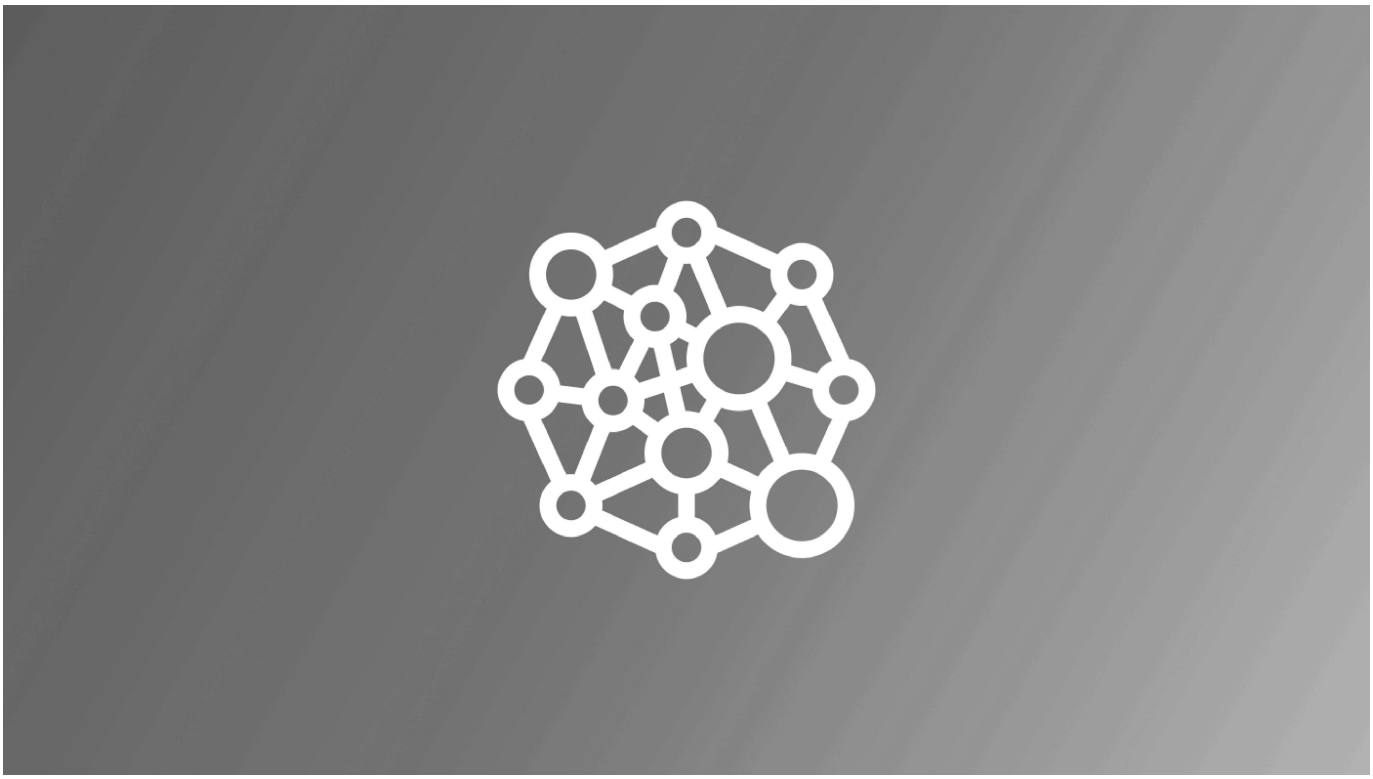
GraphRAG: New tool for complex data discovery now on GitHub

Published July 2, 2024

By Darren Edge, Senior Director; Ha Trinh, Senior Data Scientist; Steven Truitt, Principal Program Manager; Jonathan Larson, Senior Principal Data Architect



DOWNLOAD
GraphRAG

DOWNLOAD
GraphRAG Accelerator

Earlier this year, we introduced GraphRAG, a graph-based approach to retrieval-augmented generation (RAG) that enables question-answering over private or previously unseen datasets. Today, we're pleased to announce that GraphRAG is now available on GitHub, offering more structured information retrieval and comprehensive response generation than naive RAG approaches. The GraphRAG code repository is complemented by a solution accelerator, providing an easy-to-use API experience hosted on Azure that can be deployed code-free in a few clicks.

GraphRAG uses a large language model (LLM) to automate the extraction of a rich knowledge graph from any collection of text documents. One of the most exciting features of this graph-based data index is its ability to report on the semantic structure of the data prior to any user queries. It does this by detecting "communities" of densely connected nodes in a hierarchical fashion, partitioning the graph at multiple levels from high-level themes to low-level topics, as illustrated in Figure 1. Using an LLM to summarize each of these communities creates a hierarchical summary of the data, providing an overview of a dataset without needing to know which questions to ask in advance. Each community serves as the basis of a *community summary* that describes its entities and their relationships.

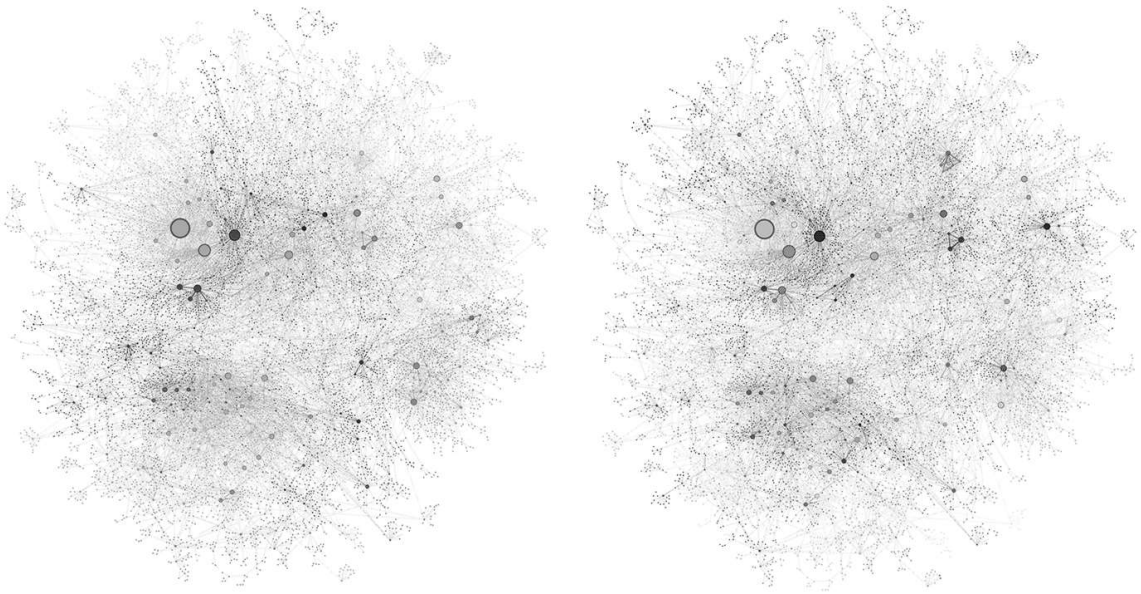


Figure 1. Knowledge graph of entity nodes and relationship edges derived from a news dataset, with different colors representing various communities. Level 0 communities (left) represent the highest-level themes of the dataset, while level 1 communities (right) show the emergence of more granular topics within these themes.

Advantages of community summaries for "global questions"

In a recent preprint, we explore how these community summaries can also help answer *global questions*—which address the entire dataset rather than focusing on specific chunks of text—where naive RAG approaches based on vector search fall short. For example, consider the question "What are the main themes in the dataset?" This is a reasonable starting point but one where naive RAG will always give misleading answers. This is because it generates answers from chunks of text semantically similar to the question, not necessarily from the subset of input texts needed to answer it.

However, if a question addresses the entire dataset, *all* input texts should be considered. Since naive RAG only considers the top-*k* most similar chunks of input text, it fails. Even worse, it will match the question against chunks of text that are superficially similar to that question, resulting in misleading answers. Community summaries help answer such global questions because the graph index of entity and relationship descriptions has already considered all input texts in its construction. Therefore, we can use a map-reduce approach for question answering that retains all relevant content from the global data context:

1. Group community reports up to the LLM context window size.
2. Map the question across each group to create community answers.
3. Reduce all relevant community answers into a final global answer.

Evaluation and results

To evaluate this approach against naive RAG and hierarchical source-text summarization, we used the LLM GPT-4 to generate a diverse set of activity-centered sense-making questions from short descriptions of two datasets: podcast transcripts and news articles. We then selected three metrics for head-to-head comparison of generated answers, as evaluated by an LLM judge: comprehensiveness (covers all aspects in detail), diversity (provides different perspectives), and empowerment (supports informed decision making).

The results show that GraphRAG, when using community summaries at any level of the community hierarchy, outperforms naive RAG on comprehensiveness and diversity (~70–80% win rate). GraphRAG using intermediate- and low-level community summaries also performed better than source text summarization on these metrics at lower token costs (~20–70% token use per query). Performance was competitive with hierarchical source text summarization for the highest-level communities at substantially lower token costs (~2–3% token use per query). This is shown in Figure 2.

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

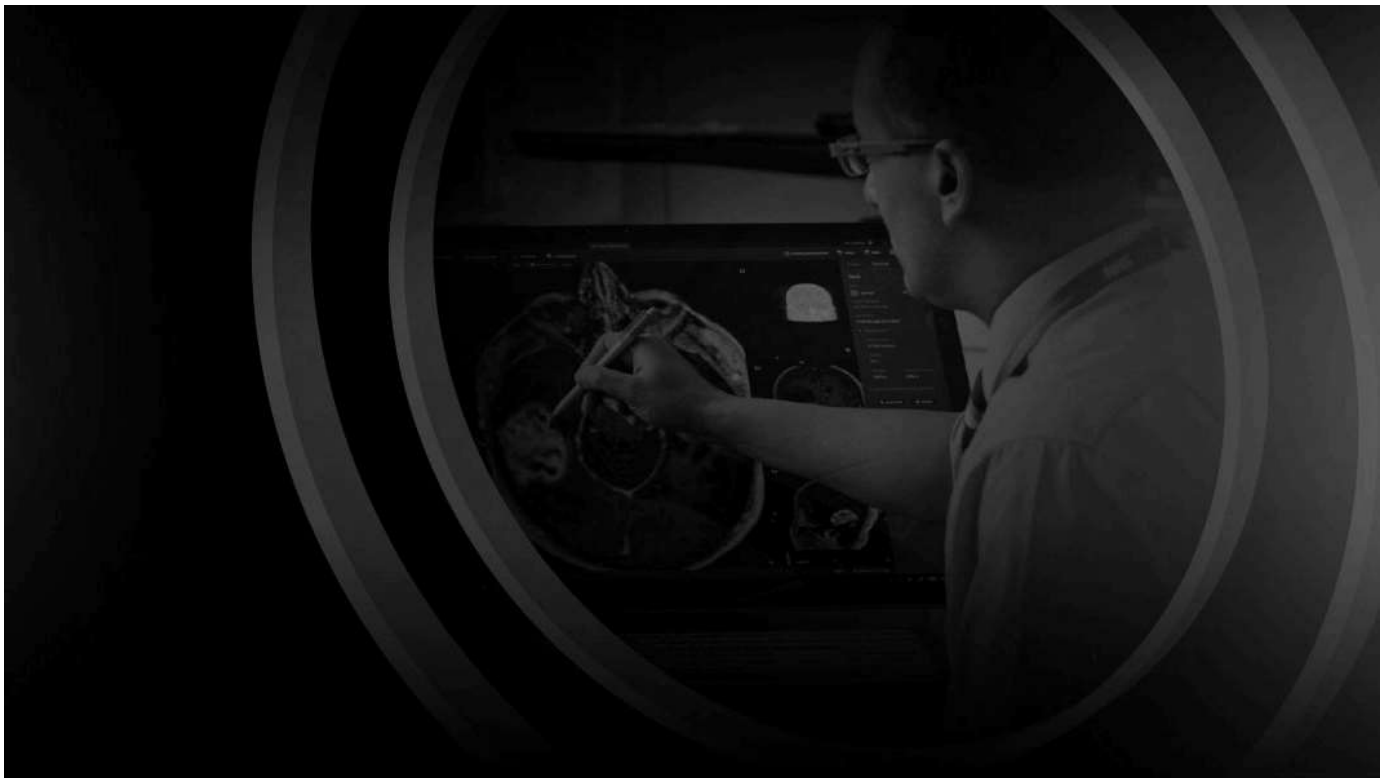
GraphRAG response

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.

Figure 2. Comparison of naive RAG and GraphRAG responses to a global question about a news dataset indicates that GraphRAG outperformed naïve RAG in terms of comprehensiveness, diversity, and empowerment.



Research insights and future directions

Through the initial research cycle, we demonstrated that LLMs can successfully derive rich knowledge graphs from unstructured text inputs, and these graphs can support a new class of global queries for which (a) naive RAG cannot generate appropriate responses, and (b) hierarchical source text summarization is prohibitively expensive per query. The overall suitability of GraphRAG for any given use case, however, depends on whether the benefits of structured knowledge representations, readymade community summaries, and support for global queries outweigh the upfront costs of graph index construction.

We're currently exploring various approaches to reduce these costs while maintaining response quality. Our latest work on automatically tuning LLM extraction prompts to the problem domain is an example of how we are reducing the upfront effort required to customize these prompts, enumerate entity types, create few-shot examples, and so on. To enable evaluation of GraphRAG with minimal upfront indexing costs, we're also investigating NLP-based approaches to approximating the knowledge graph and community summaries that would be generated by a full indexing process. Our goal is to ensure that, whatever the constraints of the deployment context, there is a GraphRAG configuration that can accommodate these constraints while still delivering exceptional response quality.

DOWNLOAD

GraphRAG Accelerator

By making GraphRAG and a solution accelerator publicly available, we aim to make graph-based RAG approaches more accessible for users and use cases where it's critical to understand data at a global level. We encourage community feedback and suggestions on both the code repository and solution accelerator as we work together to enable the next generation of RAG experiences.

Acknowledgements

Joshua Bradley, Christine Caggiano, Mónica Carvajal, Alex Chao, Newman Cheng, Ed Clark, Ben Cutler, Andres Morales Esquivel, Nathan Evans, Alonso Guevara Fernández, Amber Hoak, Kate Lytvynets, Gaudy Blanco Meneses, Apurva Mody, Robert Ness, Gabriel Nieves-Ponce, Douglas Orbaker, Richard Ortega, Rodrigo Racanicci, Billie Rinaldi, Katy Smith, Sarah Smith, Shane Solomon, Dayenne Souza, David Tittsworth, Chris Trevino, Derek Worthen

Related publications

From Local to Global: A Graph RAG Approach to Query-Focused Summarization >

Meet the authors