Perhaps the greatest challenge – and opportunity – of LLMs is extending their powerful capabilities to solve problems beyond the data on which they have been trained, and to achieve comparable results with data the LLM has never seen. This opens new possibilities in data investigation, such as identifying themes and semantic concepts with context and grounding on datasets. In this post, we introduce **GraphRAG**, created by Microsoft Research, as a significant advance in enhancing the capability of LLMs.

Retrieval-Augmented Generation (RAG) is a technique to search for information based on a user query and provide the results as reference for an AI answer to be generated. This technique is an important part of most LLM-based tools and the majority of RAG approaches use vector similarity as the search technique. GraphRAG uses **LLM-generated knowledge graphs** to provide substantial improvements in question-and-answer performance when conducting document analysis of complex information. This builds upon our recent research, which points to the power of **prompt augmentation** when performing discovery on *private datasets*. Here, we define *private dataset* as data that the LLM is not trained on and has never seen before, such as an enterprise's proprietary research, business documents, or communications. *Baseline RAG*[1] was created to help solve this problem, but we observe situations where baseline RAG performs very poorly. For example:

- Baseline RAG struggles to connect the dots. This happens when answering a question requires traversing disparate pieces of information through their shared attributes in order to provide new synthesized insights.
- Baseline RAG performs poorly when being asked to holistically understand summarized semantic concepts over large data collections or even singular large documents.

To address this, the tech community is working to develop methods that extend and enhance RAG (e.g., LlamaIndex). Microsoft Research's new approach, GraphRAG, uses the LLM to create a knowledge graph based on the private dataset. This graph is then used alongside graph machine learning to **perform prompt augmentation at query time**. GraphRAG shows substantial improvement in answering the two classes of questions described above, demonstrating intelligence or mastery that outperforms other approaches previously applied to private datasets.

Applying RAG to private datasets

To demonstrate the effectiveness of GraphRAG, let's start with an investigation using the Violent Incident Information from News Articles (VIINA) dataset. This dataset was chosen due to its complexity and the presence of differing opinions and partial information. It is a messy real-world test case that was recent enough not to be included in the LLM base model's training.

For this research, we use thousands of news articles from both Russian and Ukrainian news sources for the month of June 2023, translated into English, to create a private dataset on which we will perform our LLM-based retrieval. The dataset is far too large to fit into an LLM context window, thus demanding a RAG approach.

Query: *"What is Novorossiya?"*

Baseline RAG	GraphRAG
The term "Novorossiya" is not mentioned in the provided context. However, to provide some information, Novorossiya, or New Russia, is a historical term referring to a region in the south of Ukraine that was conquered by the Russian Empire in the 18th century. In recent years, the term has been revived by Russian-backed separatists in Eastern Ukraine to refer to the territories they claim, but it is not officially recognized as a separate entity by international bodies.	Novorossiya, also known as New Russia, is a term with historical and contemporary significance. Historically, it referred to a region in the southern part of the Russian Empire that is now part of Ukraine. In the contemporary context, it has been used by pro-Russian separatists to refer to certain areas of Ukraine [Data: Entities (6494, 912, 6493)]. The term Novorossiya has been used in the context of a political movement in Ukraine, particularly by Russian nationalists. This movement has been associated with efforts to destabilize Ukraine and create a so-called "Novorossiya" [Data: Entities (912); Relationships (15211, 15206)]. The term has also been used to denote a separatist state in Ukraine, known as the Federal Republic of Novorossiya (FRN) [Data: Entities (7690)].

In these results, we can see both systems perform well – highlighting a class of query on which baseline RAG performs well. Let's try a query that requires connecting the dots:

Query: *"What has Novorossiya done?"*

Baseline RAG	GraphRAG
The text does not provide specific information on what Novorossiya has done.	Novorossiya, a political movement in Ukraine, has been involved in a series of destructive activities, particularly targeting various entities in Ukraine [Entities (6494, 912)]. The movement has been linked to plans to destroy properties of several Ukrainian entities, including Rosen, the Odessa Canning Factory, the Odessa Regional Radio Television Transmission Center, and the National Television Company of Ukraine [Relationships (15207, 15208, 15209, 15210)]. In addition to these, Novorossiya has also been associated with plans to destroy properties of Roshen, the Odessa Conservatory Factory, the Odesa Regional Radio Television Transmission Center, and the National Telecommunications Company of Ukraine [Relationships (8329, 8330, 8331, 8332)]. Furthermore, Novorossiya has been implicated in plans to blow up ATMs of PrivatBank, the largest state bank in Ukraine [Relationships (8335, 8335)]. The Office of the General Prosecutor in Ukraine has reported on the creation of Novorossiya, indicating the government's awareness and potential concern over the activities of this movement [Relationships (8264, 8264)]. The destructive activities of Novorossiya have had significant impacts on the targeted entities and have likely contributed to the ongoing tensions and conflicts in the region. In summary, Novorossiya has been involved in a series of destructive activities targeting various entities in Ukraine. These activities have had significant impacts on the targeted entities and have likely contributed to the ongoing tensions and conflicts in the region. The Ukrainian government, through the Office of the General Prosecutor, has

	acknowledged the existence and activities of Novorossiya, indicating a level of concern over the movement's actions.

Baseline RAG fails to answer this question. Looking at the source documents inserted into the context window (Figure 1), none of the text segments discuss Novorossiya, resulting in this failure.

Relevant chunks of source documents:

	text	source
0	The substance did not go beyond the enterprise.*The activities of Meta (social networ	ria
1	The problems that the West is causing her. The sanctions list, published on the depar	ria
2	They ignore Kiev's constant refusals to negotiate in the West. Earlier, the official repre	ria
3	They plan to export grain to Russia. Russian military are robbing Ukrainians - what is	unian
4	Countries have begun to fight not only against Soviet history, but also against everyth	ria
5	Energy, defense industry, military administration, and communications in Ukraine. A:	ria
6	About anti-Ukrainian content on the air of the Russian channel "Soloviev.live" and fu	unian
7	and dentistry, resuscitation departments and operating rooms. The hospital grounds	ria
8	"of the kind that has been going on for over a year," Medvedev stated. It should be no	unian
9	They are deploying their weapons directly in the cities of people on the edge of the se	unian

Figure 1: Baseline RAG retrieved context [2]

In comparison, the GraphRAG approach **discovered an entity in the query**, Novorossiya. This allows the LLM to ground itself in the graph and results in a superior answer that contains provenance through links to the original supporting text. For example, Figure 2 below shows the exact content the LLM used for the LLM-generated statement, "Novorossiya has been implicated in plans to blow up ATMs." We see the snippet from the raw source documents (after English translation) that the LLM used to support the assertion that a specific bank was a target for Novorossiya via the relationship that exists between the two entities in the graph.

NOVOROSIYA - PRIVATBANK

Description

Text

source	date	text
interfaxua	2023-06-08	and the creation of the so-called "Novorossiya". In April-June 2014, he recruited two residents of Odessa to commit acts of terrorism in the city," the Prosecutor General's Office reported on its Telegram channel. In particular, the criminal blew up buildings of military commissariats, ATMs and branches of "PrivatBank", property of the "Roshen" company, PJSC "Odessa Canning Factory", SE "Odessa Regional Radio and Television Broadcasting Center" and the Odessa branch of the National Television Company of Ukraine. "Under the curator's instructions, the perpetrators carried out photo and video recording of administrative buildings and military commissariats, searched for explosives and devices for explosions and arson. To prepare for sabotage and terrorist acts in Odessa, the perpetrators received a cash reward, as well as laptops and flash drives with software for creating cryptographically encrypted connections," the prosecutor's office notes. Thanks to operational measures, law enforcement officers prevented the criminal activity of the group. The court sentenced the organizer to a punishment of 10 years imprisonment. He is in

Figure 2: GraphRAG provenance

By using the LLM-generated knowledge graph, GraphRAG vastly improves the "retrieval" portion of RAG, populating the context window with higher relevance content, resulting in better answers and capturing evidence provenance.

GraphRAG provides the provenance, or source grounding information, as it generates each response. It demonstrates that an answer is grounded in the dataset. Having the cited source for each assertion readily available also enables a human user to quickly and accurately audit the LLM's output directly against the original source material.

However, this isn't all that's possible using GraphRAG.

Whole dataset reasoning

Baseline RAG struggles with queries that require aggregation of information across the dataset to compose an answer. Queries such as "**What are the top 5 themes in the data?**" perform terribly because baseline RAG relies on a vector search of semantically similar text content within the dataset. There is nothing in the query to direct it to the correct information.

However, with GraphRAG we *can* answer such questions, because the structure of the LLM-generated knowledge graph tells us about the structure (and thus themes) of the dataset as a whole. This allows the private dataset to be organized into meaningful semantic clusters that are pre-summarized. The LLM uses these clusters to summarize these themes when responding to a user query.

We illustrate whole-dataset reasoning abilities by posing the following question to the two systems:

Query: "*What are the top 5 themes in the data?*"

Baseline RAG	**GraphRAG**

1. Improving the quality of life in cities and urban development. 2. The meeting between Vladimir Putin and the President of the United Arab Emirates, Mohammed bin Zayed Al Nahyan. 3. The National Rating of the Investment Climate in the Regions of Russia. 4. The state of the Russian economy and the growth of domestic brands. 5. The discussion on the meaning of life and the role of science in understanding it.	The data presents a wide range of themes, but the top five most prevalent themes can be identified as follows: 1. **Conflict and Military Activity**: A significant portion of the data revolves around the ongoing conflict in Ukraine, with numerous reports detailing military activities, occupations, and attacks by various entities. This theme includes the actions of the Ukrainian military, Russian forces, and various other military units and entities [Data: Reports (513, 241, 151, 493, 343, +more)]. 2. **Political and Governmental Entities**: Many reports focus on political figures and governmental entities, highlighting their roles, relationships, and actions in the context of the current situation. This includes entities like the Kremlin, the President of Ukraine, the US State Department, and various regional administrations [Data: Reports (98, 493, 265, 129, 196, +more)]. 3. **Infrastructure and Environmental Concerns**: Several reports highlight potential environmental impacts and infrastructure threats. These include potential environmental damage due to industrial activities, flooding crises, and threats to infrastructure due to military conflicts [Data: Reports (308, 154, 201, 348, 216, +more)]. 4. **Community Analysis and Threat Assessment**: Several reports provide detailed analyses of specific communities, often centered around a particular location or entity. These analyses typically discuss the key entities within the community, their relationships, and any associated threats. Many reports also include assessments of potential threats associated with the entities or situations they discuss [Data: Reports (323, 596, 518, 583, 179, +more)]. 5. **Health, Humanitarian Concerns, and Emergency Situations**: Some reports focus on health and humanitarian issues, such as the Shereshevsky-Turner Syndrome community and the potential risk of intestinal diseases in the Odessa region. The data also covers emergency situations, including natural disasters, accidents, and attacks [Data: Reports (204, 149, 372, 154, 532, +more)]. These themes provide a comprehensive overview of the current situation, highlighting the complexity and interconnectedness of the issues at hand.

Looking at the results from baseline RAG, we see that none of the listed themes has much to do with the war between the two countries. As anticipated, the vector search retrieved irrelevant text, which was inserted into the LLM's context window. Results that were included were likely keying on the word "theme," resulting in a less than useful assessment of what is going on in the dataset.

Observing the results from GraphRAG, we can clearly see that the results are far more aligned with what is going on in the dataset as a whole. The answer provides the five main themes as well as supporting details that are observed in the dataset. The referenced reports are pre-generated by the LLM for each semantic cluster in GraphRAG and, in turn, provide provenance back to original source material.

https://github.com/microsoft/graphrag/tree/main

https://neo4j.com/blog/knowledge-graph-vs-vectordb-for-retrieval-augmented-generation/?source=post_page-----b1f7db88edd3----------

https://blog.langchain.dev/enhancing-rag-based-applications-accuracy-by-constructing-and-leveraging-knowledge-graphs/

https://blog.langchain.dev/graph-based-metadata-filtering-for-improving-vector-search-in-rag-applications/

https://blog.langchain.dev/query-construction/

What is GraphRAG?
GraphRAG uses LLMs to construct knowledge graphs from data and answer user queries based on private datasets. Unlike traditional RAG methods that rely on vector similarity, GraphRAG leverages LLM-generated knowledge graphs to significantly enhance question-and-answer performance, especially for complex document analysis.

Why graphs?
Graphs excel at connecting different pieces of information through their relationships, making them ideal for synthesizing insights and performing complex analytical tasks, unlike flat, unconnected data.

When should you use it?
👉 GraphRAG could be ideal for scenarios requiring deep information discovery and analysis across multiple noisy documents or datasets containing mis/dis-information.
👉 GraphRAG can also be beneficial in scenarios where a straightforward semantic search based solely on the query may not suffice. The answer may reside in related information rather than the query term alone.

When to skip it?
👉 GraphRAG's effectiveness depends on well-structured indexing, particularly in unique datasets with domain-specific concepts, indexing can be resource-intensive, requiring careful setup and testing before extensive use.
👉 GraphRAG's effectiveness also hinges on its ability to accurately identify nodes related to a query, a task that is super straightforward in semantic search.