Gartner®

Quick Answer: What Impact Will Generative AI Have on Search?

Published 14 June 2023 - ID G00788353 - 7 min read

Hao Yin

Initiatives: Artificial Intelligence

ChatGPT has catapulted generative AI from obscurity to the everyday. Data and analytics leaders must prepare for the transformation of search experiences as vendors reengineer their products with technology that enables semantic search at scale.

Quick Answer

What impact will generative AI have on search?

- **As an experience,** search will recede behind the user interface for the majority of citizens, customers and employees. Their need for insight — the understanding required to know, decide and act — will be increasingly met through proactive synthesis of information delivered through conversational interfaces and visual navigation: facts in context, delivered automatically.

- **As a technology,** search will augment generative AI to power the synthesis of information to humans, or data to machines. This represents a sharp and momentous shift from the pre-ChatGPT era in which AI augmented search. Generative AI, in isolation, is neither an alternative to nor replacement for current search technologies.

More Detail

Search is a wide and complex topic representing many constituencies and their interests. Broadly speaking, it can be approached as both an experience and as a technology.

Search as an Experience

When we talk about search, it is important to recognize that search experiences vary across many constituents and their use cases, which are served by numerous vendors, many of whom have offerings across several use cases:

- **Public search** — Publicly accessible sources served by vendors such as Baidu, Microsoft, Google and Yahoo, among others.

- **Website search** — Search across a single or group of websites by vendors such as Cludo, Elastic, SearchBlox and Squiz, among others.

- **Digital commerce search and product discovery** — Augmentation of digital commerce solutions to serve product discovery by Algolia, Bloomreach, Coveo, Lucidworks and Yext, among others.

- **Digital workplace search** — Enterprisewide, intranet and localized search to serve workers within a digital workplace by vendors such as Ayfie, Glean Technologies, IntraFind, Microsoft and S&P Global, among others.

- **Application search** — Via third-party or custom-made applications by vendors such as EPAM, Expert.ai, IBM, Mindbreeze, OpenText, Sinequa and Squirro, among others.

- **Device search** — Localized to device and networked storage by vendors such as Axonic, Copernic, dtSearch and X1, among others.

Search delivers information to provide the insight needed to know, decide and act. This typically entails the retrieval of what exists, such as a webpage or document, but increasingly also the synthesis of something new, such as an information card or composed answer. Both retrieval and synthesis can be delivered reactively, interactively or proactively.

> **Generative AI is shifting both capability and expectation from retrieval to synthesis of information; reactive to proactive delivery.**

Combining generative AI with search demonstrates various forms of information synthesis:

- Answers to questions, delivering Q&A

- Context-sensitive responses, delivering dialogue

- Transient information assets, delivering summaries and articles

- Next best actions, delivering guidance and automation

Moreover, such responses can be delivered proactively:

- Relevant questions to users, to initiate or steer conversation
- Recommendations, to match user context
- Automation, to act on next best actions

Given such advances, the experience of search can change from the traditional search bar plus results to become an integral part of conversational and navigational user interfaces. For many users, this would be nothing short of the death of the traditional keyword-based query-response experience of search.

However, the novelty of generative AI's ability to proactively synthesize information is tempered by the need for accuracy, verification, low latency and explanation, as well as relevance, especially in domains where decision and action are constrained by standards, regulations and laws (e.g., healthcare, government, pharmaceutical, finance). Consequently, the need for retrieval will persist to steer and ground responses, as well as serve as a modality in its own right for certain use cases (e.g., forensic search).

Search results follow well-established UI patterns such as document lists or product grids, facets, and filters. What is likely to appear are new, hybrid UIs in which the conversational dialogue can be followed, but that also deliver rich visual results and navigation (especially in the context of product discovery). Some vendors have already experimented with hybrid conversational UIs, but have found that enterprises are very cautious when considering a new UI. Indeed, one impact of ChatGPT's initial hype may be that there is now an *expectation* that search user journeys will become conversational, and thus be open to such new UI paradigms.

Search as a Technology

Generative Pre-trained Transformer (GPT), the model underpinning the ChatGPT application, is one of a class of AI foundational models. These transformer-architecture-based large language models (LLMs) enable semantic search that performs beyond the limits reached with rule-based approaches. Furthermore, foundational models can be combined or substituted with other classes of model, such as convolutional models for vision, to enable multimodal semantic search. But while experiences with foundational models are astonishing, neither they nor the GPT models deliver scalable semantic search in isolation. In their current form, none are an alternative to nor replacement for current search technologies.

Foundational models can be used in conjunction with search technologies to power semantic, conversational search.

Semantic search analyzes the relationships between words as well as words themselves during indexing and query to deliver more relevant search experiences. This requires either:

- A rule-based approach (e.g., knowledge graphs, taxonomies), or
- A math-based approach, such as vector search (e.g., Word2Vec, BERT), or
- A combination of both approaches

These approaches use rule-sets and models that can be tailored to localized corpora, and therefore — within reason — scaled to the enterprise. But only just. For many enterprises, these approaches are a step too far. Both enable semantic search, but their model size limits the benefits, requiring careful trade-offs.

GPT, on the other hand, is many orders of magnitude larger, delivering better performance. But it is not trained on the corpora of enterprise content and data across which search is needed. In addition, its recency matches that of the content and data used to train it: no content nor data created after the last cycle of training is included in the model. For both GPT-3.5 (the model used for the public version of ChatGPT) and GPT-4 (the latest model), the cut-off for training data was September 2021, despite different models and release dates. Furthermore, factual inaccuracies, biases and other artifacts in that training data are now embedded in the model leading to the potential for inaccurate and inappropriate responses.

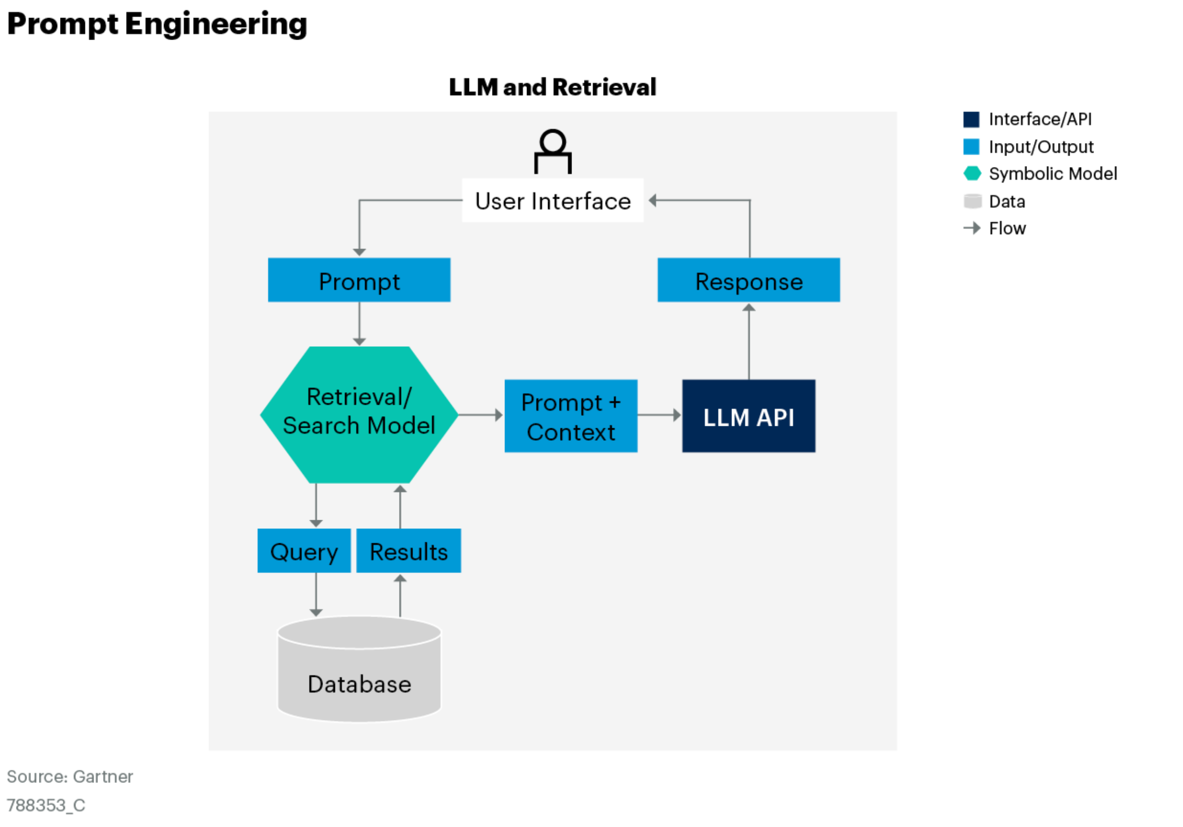
Use a Composite AI-Based Approach

To overcome the impasse, a composite AI-based approach is necessary. Composite AI combines explicit rule-based with implicit math-based approaches.

A technique known as prompt engineering is used to combine localized semantic search with LLMs such as GPT (see Figure 1). Engineering the prompt involves expanding the prompt submitted to the LLM with relevant content and data drawn from an initial, localized semantic search. This is known as “grounding” — that is, grounding Generative AI workflows with context and recency.

The prompt is then submitted to the LLM to generate a response that synthesizes relevant and well-formed information for people, or data for machines. Moreover, such responses offer some level of verification and explanation through the inclusion of references to the sources used to engineer the prompt (see Figure 1). [1]

Figure 1: Prompt Engineering



Implicitly, there is a fundamental shift at play here: Search augments AI, rather than the reverse, to provide more relevant search experiences. And it's not just with respect to the synthesis of information and direct experiences of search. Generative AI serves a range of search-related tasks, including:

- **Summarization** — Generating summaries to be returned in response to searches
- **Classification** — Capturing and attributing metadata to digital assets and their parts
- **Query parsing** — Reformulating user queries for processing against the index
- **Response quality** — Checking the relevance of responses against the initial query

As well as facilitating search, such capabilities underpin effective content, information and knowledge management. The capabilities rely on the interplay of search and generative AI. Search also facilitates the use of generative AI for other use cases in addition to search. For instance, security trimming ensures that only content a user is authorized to access is used to build context as the prompt is grounded.

Aside from the engineering challenges, other obstacles include the need to share classified content and data returned by search with third parties, such as OpenAI in the case of ChatGPT, and the governance issues this raises. Microsoft and OpenAI offer both ChatGPT and GPT-4 as cloud APIs within Azure that can in part address these governance issues, albeit at a cost. Moreover, foundational models can be fine-tuned with organizational data to create custom models, which have better accuracy and potentially less falsehood — wrongly referred to as "hallucination." However, the feasibility of fine-tuning decreases as model size increases.

Recommended by the Authors

Innovation Insight for Artificial Intelligence Foundation Models

ChatGPT Research Highlights

Magic Quadrant for Insight Engines

AI Design Patterns for Large Language Models

Evidence

[1] Evaluating Verifiability in Generative Search Engines, Arxiv, Cornell University