

LLMS — MICROSOFT — NEWS

'Enormous business potential': Microsoft on why GraphRAG outperforms naive RAG

Redmond opens up to discuss its new tool, which can extract data from unstructured text using large language models.

JASPER HAMILL

July 16, 2024 · 3:14 PM — 4 min read



GraphRAG is a new way of doing Retrieval-Augmented Generation (RAG) (Image: Microsoft)

Microsoft's GraphRAG is a new approach to Retrieval-Augmented Generation (RAG) that Redmond has described as a "[significant advance in enhancing the capability of LLMs.](#)"

The tool is designed to perform response generation and structured information retrieval more effectively than "naive RAG" - an early approach to Generative AI indexing, retrieval, and generation.

At the beginning of July, GraphRAG made it onto [Github](#), where it's described as a "data pipeline and transformation suite that is designed to extract meaningful, structured data from unstructured text using the power of large language models."

We spoke to Microsoft to understand more about GraphRAG and how it improves upon existing RAG methodologies.

Jonathan Larson, Senior Principal Data Architect at Microsoft Research, told *The Stack* that GraphRAG has been "seeing a lot of interest" since making its Github debut - especially among "startups trying to tackle difficult research problems in spaces from cancer treatment to chemistry."

"Naïve RAG is great for queries where an embedding nearest neighbour search will help you arrive at a result quickly," Larson explained. "In other words, naïve RAG is better at finding specific phrases rather than more abstract ideas and concepts. It is difficult for naïve RAG to retrieve all relevant parts of abstract ideas and concepts. It has no understanding of the dataset as a whole and can't reason holistically over it."



Question: Which public figures are repeatedly mentioned across various entertainment articles?

Naive RAG response

Public figures who are repeatedly mentioned across various entertainment articles include <Public figures 1, 2, 3, and 4>. These individuals have been highlighted for various reasons, ranging from their professional achievements to their personal lives.

<Public figure 1> [...]
<Public figure 2> [...]
<Public figure 3> [...]
<Public figure 4> [...]

These figures are frequently covered due to their high-profile status and the public's interest in their careers and personal lives. Their activities, whether in music, sports, or personal relationships, have significant cultural and economic impacts, as evidenced by the media coverage and public reactions.

LLM evaluation

Comprehensiveness: Winner = GraphRAG
Diversity: Winner = GraphRAG
Empowerment: Winner = GraphRAG

GraphRAG response

The entertainment industry is vast and diverse, encompassing film, television, music, sports, and digital media. Certain public figures stand out due to their significant contributions and influence across these sectors. The following summary highlights key individuals who are repeatedly mentioned in various entertainment articles, reflecting their impact and presence within the industry.

Actors and Directors [... 20 figures ...]
Public Figures in Controversy [... 3 figures ...]
Musicians and Executives [... 5 figures ...]
Athletes and Coaches [... 7 figures ...]
Influencers and Entrepreneurs [... 3 figures ...]

The repeated mention of these figures in entertainment articles signifies their ongoing relevance and the public's interest in their work. Their influence spans across various aspects of entertainment, from shaping cultural narratives in film and television to driving trends in music and digital media. These individuals not only contribute to their respective fields but also influence the broader cultural landscape, often becoming central figures in social discussions and public discourse.

A comparison between GraphRAG and naive RAG

One question that traditional naive RAG approach can answer is a query such as: 'How many models of Product XYZ are we currently selling to Customer ZYX?'

However, naive models do not work so well with deeper questions such as: "Tell me about all of my customers and give me a summary of the status for each."

"Naive RAG will fall short on this type of question as it doesn't have the ability to holistically analyze the dataset," Larson continued.

Moving on from naive RAG

GraphRAG enters the fray by improving on naive RAG approaches based on vector search—a method of information retrieval in which queries and documents are mathematically represented as vectors instead of plain text.

GraphRAG uses an LLM to automate the extraction of a "rich knowledge graph" from a collection of text documents. It reports on the semantic structure of the data before answering user queries by detecting "communities" of nodes and then creating a hierarchical summary of the data to provide an overview of a dataset, with each community able to summarise its entities and their relationships.

Larson said: "GraphRAG enables a variety of new scenarios that naive RAG fails to address. We see enormous potential for business productivity as GraphRAG takes us beyond the limitations of naive RAG, allowing us to reason holistically and to get past the limitations of vector search.

"For example, suppose I look at a tranche of enterprise project and design documents and ask the question: 'What are the major projects that are being worked on? Give me details of each project and a listing of everyone mentioned to be working on it.'

Read more: [Meta's new "CRAG" benchmark exposes depth of RAG challenge](#)

"Using naive RAG, this type of question fails as it lacks specificity to locate relevant results in its initial retrieval. With an LLM memory representation, GraphRAG has a holistic understanding of the dataset and uses that to answer the question with high fidelity."

In contrast to naive approaches, GraphRAG builds a memory representation of the dataset which allows it to "clearly see and reason over its contents and their relationships", Larson went on. "This allows you to ask questions like 'which are the most popular products across all of our customers' for which naïve RAG would struggle," he said.

Microsoft's own research found that GraphRAG "outperforms" RAG on comprehensiveness and diversity when using community summaries "at any level of the community hierarchy", with a win rate of between 70% and 80%.

What is the industry saying about GraphRAG?

To get an outside perspective on GraphRAG, we spoke to Jim Webber, Chief Scientist at database and analytics firm Neo4j, who told us that "simple RAG based on vector search has lower accuracy since it merely matches chunks of text based on geometric closeness which is similarity only at a superficial level", hailing Microsoft's GraphRAG as "a welcome addition to the generative AI world."

"Conversely, knowledge graphs are a rich network of facts that increase retrieval accuracy and by further curating graphs with algorithms (hierarchical clustering in this case), accuracy can be further improved by implicitly enriching local retrieval with global knowledge through the topology," he said.

"The industry has known about these possibilities for some time, but with the might of Microsoft Research backing up these findings, it will open a path for many users to follow."

Dom Couldwell, Head of Field Engineering EMEA at DataStax, also explained: "RAG is about leveraging your own data in generative AI systems. When you have a batch of data from your enterprise, you want to use that data to improve the responses that you deliver to customers or users when it is useful and relevant. RAG also helps you to get over the issues of training data cut-offs, so you can get real-time data into those responses.

"One challenge around this is where you have a lot of files in your data that have very similar information. How do you help your RAG system find that data when the search is looking at files with very similar semantic information?"

"GraphRAG helps you overcome that problem - you can create a knowledge graph that connects up your data in a different way to make it easier to deliver that needed information into your generative AI system. By combining different data management techniques, you can overcome that potential performance issue."

[READ MORE: Microsoft unveils a large language model that excels at encoding spreadsheets](#)

Related





LLMS

AI-generated data causes LLM model collapse: Researchers

JASPER HAMILL
July 26, 2024



LLMS

Microsoft unveils a large language model that excels at encoding spreadsheets

JASPER HAMILL
July 15, 2024



LLMS

OpenAI sends frontier model LLMs to fight bioterrorism at Los Alamos Lab

JASPER HAMILL
July 11, 2024



LLMS

Nvidia intern's team beats OpenAI's GPT-4 LLM in 'knowledge-intensive' showdowns

JASPER HAMILL
July 9, 2024

Latest



GITHUB

'Rotate your keys now': Sensitive data could be accessible in deleted or private Github repositories

JASPER HAMILL
July 26, 2024



LLMS

AI-generated data causes LLM model collapse: Researchers

JASPER HAMILL
July 26, 2024



CISO

"We're becoming scapegoats": How have CISOs responded to SEC cyber risk disclosure rules?

JASPER HAMILL
July 26, 2024



FORD

Ford's CEO eyes \$1 billion in software sales

THE STACK
July 25, 2024

THE STACK



[Sign up](#)

© 2024 The Stack – Published with [Ghost](#) & [Tripoli](#)