



Research Our research

All Microsoft

[Return to Blog Home](#)

Microsoft Research Blog

LazyGraphRAG: Setting a new standard for quality and cost

Published November 25, 2024

By [Darren Edge](#), Senior Director; [Ha Trinh](#), Senior Data Scientist; [Jonathan Larson](#), Partner Data Architect

Share this page

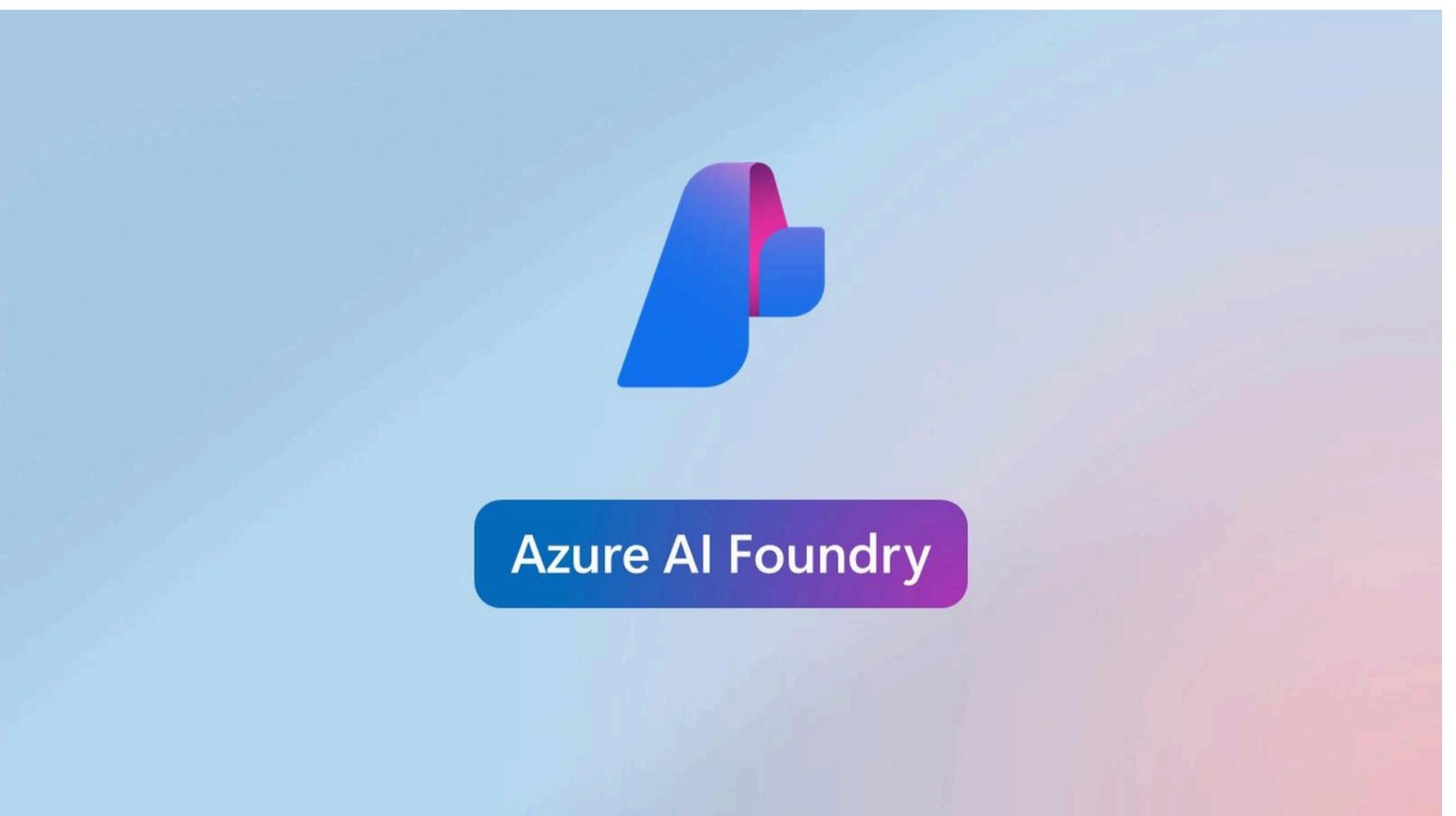


Editor's note, June 6, 2025 – If you would like to use LazyGraphRAG, the technology has been integrated into [Microsoft Discovery](#), an agentic platform for scientific research built on Azure, and into [Azure Local](#) services as part of a [public preview](#).

Affordable GraphRAG for every use case

The [GraphRAG project](#) aims to expand the class of questions that AI systems can answer over private datasets by leveraging the implicit relationships within unstructured text.

A key advantage of GraphRAG over conventional vector RAG (or “semantic search”) is its ability to answer *global queries* that address the entire dataset, such as “what are the main themes in the data?”, or “what are the most important implications for X?”. Conversely, vector RAG excels for *local queries* where the answer resembles the query and can be found within specific text regions, as is typically the case for “who”, “what”, “when”, and “where” questions.



Azure AI Foundry Labs

Get a glimpse of potential future directions for AI, with these experimental technologies from Microsoft Research.

[Azure AI Foundry >](#)

In recent blog posts, we have shared two new query mechanisms that exploit the rich, summary-based data index created by GraphRAG to improve [local search performance](#) and [global search costs](#), respectively.

In this blog post, we introduce a radically different approach to graph-enabled RAG that requires no prior summarization of the source data, avoiding the up-front indexing costs that may be prohibitive for some users and use cases. We call this approach “LazyGraphRAG”.

A key advantage of LazyGraphRAG is its inherent scalability in terms of both cost and quality. Across a range of competing methods (standard vector RAG, [RAPTOR](#), and GraphRAG [local](#), [global](#), and [DRIFT](#) search mechanisms), LazyGraphRAG shows strong performance across the cost-quality spectrum as follows:

- LazyGraphRAG data indexing costs are identical to vector RAG and 0.1% of the costs of full GraphRAG.
- For comparable query costs to vector RAG, LazyGraphRAG outperforms all competing methods on local queries, including long-context vector RAG and GraphRAG [DRIFT](#) search (our recently introduced RAG approach shown to outperform vector RAG) as well as GraphRAG local search.
- The same LazyGraphRAG configuration also shows comparable answer quality to GraphRAG Global Search for global queries, but *more than 700 times lower query cost*.
- For 4% of the query cost of GraphRAG global search, LazyGraphRAG *significantly outperforms all competing methods* on both local and global query types, including GraphRAG global search at the C2 level (the third level of the community hierarchy recommended for most applications).

Blending vector RAG and Graph RAG with deferred LLM use

LazyGraphRAG aims to blend the advantages of vector RAG and Graph RAG while overcoming their respective limitations:

- Vector RAG is a form of *best-first search* that uses the similarity with the query to select the best-matching source text chunks. However, it has no sense of the breadth of the dataset to consider for global queries.
- GraphRAG global search is a form of *breadth-first search* that uses the community structure of source text entities to ensure that queries are answered considering the full breadth of the dataset. However, it has no sense of the best communities to consider for local queries.

LazyGraphRAG combines best-first and breadth-first search dynamics in an iterative deepening manner (Table 1). Compared to the global search mechanism of full GraphRAG, this approach is “lazy” in ways that defer LLM use and dramatically increase the efficiency of answer generation. Overall performance can be scaled via a single main parameter – the *relevance test budget* – that controls the cost-quality trade-off in a consistent manner.

	GraphRAG	LazyGraphRAG
Build index	Uses an LLM to extract and describe entities and their relationships, b) uses an LLM to summarize all observations of each entity and relationship, c) uses graph statistics to optimize the entity graph and extract hierarchical community structure	Uses NLP noun phrase extraction to extract concepts and their co-occurrences, b) uses graph statistics to optimize the concept graph and extract hierarchical community structure
Summarize index	Uses an LLM to summarize entities and relationships in each community	None – the “lazy” approach defers all LLM use until query time
Refine query	None – the original query is used throughout	Uses an LLM to a) identify relevant subqueries and recombine them into a single expanded query, b) refine subqueries with matching concepts from the concept graph
Match query	None – all queries are answered using all community summaries (<i>breadth first</i>)	For each of q subqueries [3-5]: – Uses text chunk embeddings and chunk-community relationships to first rank text chunks by similarity to the query, then rank communities by the rank of their top- k text chunks (<i>best first</i>) – Uses an LLM-based sentence-level relevance assessor to rate the relevance of the top- k untested text chunks from communities in rank order (<i>breadth first</i>) – Recurses into relevant sub-communities after z successive communities yield zero relevant text chunks (<i>iterative deepening</i>) – Terminates when no relevant communities remain or <i>relevance test budget / q</i> is reached
Map answers	Uses an LLM to answer the original query over random batches of community summaries in parallel	For each of q subqueries [3-5]: – Builds a subgraph of concepts from the relevant text chunks – Uses the community assignments of concepts to group related chunks together – Uses an LLM to extract subquery-relevant claims from groups of related chunks as a way of focusing on relevant content only – Ranks and filters extracted claims to fit a pre-defined context window size
Reduce answers	Uses an LLM to answer the original query using the mapped answers	Uses an LLM to answer the expanded query using the extracted map claims

LazyGraphRAG answer quality is state-of-the-art

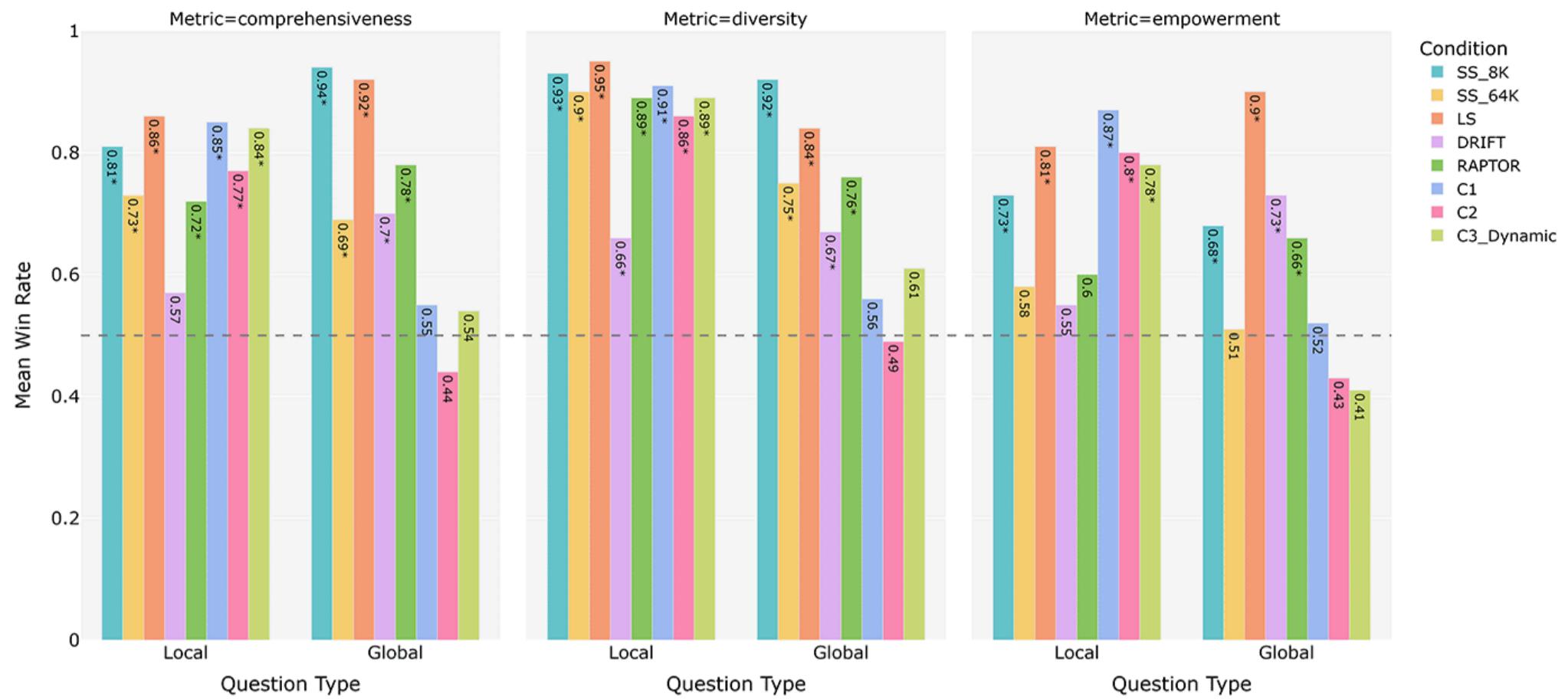
We compared LazyGraphRAG at varying levels of relevance test budget to a range of competing methods, as follows:

- Dataset: 5,590 AP news articles (used with license)
- Queries: 100 synthetic queries (50 local and 50 global), generated using a new method to be described in a future blog post
- Metrics: Comprehensiveness, Diversity, Empowerment (as described [here](#), with an LLM used to compare pairs of answers head-to-head on each metric)
- Conditions: Includes LazyGraphRAG with three relevance test budget settings, in addition to eight competing conditions from GraphRAG and literature (Table 2).

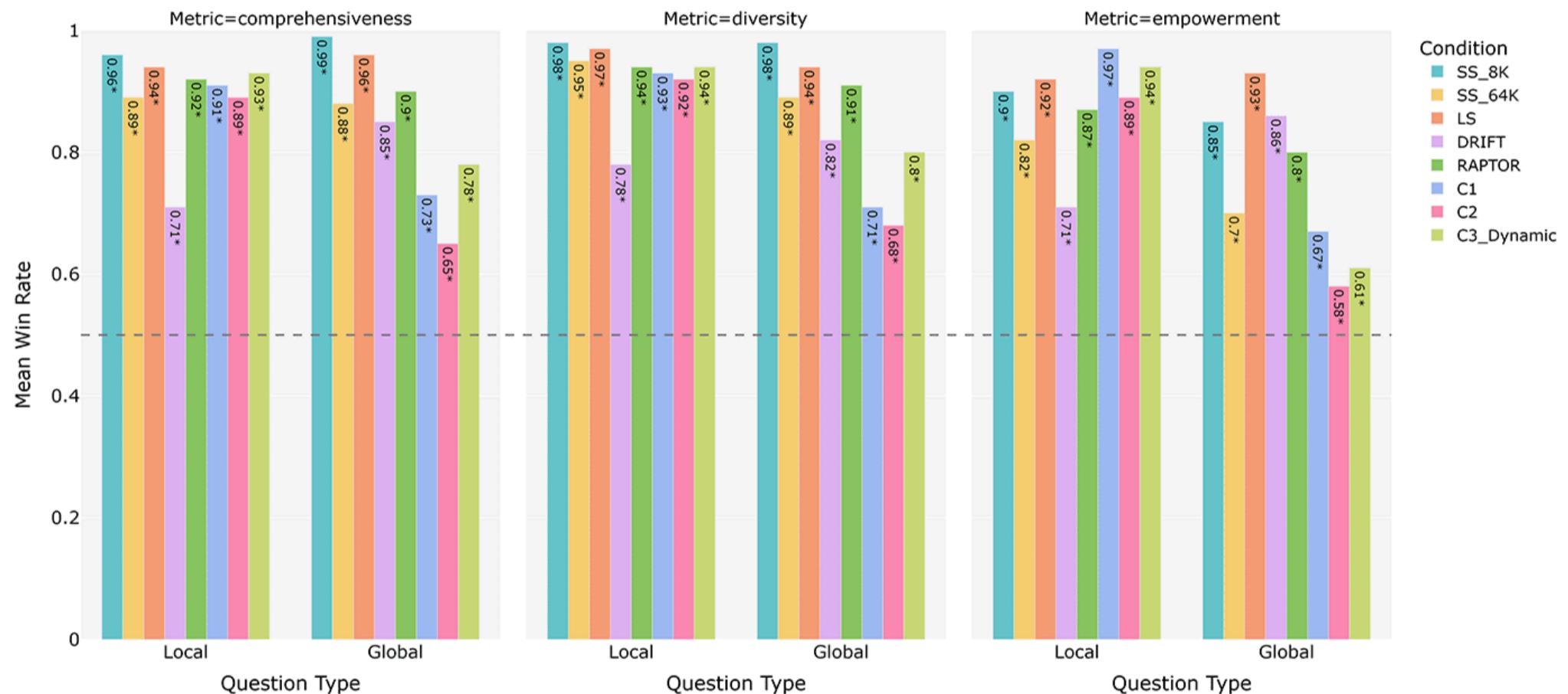
Condition	Description
Z100_Lite	LazyGraphRAG with a relevance test budget of 100 and using a low-cost LLM model at all steps
Z500	LazyGraphRAG with a relevance test budget of 500, using a low-cost LLM for relevance tests and a more advanced (higher cost) LLM for query refinement and map/reduce answer generation
Z1500	LazyGraphRAG with a relevance test budget of 1,500, using a low-cost LLM for relevance tests and a more advanced (higher cost) LLM for query refinement and map/reduce answer generation
C1	GraphRAG Global Search at community level 1
C2	GraphRAG Global Search at community level 2
C3_Dynamic	GraphRAG Global Search at community level 3 using dynamic community selection (a cost saving solution introduced in our previous blog post)
LS	GraphRAG Local Search
DRIFT	GraphRAG DRIFT Search
SS_8K	Semantic Search with vector RAG retrieving 8k tokens of input
SS_64K	Semantic Search with vector RAG retrieving 64k tokens of input (included to examine the effect of long-context window on vector RAG)
RAPTOR	A hierarchical clustering-based approach to RAG that shares some similar features to LazyGraphRAG

Figure 1 shows LazyGraphRAG win rates against each of the eight competing conditions. At the lowest budget level of 100 relevance tests using a low-cost LLM model (same cost as SS_8K), LazyGraphRAG significantly outperforms all conditions on local and global queries, except for GraphRAG global search conditions for global queries. At an increased budget of 500 relevance tests using a more advanced LLM model (4% of the query cost of C2), LazyGraphRAG significantly outperforms all conditions on both local and global queries. LazyGraphRAG's win rates continue to increase as the relevance test budget increases to 1,500, demonstrating its scalability in terms of cost vs. quality.

Mean Win Rates of LazyGraphRAG Z100_Lite vs. Other Conditions by Question Type and Evaluation Metric (*Significant at p=0.05)



Mean Win Rates of LazyGraphRAG Z500 vs. Other Conditions by Question Type and Evaluation Metric (*Significant at p=0.05)



Mean Win Rates of LazyGraphRAG Z1500 vs. Other Conditions by Question Type and Evaluation Metric (*Significant at p=0.05)

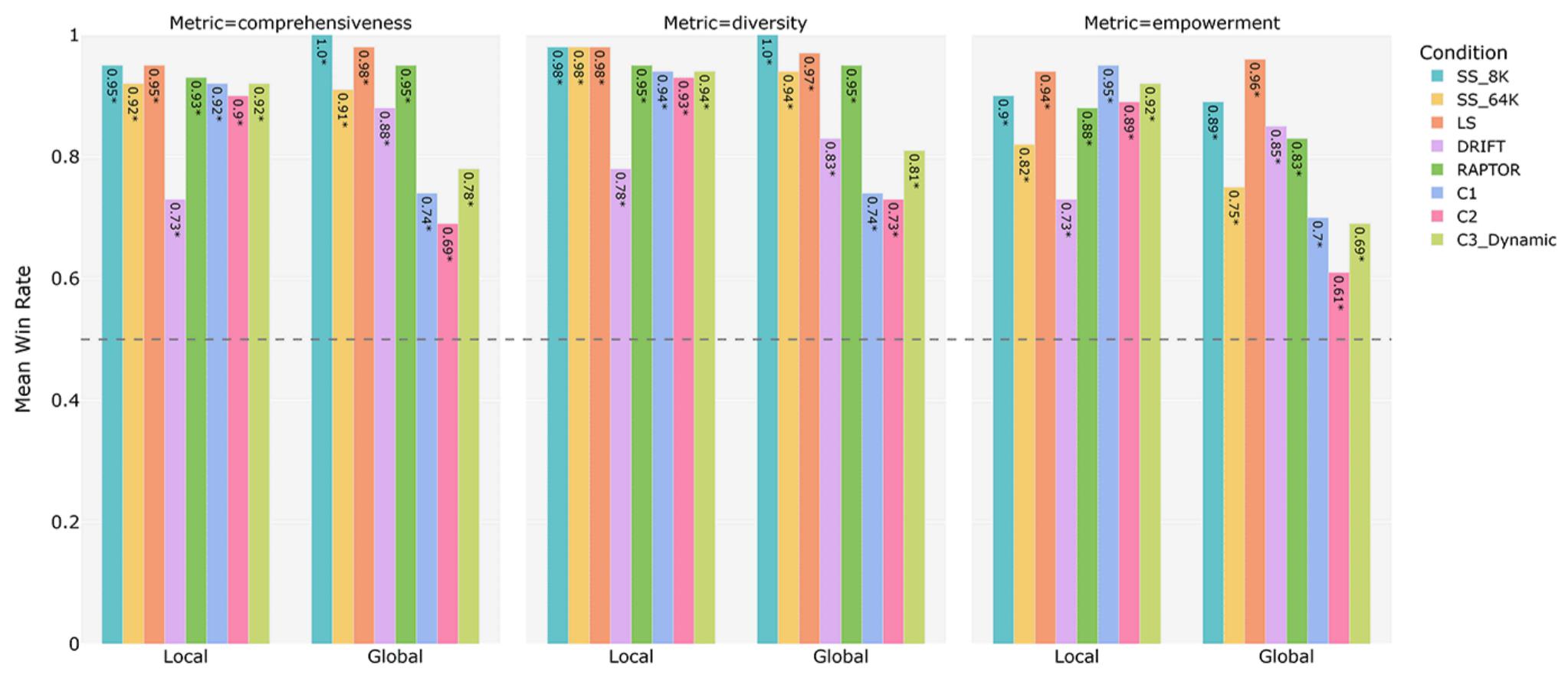


Figure 1. Clustered bar charts showing win rates of LazyGraphRAG with relevance test budgets of (100, 500, 1,500) over competing conditions for each combination of (local, global) queries and (comprehensiveness, diversity, empowerment) metrics

An example comparison is shown in Table 3.

Query: Describe what factors individuals in the United States should consider when selecting a health insurance plan during the open enrollment period for 2024 health insurance plans.

Vector RAG (64K context)	LazyGraphRAG
# Factors to Consider When Selecting a Health Insurance Plan During Open Enrollment for 2024	<p>### Types of Health Insurance Plans</p> <p>When selecting a health insurance plan during the open enrollment period for 2024, individuals in the United States have several options to consider:</p>
Selecting a health insurance plan during the open enrollment period can be a complex process, but understanding the key factors to consider can help individuals make informed decisions. Here are some important aspects to keep in mind:	<ol style="list-style-type: none"> 1. Marketplace Plans: These are available under the Affordable Care Act (ACA) and are categorized into platinum, gold, silver, and bronze levels. Bronze plans generally have lower premiums but higher out-of-pocket costs, while platinum plans have higher premiums but lower out-of-pocket costs [Data: Sources (47411, 47420, 47421)].
## Coverage and Networks	<ol style="list-style-type: none"> 2. Low-Cost Marketplace Plans: These often come with high deductibles, requiring patients to pay thousands of dollars before most coverage kicks in. They also have annual out-of-pocket maximums that can exceed \$9,000 for individuals and \$18,000 for families [Data: Sources (47415, 47416, 47414)].
One of the primary considerations is the coverage network of the plan. Insurers form coverage networks of hospitals and doctors, and it is crucial to ensure that your preferred healthcare providers are included in the network of the plan you are considering. Many marketplace plans, such as Health Maintenance Organizations (HMOs) and Exclusive Provider Organizations (EPOs), do not cover claims for most non-emergency care outside their networks [Data: Sources (47419, 47421, 47420, 47410, 47408)].	<ol style="list-style-type: none"> 3. Exclusive Provider Organizations (EPOs) and Health Maintenance Organizations (HMOs): EPOs and HMOs generally restrict patients to a network of doctors and require a primary care doctor to direct care. They tend to be cheaper but lack out-of-network flexibility [Data: Sources (47420, 47421, 43218, 43217)].
## Costs: Premiums, Deductibles, and Out-of-Pocket Maximums	<ol style="list-style-type: none"> 4. Preferred Provider Organizations (PPOs): These plans allow for out-of-network services but at a higher cost. They offer more flexibility compared to HMOs and EPOs [Data: Sources (43217)].
Understanding the costs associated with a health insurance plan is essential. This includes the monthly premiums, deductibles, copays, and coinsurance. Plans are typically categorized into different levels such as platinum, gold, silver, and bronze, with bronze plans generally having lower premiums but higher out-of-pocket costs [Data: Sources (47411, 47412, 47416, 47414, 47418)]. Additionally, it is important to consider the out-of-pocket maximum, which is the most you will have to pay in a year for covered services before the insurance starts covering all costs [Data: Sources (47416, 47415)].	<ol style="list-style-type: none"> 5. High-Deductible Health Plans (HDHPs): Defined as plans with a deductible of at least \$1,600 for individual coverage or \$3,200 for family coverage, with out-of-pocket maximums of no more than \$8,050 or \$16,100, respectively. HDHPs usually have lower premiums, and sometimes companies contribute to a health savings account (HSA) to help cover the deductible [Data: Sources (43227, 43226)].
## Tax Credits and Cost-Sharing Reductions	<ol style="list-style-type: none"> 6. Medicare Advantage: These are privately run versions of the federal government's Medicare program, mostly for people aged 65 and over. They often include prescription drug coverage and may offer additional benefits like dental or vision coverage not provided by traditional Medicare [Data: Sources (72063, 72061, 72060, 72062)].
Individuals with lower incomes may qualify for income-based tax credits to help reduce the cost of premiums. It is important to accurately estimate your income for the coming year when applying for these credits to avoid having to repay any excess credits received [Data: Sources (47412, 47413)]. Additionally, those with incomes below 250% of the federal poverty level may qualify for cost-sharing reductions if they select a silver plan, which can lower deductibles and out-of-pocket maximums [Data: Sources (47417, 47418)].	<ol style="list-style-type: none"> 7. Short-Term Health Insurance Plans: These plans are limited to three months and can only be renewed for a maximum of four months under new rules. They are intended for temporary coverage but often lack comprehensive benefits [Data: Sources (97999, 97995, 97996, 97997)].
## Special Considerations for Medicaid and Medicare	<h3>### Cost Factors: Premiums, Deductibles, Co-pays, and Out-of-Pocket Maximums</h3>
For those transitioning from Medicaid, it is important to understand that marketplace plans may come with higher premiums and copays compared to Medicaid. Additionally, many people who have been removed from Medicaid may be eligible to enroll in marketplace plans through an extended enrollment period [Data: Sources (6284, 6285, 6286, 6287)]. For seniors, the Medicare open enrollment period allows for switching between traditional Medicare and Medicare Advantage plans, with considerations for prescription drug coverage and additional benefits like dental and vision [Data: Sources (72060, 72061, 72062, 72063, 72064)].	<p>The overall cost of health insurance plans in 2024 is influenced by several factors:</p> <ul style="list-style-type: none"> - Premiums: This is the set monthly cost you pay for your health insurance plan. Premiums have been rising, with a notable increase of 7% for both family and single plans in 2023, partly due to inflation [Data: Sources (83383, 83382, 83384, 83385, 83381, +more)].
## Assistance and Resources	<ul style="list-style-type: none"> - Deductibles: The amount you pay out-of-pocket for health care services before your insurance starts to pay. For HDHPs, the deductible is at least \$1,600 for individual coverage or \$3,200 for family coverage [Data: Sources (43226, 43225)].
Navigating the complexities of health insurance can be challenging, and seeking assistance from insurance agents or government-funded navigators can be beneficial. These professionals can help individuals understand their options and make informed decisions without recommending specific plans [Data: Sources (47424, 47425, 47426)].	<ul style="list-style-type: none"> - Co-pays and Co-insurance: These are the costs you pay each time you receive a medical service. Co-pays are fixed amounts, while co-insurance is a percentage of the service cost.
## Conclusion	<ul style="list-style-type: none"> - Out-of-Pocket Maximums: This is the maximum amount you will pay for covered services in a year. For example, HDHPs have out-of-pocket maximums of no more than \$8,050 for individual coverage or \$16,100 for family coverage [Data: Sources (43227, 43226)].
	<h3>### Provider Networks: In-Network vs. Out-of-Network</h3>
Selecting the right health insurance plan requires careful consideration of various factors including coverage networks, costs, eligibility for tax credits and cost-sharing reductions, and available assistance. By thoroughly evaluating these aspects, individuals can choose a plan that best meets their healthcare needs and financial situation for the upcoming year.	<p>The network of healthcare providers is a crucial factor in selecting a health insurance plan:</p> <ul style="list-style-type: none"> - In-Network Providers: These are doctors and hospitals that have agreements with your insurance plan to provide services at lower rates. Ensuring your preferred

Vector RAG (64K context)	LazyGraphRAG
	<p>doctors and specialists are in-network can save you significant costs [Data: Sources (43216, 47419)].</p> <ul style="list-style-type: none"> – **Out-of-Network Providers**: Services from these providers are usually more expensive and may not be covered at all, except in emergencies. PPO plans offer some out-of-network coverage but at a higher cost, while HMOs and EPOs generally do not cover non-emergency out-of-network care [Data: Sources (43217, 47421)]. <p>### Specific Medical Needs and Services</p> <p>When selecting a health insurance plan, individuals should consider their specific medical needs:</p> <ul style="list-style-type: none"> – **Prescription Drugs**: Ensure that your medications are covered by the plan's formulary, as drug coverage can change annually [Data: Sources (43220, 43218, 43219)]. – **Mental Health Services**: Coverage for mental health treatments is essential, especially with new rules pushing insurers to increase their coverage of these services [Data: Sources (97031, 97028, 97027, 97030, 97033, +more)]. – **Chronic Conditions**: Plans should cover ongoing treatments and medications for chronic conditions. Medicare Supplement Insurance (Medigap) can help cover gaps in Medicare for chronic disease management [Data: Sources (93367, 93368)]. – **Preventive Care**: Coverage for preventive services like cancer screenings and HIV prevention is mandated under the ACA, though its future is uncertain due to ongoing legal battles [Data: Sources (71106, 71109, 71098, 71099, 71100, +more)]. <p>### Key Dates and Steps for Open Enrollment</p> <p>The open enrollment period for 2024 health insurance plans involves several key dates and steps:</p> <ul style="list-style-type: none"> – **Marketplace Plans**: Open enrollment starts on November 1, 2023, and runs through mid-December in most states, ending on January 16, 2024 [Data: Sources (47419, 47411, 47416, 47421, 47409, +more)]. – **Medicare**: Open enrollment for Medicare runs from October 15, 2023, to December 7, 2023. During this period, individuals can choose between traditional Medicare, Medicare Advantage plans, and prescription drug plans [Data: Sources (72061, 72063, 72060, 72062)]. – **Special Enrollment Periods**: Individuals who lose coverage due to life events like job loss or moving may qualify for special enrollment periods. For example, those removed from Medicaid may enroll in marketplace plans through July 2024 [Data: Sources (6288, 6289)]. <p>By considering these factors, individuals can make informed decisions about their health insurance coverage for 2024, ensuring they select plans that best meet their medical needs and financial situations.</p>

Looking forward

LazyGraphRAG shows that it is possible for a single, flexible query mechanism to substantially outperform a diverse range of specialized query mechanisms across the local-global query spectrum, and to do so without the up-front costs of LLM data summarization. Its very fast and almost-free indexing make LazyGraphRAG ideal for one-off queries, exploratory analysis, and streaming data use cases, while its ability to smoothly increase answer quality with increasing relevance test budget makes it a valuable tool for benchmarking RAG approaches in general (e.g., "RAG approach X beats LazyGraphRAG with budget Y for task Z").

Does this mean that all graph-enabled RAG should be lazy? We believe the answer is no, for three reasons:

1. A GraphRAG data index of entity, relationship, and community summaries has use value beyond question answering (e.g., reading and sharing as reports).
2. A GraphRAG data index of entity, relationship, and community summaries, combined with a LazyGraphRAG-like search mechanism, is likely to achieve better results than LazyGraphRAG alone.
3. A new kind of GraphRAG data index designed to support a LazyGraphRAG-like search mechanism (e.g., through pre-emptive claim and topic extraction) is likely to achieve the best possible results.

We will be exploring these directions in the coming period. Stay tuned!

Meet the authors



Darren Edge

Senior Director



Ha Trinh

Senior Data Scientist

[Learn more >](#)

[Learn more >](#)

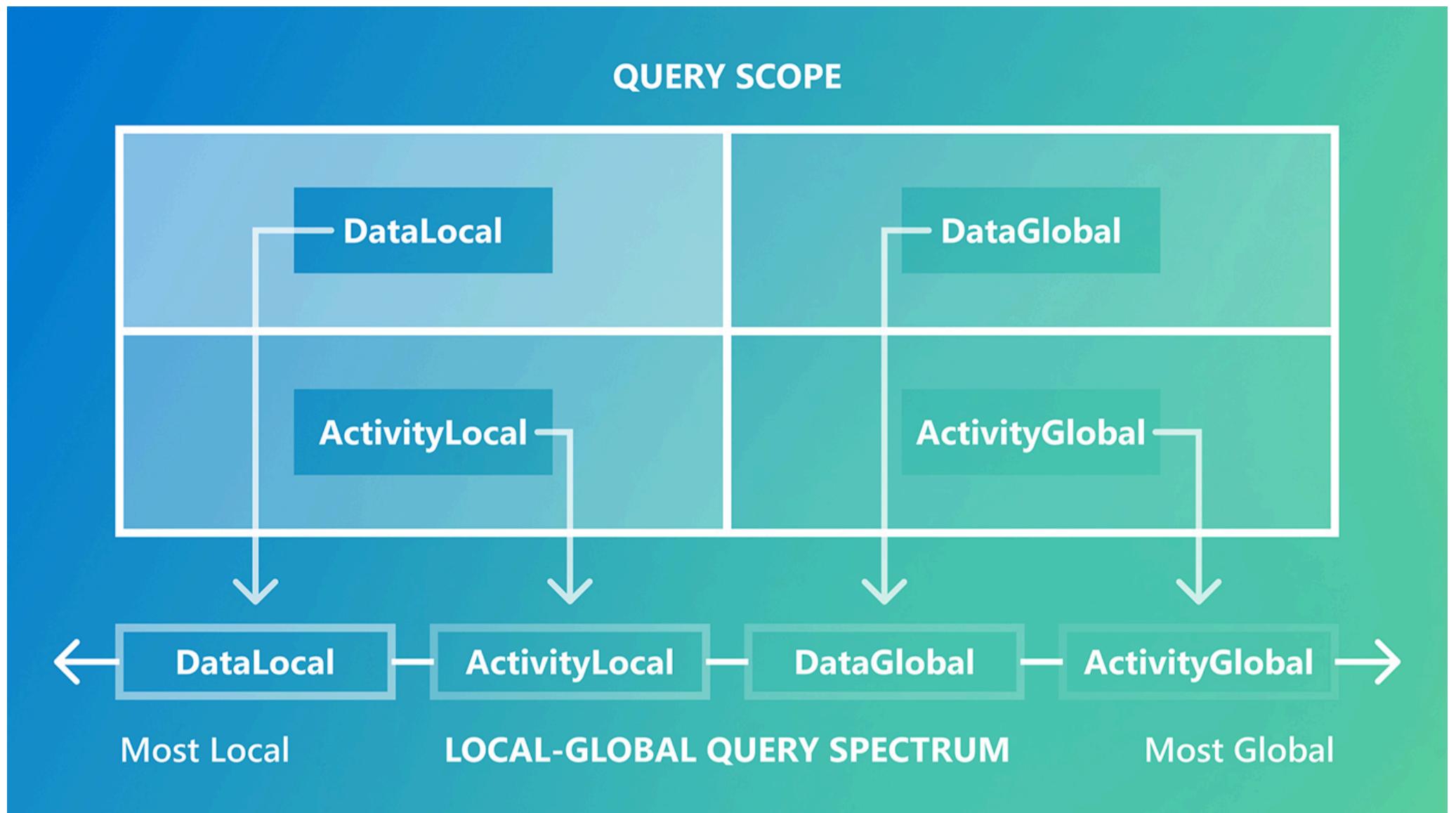


Jonathan Larson

Partner Data Architect

[Learn more >](#)

Continue reading



June 5, 2025

[BenchmarkQED: Automated benchmarking of RAG systems >](#)



December 16, 2024

[Moving to GraphRAG 1.0 – Streamlining ergonomics for developers and users >](#)



July 2, 2024

[GraphRAG: New tool for complex data discovery now on GitHub >](#)



February 13, 2024

[GraphRAG: Unlocking LLM discovery on narrative private data >](#)

[See all blog posts >](#)

Research Areas
 Artificial intelligence
Related tools

GraphRAG

Related projects

Project GraphRAG

Follow us:      Share this page:    

What's new	Microsoft Store	Education	Business	Developer & IT	Company
Surface Pro	Account profile	Microsoft in education	Microsoft Cloud	Azure	Careers
Surface Laptop	Download Center	Devices for education	Microsoft Security	Microsoft Developer	About Microsoft
Surface Laptop Studio 2	Microsoft Store support	Microsoft Teams for Education	Dynamics 365	Microsoft Learn	Company news
Copilot for organizations	Returns	Microsoft 365 Education	Microsoft 365	Support for AI marketplace apps	Privacy at Microsoft
Copilot for personal use	Order tracking	How to buy for your school	Microsoft Power Platform	Microsoft Tech Community	Investors
AI in Windows	Certified Refurbished	Educator training and development	Microsoft Teams	Azure Marketplace	Diversity and inclusion
Explore Microsoft products	Microsoft Store Promise	Deals for students and parents	Microsoft 365 Copilot	AppSource	Accessibility
Windows 11 apps	Flexible Payments	AI for education	Small Business	Visual Studio	Sustainability

 Your Privacy Choices

Consumer Health Privacy

[Sitemap](#) [Contact Microsoft](#) [Privacy](#) [Terms of use](#) [Trademarks](#) [Safety & eco](#) [Recycling](#) [About our ads](#) [© Microsoft 2025](#)