and dialogue). There is more than just a basic tension between automation and control. In reality, it is not a zero-sum game. Agent designers need to ensure human control while increasing automation.[27] New frameworks for multi-agent task completion are moving in this direction,[41] with agents and humans working together synergistically to decompose and tackle complex tasks.

Overall, these are just a few of the challenges that affect the viability of AI agents in search settings. There are other challenges, such as searchers' deeply ingrained search habits, that may be a barrier to their adoption of new search functionality despite the clear benefits to them from embracing agent technologies.

Opportunities

For some time, scholars have argued that the future of information interaction will involve personal search assistants with advanced capabilities, including natural language input, rich sensing, user/task/world models, and reactive and proactive experiences.[38] Technology is catching up with this vision. Opportunities going forward can be grouped into four areas: model innovation, next-generation experiences, measurement, and broader implications. The opportunities are summarized in Figure 6. There are likely more such opportunities not listed here, but the long list shown in the figure is a reasonable starting point for scientists and practitioners interested in working in this area.

Model innovation. There are many opportunities to better model search situations and augment and adapt foundation models to better align with searchers' tasks and goals and provide more-accurate answers. Agents can leverage these model enhancements to improve the support that they provide for complex search tasks. We now present more detail on each opportunity.

Task modeling: Build richer task models that more fully represent tasks and task contexts. This includes how we infer tasks and intent—for example, from the textual content of the search process, from user-system interactions, and from other situational and contextual information such as location, time, and application usage—and how we represent those tasks internally—for example, as a hierarchy (Figure 1) or a more abstract representation (semantic vectors, graph embeddings, Markov models, and so on). We also need to be able to estimate key task

characteristics such as task complexity, which, in one use, can help search systems route requests to the most appropriate interaction modality. In addition, we need to find ways for agents to collect more, and more-accurate, user and world knowledge, both in general and specifically related to the task at hand. A better understanding of the short- and long-term task context will help agents more accurately model the tasks themselves.

Alignment: Develop methods to continuously align agents to tasks, goals, and values via feedback. Here, feedback includes conversation content, such as searchers expressing positive sentiments like gratitude to the agent, or explicit feedback on agent answers via likes and dislikes. The performance of agents lacking alignment will remain fixed over time. Agents need application-aligned feedback loops to better understand searcher goals and tasks and must use that feedback to continuously improve answer accuracy and relevance. Beyond research on fine-tuning foundation models from human feedback (for example, likes and dislikes),[45] we can also build on learnings from research on implicit feedback in IR, including work on improving ranking algorithms via SERP clicks[16] and developing specialized interfaces to capture user feedback.[38]

Augmentation: Enhance agents with relevant external knowledge and enhanced tools and capabilities. As mentioned earlier, RAG is a common form of knowledge injection for foundation models. Relevance models are tuned to maximize user benefit, not for agent consumption. We need to evaluate whether this difference is meaningful practically, and if so, develop new ranking criteria that consider the intended consumer of the search results (human or machine). Despite their incredible capabilities, foundation models still have shortcomings that manifest in the agents that use them. We need to understand these shortcomings through principled evaluation and find ways to leverage external skills or plug-ins to address them. Agents must find and recommend skills per task demands,[39] for example, invoking Wolfram for computational assistance. We can also integrate tool use directly into tool-augmented models, such as Toolformer,[25] that can teach themselves to use tools. Models of task context may also be incomplete, so we should invest in ways to better *ground* agent responses via context using, for example, richer sensing, context filtering, and dynamic prompting.

We need to find ways for agents to collect more, and more-accurate, user and world knowledge, both in general and specifically related to the task at hand.

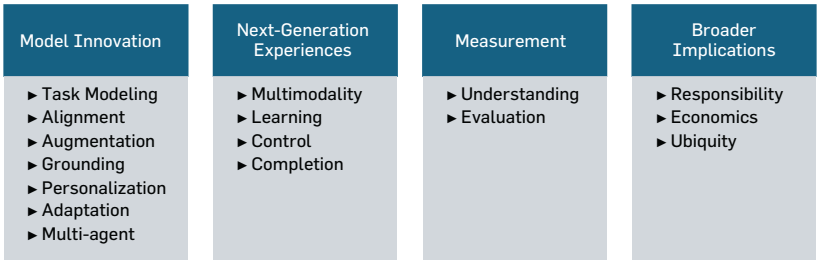
Grounding: Apply use-case-specific information to reduce hallucinations, build trust, and support content creators. It is in the interests of agents, searchers, and content creators (and providers and advertisers) to consider the source of the data used in generating answers. Provenance is critical. Agents should provide links back to relevant sources, preferably with specific details or URLs, to help establish and maintain user trust, provide attribution for content creators, and drive engagement for content providers and advertisers. To build trust and support learning, it is also important for agents to practice faithful reasoning[10] and provide interpretable reasoning traces (for example, explanations with chain-of-thought descriptions) along with their answers. We should also consider how we can integrate agents within existing experiences (for example, in other applications) to ground answers in their context of use.

Personalization: Develop personal agents that can understand searchers and their tasks while using personal data, privately and securely. Searchers bring their personal tasks to search systems, as they will with agents. Here are some example personal prompts that describe the types of personal tasks that searchers might expect an agent to handle:

- Write an email to my client in my personal style with a description of the quote in the attached doc.
- Tell me what's important for me to know about the company town hall that I missed.
- Where should I go for lunch today?

These tasks, spanning creation, summarization, and recommendation, quickly illustrate the wide range of expectations people may have from their *personal* agents. As part of developing such personalized AI support, we need to do two key things. The first is to study foundation model *capabilities*, including their ability to identify task-relevant information in personal data and activity histories, and model

Figure 6. Selected opportunities for progress on AI agents in search settings. These opportunities are only a first step in this direction but the figure highlights the many avenues for impact in this area.



user knowledge in the current task and topic. The second is to develop core *technologies*, including the following:

▸ Infinite memory, using relevant long-term activity (in IR, there has been considerable research on relevant areas such as re-finding[34] and personalization[30])

▸ Context compression, to fit more context into finite token limits (for example, using turn-by-turn summarization rather than raw conversational content)

▸ Privacy, including mitigations such as differential privacy and federated learning, as well as machine unlearning[5] to intentionally forget irrelevant information over time, including sensitive information that the searcher may have explicitly asked to be removed from trained models, and also remove irrelevant or unwanted data from agent memory.

Adaptation. Two main forms of adaptation that we consider here are model specialization and so-called adaptive computation:

▸ *Model specialization: Develop specialized foundation models for search tasks that are controllable and efficient.* Large foundation models are generalists and have a wide capability surface. Specializing these models for specific tasks and applications discards useless knowledge, making the models more accurate and efficient for the task at hand. Recent advances in this area have yielded strong performance; for example, the Orca-13B model[22] uses explanation-based tuning, where the model explains the steps used to achieve its output and those explanations are used to train a small language model (SLM), to outperform state-of-the-art models of a similar size, such as Vicuna-13B.[9] Future work could explore guiding specialization via search data, including anonymized large-scale search logs as well as algorithmic advances in searcher-preference modeling and continual learning.

▸ *Adaptive computation: Develop methods to adaptively apply different models per task and application demands.* Adaptive compute involves using multiple foundation models (for example, GPT and a specialized model), each with different inference-time constraints, primarily around speed, capabilities, and cost, and learning which model to apply for a given task. The specialized model can back off to one or more larger models as needed, per task demands. The input can be the task plus the constraints of the application scenario under which the model must operate. Human feedback can also be used to refine the adaptation strategy over time.[43]

These adaptation methods will yield more-effective and more-efficient AI capabilities that agents can use to help searchers across a range of settings, including in offline settings (for example, on-device only).

Multi-agent: Utilize multiple specialized agents working together and with humans to help complete a search task. Multiple agents have been shown to help encourage divergent thinking, improve factuality and reasoning, and provide guardrails for AI systems. Multi-agent systems such as AutoGen[41] orchestrate communication between agents to help users complete tasks more effectively. These systems could be used in search settings to, for example, retrieve relevant resources from diverse sources, improve answer correctness by critiquing and refining AI-generated output, improve human decision making by presenting alternative solutions, and even automate the completion of some tasks or sub-tasks, with humans in the loop throughout.

Next-generation experiences. Advancing models is necessary but not sufficient given the central role that interaction plays in the search process.[38] There are many opportunities to develop new search experiences that capitalize on agent capabilities while keeping searchers in control.

Multimodality: Develop experiences bridging (at least) the search and agent (chat) modalities, offering explanations and suggestions. Given how entrenched and popular traditional search is, it is likely that some form of query-result interaction will remain a core part of how we find information online. Future, agent-enhanced experiences may reflect a more-seamless combination of the two interaction modalities in a unified experience. Both Google and Bing are taking a step in that direction by unifying search results and agent answers in a single interface. Explanations on which modality and style (for example, creative, balanced, or precise) perform best and when will help searchers make decisions about which modalities and settings to use. Modality recommendation based on task is also worth exploring: Simple tasks may need only traditional search, whereas complex tasks may need agents. Contextualization and personalization will also play an important part in deciding how much information is needed from the searcher (incurring interaction cost but yielding greater control) and how much can be reliably inferred from signals already available to the system. Related to this are opportunities around conversation-style suggestion given the current task. For example, a simple fact-finding task needs short, precise replies (when generative-AI-powered agents can often be verbose), while generating new content needs creativity (when agent responses can often be unoriginal or bland). Search providers could also consider offering a single point of entry and an automatic routing mechanism to direct requests to the correct modality given inferences about the underlying task

(see "Task modeling" section above) and the appropriateness of each of the modalities for that task. Beyond search and chat, other modalities to help support complex search tasks may include third-party tools and applications, bespoke user interfaces (for example, tailored dynamically by the agent to the task at hand), interactive visualizations, and proactive recommendations.

Human learning: Develop agents that can detect learning-related search tasks and support relevant learning activities. As mentioned earlier, agents can remove or change human learning opportunities by their automated generation and provision of answers. Learning is a core outcome of information seeking.[11,21,35] We need to develop agents that can detect learning and sensemaking tasks, and support relevant learning activities via agent experiences that, for example, provide detailed explanations and reasoning, offer links to learning resources such as instructional videos, enable deep engagement with task content via relevant sources, and support specifying and attaining learning objectives. A good example of all of this is Microsoft's recently announced partnership with the Khan Academy.[g]

Human control: Better understand control and develop agents with control while growing automation. Control is an essential aspect of searcher interaction with agents. Agents should consult humans to resolve or codify value tensions. Agents should be in collaboration mode by default and must only take control with the permission of stakeholders. Experiences that provide searchers with more agency, such as adjusting the specificity or diversity in agent answers, are critical, leading to less generality and less repetition. As mentioned in the "Grounding" section above, citations in answers are important. Humans need to be able to verify citation correctness in a lightweight way, ideally without leaving the user experience; Gemini now offers an ability to manually dive deeper and verify answers. We also need a set of user studies to understand the implications of providing less control over some aspects, such as answer generation; more control over other aspects, such as macrotask specification; and control over new aspects, such as conversation style and tone (as with Copilot).

Completion: Agents should help searchers complete tasks while keeping searchers in control. We need to both expand the search frontier by adding or discovering more capabilities of foundation models that can be surfaced through agents *and* deepen task capabilities so that agents can help searchers better complete more tasks. We are moving from a world equipped with only search engines, to one also equipped with AI-powered answer engines, with agents that provide relevant information, synthesized from several sources, and action engines that can perform actions to complete tasks, or help find agents to do so. We can view skills and plugins as actuators of the digital world and should help foundation models fully utilize them. We need to start simple (for example, reservations), learn and iterate, and increase task complexity as model capabilities improve with time.

The standard mode of engagement with AI agents is reactive; users send requests and the agents respond. Agents should ideally have a *dynamic interaction model* that tailors the interface to the task and the context. With this model, agents can take initiative, with permission, and provide updates for standing tasks, or they can offer proactive suggestions or take actions directly when agent uncertainty is low. Agents can also help support task planning (decomposition, prioritization, and so on) for complex tasks such as travel or events. AI can already help complete repetitive tasks (for example, action transformers trained on digital tools[h]) and create and apply "tasklets" (user interface scripts) learned from websites.[19]

Given the centrality of information interaction in search task completion, it is important to focus sufficient attention on interaction models and experiences in AI agents. In doing so, we must also carefully consider the implications of critical decisions on issues that affect AI in general, such as control and automation.

Measurement. Another important direction is in measuring AI agent performance, understanding agent impact and capabilities, and tracking agent evolution over time. Many of the challenges and opportunities in this area, such as non-determinism, saturated benchmarks, and inadequate metrics, also affect the evaluation of foundation models in general.

Understanding: Deeply understand agent capabilities and agent impact on searchers and their tasks. We have only scratched the surface in understanding AI agents and their impact. Gaining a deeper understanding could take a few forms. The first is *user understanding*, covering the mental models of agents and the effects of bias (for example, functional fixedness) on how agents are adopted and used in search settings. It also covers changes in search behavior and information seeking strategies, including measuring changes in effects across modalities, for example, search *versus* agents and search *plus* agents. There are also opportunities in using foundation models to understand search interactions via user studies and generate intent taxonomies and classify intents from log data. The second form is *task understanding*, covering the intents and tasks for which agents are used and most effective. Finally, there is *agent understanding*, covering the capabilities and limitations of agents. This form of inquiry is similar to that found in the "Sparks of AGI" paper on GPT-4,[7] which examined foundation model capabilities in depth.

Evaluation: Identify and develop metrics for agent evaluation, while considering important factors, and find applications of agent components for IR evaluation. There are many options for AI agent metrics, including feedback, engagement, precision-recall, generation quality, and answer accuracy. Given the task focus, metrics should likely target the task holistically (for example, success, effort, satisfaction). In evaluating agents in search settings, it is also important to consider the following:

- Repeatability: Non-determinism can make agents difficult to evaluate and debug
- Interplay between search and

g https://news.microsoft.com/source/features/ai/khan-academy-and-microsoft-partner-to-expand-access-to-ai-tools

h https://www.adept.ai/blog/act-1

agents, including switching, joint task success, and so on

- Longer-term effects on user capabilities and productivity
- Task characteristics, such as complexity
- New benchmarks: Agents are affected by external data, grounding, queries, and so on.

There are also opportunities to consider applications of agent components for IR evaluation. Foundation models can predict searcher preferences[33] and assist with relevance judgments,[12] including generating explanations for judges. Also, foundation models can create powerful searcher simulations that can mimic human behavior and values, expanding on early work on searcher simulations in IR.

Measuring agent performance is essential in understanding their utility and improving their performance over time. Agents do not function in a vacuum, and we must consider the broader implications of their deployment for complex tasks in search settings.

Broader implications. AI agents must operate in a complex and dynamic world. There are several opportunities beyond advances in technology and deepening our understanding of agent performance and capabilities.

Responsibility: Understand factors affecting reliability, safety, fairness, and inclusion in agent usage in search. The broad reach of search engines means that AI agents have a critical obligation to act responsibly. Research is needed on ways to improve answer accuracy via better grounding in more-reliable data sources; developing guardrails; understanding biases in foundation models, prompts, and the data used for grounding; and understanding how well agents work in different contexts, with different tasks, and with different people and cohorts. Red teaming, user testing, and feedback loops are all needed to determine emerging risks in agents and the foundation models that underlie them. This also builds on existing work on responsible AI, responsible IR, and FACTS-IR, which has studied biases and harms and ways to mitigate them.[24]

Economics: Understand and expand the economic impact of agents in search. This includes exploring new business models that agents will create beyond information finding. Advancing the search frontier from information finding deeper into task completion (for example, into creation and analysis) creates new business opportunities. It also unlocks new opportunities for advertising, including advertisements that are shown inline with dialogue or answers and are contextually relevant to the current conversation. There is also a need to more deeply understand the impact of agents on content creation and search engine optimization. Content attribution is vital in such scenarios to ensure that content creators (and advertisers and publishers) can still generate returns. We should avoid the so-called paradox of reuse,[36] where fewer visits to online content leads to less content being created, which in turn leads to worse models over time. Another important aspect of economics is the cost-benefit tradeoff, related to work on adaptation (see "Adaptation" section above). Large model inference is expensive and unnecessary for many applications. This cost will reduce with optimization, for which model specialization and adaptive computation can help, as does the emergence of high-performing SLMs of a few billion parameters, such as Phi, trained on highly curated data.[14]

Agents will advance the search frontier to make more tasks actionable and make inroads into the "last mile" in search interaction: task completion.

Ubiquity: Integrate agents to model and support complex search tasks. AI agents must coexist with the other parts of the application ecosystem. Search agents can be integrated into applications such as Web browsers, offering in-browser chat, editing assistance, and summarization; and productivity applications, offering support in creating documents, emails, presentations, and so on. These agents can capitalize on the application context to do a better job of answering searcher requests. Agents can also span applications through integration with the operating system. This enables richer task modeling and complex task support, since such tasks often involve multiple applications. Critically, we must do this privately and securely to mitigate risks for searchers and earn their trust.

Summary. The directions highlighted in this section are just a few examples of the opportunities afforded by the emergence of generative AI and agents in search settings. There are other areas for search providers to consider too, such as multilingual agent experiences

(foundation models could help with language translation); agent efficiency (large model inference is expensive and not sustainable at massive scale, so creative solutions are needed[43]); reducing the carbon impact from running foundation models at search-engine scale; making agents private and secure by design; and government directives (for example, the 2023 executive order from U.S. President Joe Biden on AI safety and security[i]) and legislation, among many other opportunities.

The Undiscovered Country

AI agents will transform how we search. Tasks are central to people's lives, and more support is needed for complex tasks in search settings. Some limited support for these tasks already exists in search engines, but agents will advance the search frontier to make more tasks actionable and make inroads into the "last mile" in search interaction: task completion.[38] Moving forward, search providers should invest in "better together" experiences that utilize agents plus traditional search (plus more modalities going forward), make these joint experiences more seamless for searchers, and add more support for their use in practice, for example, helping people to quickly understand agent capabilities and/or recommending the best modality for the current task or task stage. This includes experiences where both modalities are offered separately and can be selected by searchers and those where there is unification and the selection happens automatically based on the task and the conversation context.

The foundation models that power AI agents have other search-related applications, such as those for generating and applying intent taxonomies and those for evaluation.[12] We must maintain a continued focus on human-AI cooperation, where searchers stay in control while the degree of system support increases as needed,[27] and on AI safety and security. Searchers need to be able to trust agents but also be able to verify their answers with minimal effort.

Overall, the future is bright for IR, and AI research in general, with the advent of generative AI and the agents that build upon it. Agents will help augment, empower, and inspire searchers on their task journeys. Computer science researchers and practitioners should embrace this new era of assistive AI agents and engage across the full spectrum of exciting practical and scientific opportunities, both within search, as we focused on in this article, and onward into other important domains such as personal productivity and scientific discovery. C

References

1. Agichtein, E., White, R.W., Dumais, S.T., and Bennet, P.N. Search, interrupted: Understanding and predicting search task continuation. In *Proceedings of the 35th Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2012, 315–324.
2. Bates, M.J. Where should the person stop and the information search interface start? *Information Processing & Management 26*, 5 (1990), 575–591.
3. Belkin, N.J. Anomalous states of knowledge as a basis for information retrieval. *Canadian J. of Information Science 5*, 1 (1980), 133–143.
4. Bennett, P.N. et al. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2012, 185–194.
5. Bourtoule, L. et al. Machine unlearning. In *IEEE Symp. on Security and Privacy*. IEEE, 2021), 141–159.
6. Broder, A.Z. and McAfee, P. Delphic costs and benefits in web search: A utilitarian and historical analysis. *arXiv preprint arXiv:2308.07525* (2023).
7. Bubeck, S. et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
8. Byström, K. and Järvelin, K. Task complexity affects information seeking and use. *Information Processing & Management 31*, 2 (1995), 191–213.
9. Chiang, W.-L. et al. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *LMSYS Org.* (Mar. 30, 2023); https://vicuna.lmsys.org
10. Creswell, A. and Shanahan, M. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271* (2022).
11. Dervin, B. Sense-making theory and practice: An overview of user interests in knowledge seeking and use. *J. of Knowledge Mgmt. 2*, 2 (1998), 36–46.
12. Faggioli, G. et al. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR Intern. Conf. on Theory of Information Retrieval*, 2023, 39–50.
13. Gao, J., Xiong, C., Bennett, P., and Craswell, N. *Neural Approaches to Conversational Information Retrieval. INRE Vol. 44*. Springer Nature, 2023.
14. Gunasekar, S. et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644* (2023).
15. Awadallah, A.H. et al. Supporting complex search tasks. In *Proceedings of the 23rd ACM Intern. Conf. on Information and Knowledge Management*, 2014, 829–838.
16. Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*, 2002, 133–142.
17. Krathwohl, D.R. A revision of Bloom's taxonomy: An overview. *Theory into Practice 41*, 4 (2002), 212–218.
18. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems 33*, (2020), 9459–9474.
19. Li, Y. and Riva, O. Glider: A reinforcement learning approach to extract UI scripts from websites. In *Proceedings of the 44th Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2021, 1420–1430.
20. Liang, P.P., Wu, C., Morency, L.-P., and Salakhutdinov, R. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th Intern. Conf. on Machine Learning. PMLR 139* (2021), 6565–6576.
21. Marchionini, G. Exploratory search: from finding to understanding. *Commun. ACM 49*, 4 (2006), 41–46.
22. Mukherjee, S. et al. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv preprint arXiv:2306.02707* (2023).
23. Najork, N. *Generative information retrieval.* In *Proceedings of the 46th Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval.* 2023, 1.
24. Olteanu, A. et al. FACTS-IR: Fairness, accountability, confidentiality, transparency, and safety in information retrieval. *ACM SIGIR Forum 53*, 1 (2021), 20–43.
25. Schick, T. et al. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761* (2023).
26. Shah, C. et al. Taking search to task. In *Proceedings of the 2023 Conf. on Human Information Interaction and Retrieval*. 2023, 1–13.
27. Shneiderman, B. *Human-Centered AI*. Oxford University Press, 2022.
28. Singla, A., White, R., and Huang, J. Studying trailfinding algorithms for enhanced web search. In *Proceedings of the 33rd Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval*. 2010, 443–450.
29. Teevan, J. et al. Slow search. *Commun. ACM 57*, 8 (2014), 36–38.
30. Teevan, J., Dumais, S.T., and Horvitz, E. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval*. 2005, 449–456.
31. Teevan, J., Morris, M.R., and Bush, S. Discovering and using groups to improve personalized search. In *Proceedings of the Second ACM Intern. Conf. on Web Search and Data Mining*. 2009, 15–24.
32. ter Hoeve M. et al. Conversations with documents: An exploration of document-centered assistance. In *Proceedings of the 2020 Conf. on Human Information Interaction and Retrieval*. 2020, 43–52.
33. Thomas, P. et al. Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621* (2023).
34. Tyler, S.K. and Teevan, J. Large scale query log analysis of re-finding. In *Proceedings of the Third ACM Intern. Conf. on Web Search and Data Mining*. 2010, 191–200.
35. Vakkari, P. Searching as learning: A systematization based on literature. *J. of Information Science 42*, 1 (2016), 7–18.
36. Vincent, N. The paradox of reuse, language models edition. *nmvg*. (Dec. 2, 2022); https://nmvg.mataroa.blog/blog/the-paradox-of-reuse-language-models-edition/.
37. Wang, Y., Huang, X., and White, R.W. Characterizing and supporting cross-device search tasks. In *Proceedings of the Sixth ACM Intern. Conf. on Web Search and Data Mining*. 2013, 707–716.
38. White, R.W. *Interactions with Search Systems*. Cambridge University Press, 2016.
39. White, R.W. Skill discovery in virtual assistants. *Commun. ACM 61*, 11 (2018), 106–113.
40. White, R.W. et al. Multi-device digital assistance. *Commun. ACM 62*, 10 (2019), 28–31.
41. Wu, Q. et al. AutoGen: Enabling next- gen LLM applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155* (2023).
42. Zamani, H. et al. Generating clarifying questions for information retrieval. In *Proceedings of the Web Conf. 2020*. 2020, 418–428.
43. Zhang, J. et al. EcoAssistant: Using LLM assistant more affordably and accurately. *arXiv preprint arXiv:2310.03046* (2023).
44. Zhang, Y. et al. Learning to decompose and organize complex tasks. In *Proceedings of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, 2726–2735.
45. Ziegler, D.M. et al. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

i https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

Ryen W. White is a research scientist, general manager, and deputy lab director at Microsoft Research in Redmond, WA, USA.