#### GraphRAG methods to create optimized LLM context windows for Retrieval — Jonathan Larson, Microsoft







🔷 Ask

1,665 views Jun 27, 2025 AIEWF 2025 Complete Playlist

Jonathan Larson is a Senior Principal Data Architect at Microsoft Research working in Special Projects. He currently leads a research team focused on the intersection of graph machine learning, LLM memory representations, and LLM orchestration.

His research has led to shipping new features in Bing, Viva, PowerBI. He also shipped new tools to combat tech fraud. Many of the supporting libraries have been open sourced in collaboration on GitHub. Prior to joining Microsoft, Jonathan was Chief Scientist and Technical Fellow at Sotera Defense Solutions on assignment to DARPA, and led a variety of research across several programs. Jonathan has also led large-scale data science efforts at Google, Zillow, and the US Army. Early in his career, he also worked several startups and incubators.

#### About Jonathan Larson

Jonathan Larson is a Senior Principal Data Architect at Microsoft Research working in Special Projects. He currently leads a research team focused on the intersection of graph machine learning, LLM memory representations, and LLM orchestration.



#### From Local to Global: A GraphRAG Approach to Query-Focused Summarization

Darren Edge<sup>1†</sup> Ha Trinh<sup>1†</sup> Newman Cheng<sup>2</sup> Joshua Bradley<sup>2</sup> Alex Chao<sup>3</sup> Steven Truitt<sup>2</sup> Dasha Metropolitansky<sup>1</sup> Robert Osazuwa Ness Apurva Mody<sup>3</sup>

#### Jonathan Larson

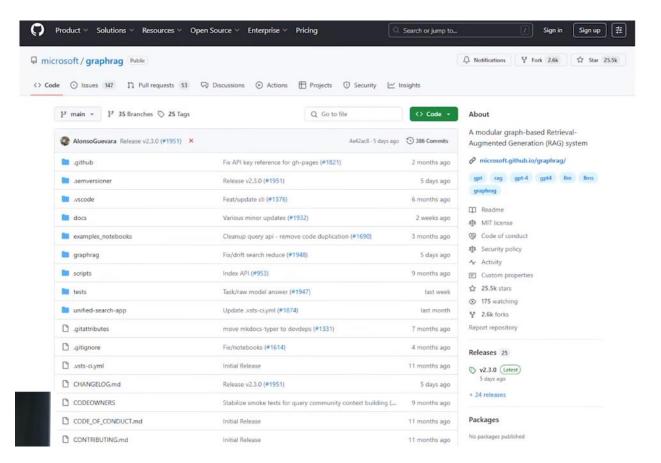
<sup>1</sup>Microsoft Research <sup>2</sup>Microsoft Strategic Missions and Technologies 3Microsoft Office of the CTO

{daedge,trinhha,newmancheng,joshbradley,achao,moapurva, steventruitt, dasham, robertness, jolarso} @microsoft.co

<sup>1</sup>These authors contributed equally to this work

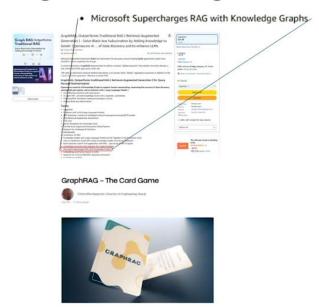
#### Abstract

The use of retrieval-augmented generation (RAG) to retrieve relevant inform the use of retrieval-augmented generation (RAZ) to retrieve reteval in intolition from an external knowledge source enables large language models (LLMs) to answer questions over private and/or previously unseen document collections. However, RAG fails on global questions directed at an entire text corpus, such as "What are the main themes in the dataset?", since this is inherently a query-focused summarization (QFS) task, rather than an explicit retrieval task. Prior QFS methods, meanwhile, do not scale to the quantities of text indexed by typ-ical RAG systems. To combine the strengths of these contrasting methods, we propose GranhRAG, a granh-based approach to question answering over private



## Inspired many other offerings, books, and games

- Neo4j-graphrag-python
- Llamaindex graphrag
- Graphrag-lite
- Nano-graphrag
- Fast-graphrag
- · Langchain-graphrag
- Tiny-graphrag
- LightRAG
- GraphRAG Local UI
- infiniflow/ragflow





### **New Horizons!**

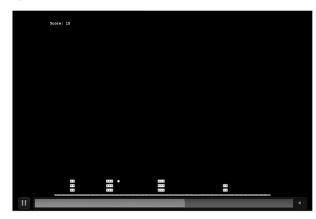
#### LLM memory with structure is a key enabler

- New Vertical: GraphRAG for Code
- New OSS Release Today: BenchmarkQED
- LazyGraphRAG Results
  - · New Benchmarks
  - Azure Local
  - · LazyGraphRAG for Science

### GraphRAG for Code - Repo Understanding An Example

Terminal based Python game where the player earns points by avoiding obstacles and loses points if they collide.

- · LLM has never seen this code before
- Small enough for human to know holistically for ground truth
- Complex enough that the LLM couldn't reason well over it
- ~200 lines of code across 7 files





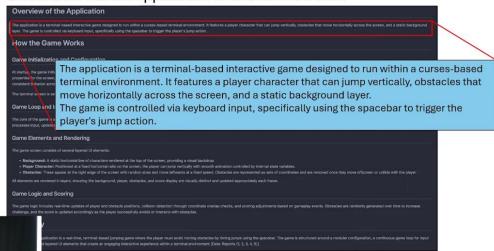
### Regular RAG - Q&A

"Describe what the application is and how it works."



## GraphRAG for Code - Q&A

"Describe what the application is and how it works."



### GraphRAG for Code - Translation

- Simply including all code directly in the context window fails
- · Python -> Rust
  - Involved language specific libraries (E.g. curses)
  - · Game codebase uses abstract classes
  - Rust has highly different semantics than Python
- Building Rust codebase generated complete working code out of the box

```
| Carried | Carr
```

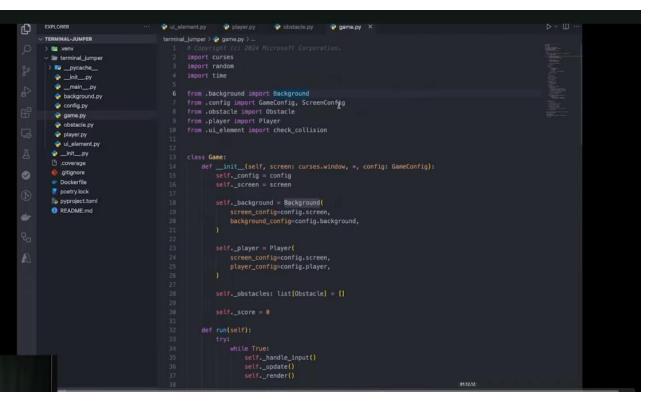
```
EXPLORER
                                                    ui_element.py
                                                                          player.py X pobstacle.py
0
       V TERMINAL-JUMPER
                                                     terminal_jumper > → player.py > ...
5 class Player(UIElement):
         > 💷 .venv

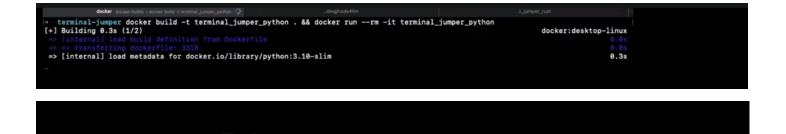
√ Im terminal_jumper

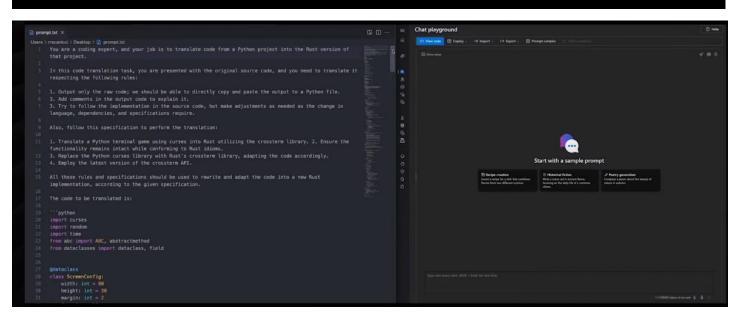
          > pycache_
            __main__py_ background.py
                                                                  def update(self):
            config.py
                                                                      if self._is_jumping:
    self._jump_offset += self._jump_direction
            obstacle.py
            player.py
                                                                            self._jump_direction = 1
elif self._jump_offset == 0:
           coverage.
                                                                                  self._is_jumping = False
self._jump_direction = -1
            .gitignore
           Dockerfile
poetry.lock
           pyproject.toml
           README.md
```

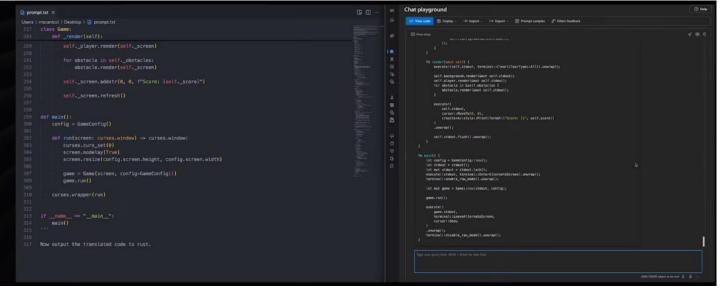
```
EXPLORER
                                                              ul_element.py 🙋 player.py
                                                                                                               obstacle.py X game.py
0
          TERMINAL - HIMPER
                                                               terminal_jumper > 😜 obstacle.py >
          > 🚾 .venv

~ 🗃 terminal_jumper
                                                                        from .config import ObstaclesConfig, ScreenConfig from .ui_element import UIElement
              __main__.py
_background.py
                                                                         class Obstacle(UIElement):
              game.py
               obstacle.py
               ui_element.py
              init_py
                                                                                     self._base_y = screen_config.max_y
self.x_width = random.randint(1, obstacle_config.max_width)
self.y_width = random.randint(1, obstacle_config.max_height)
              Dockerfile
                                                                                     self.render_character = obstacle_config.render_character
self.positions_to_render = self._create_positions_to_render()
              @ README.md
                                                                                        (x, y)
for x in range(self._base_x, self._base_x + self.x_width)
for y in range(self._base_y - self.y_width, self._base_y)
                                                                               def update(self) -> None:
    self._base_x -= self._obstacle_config.speed
```









PRO	OBLEMS 12 OUTPUT	DEBUG CONSOLE TERMINAL PORTS AZURE SPELL CHECKER ··· Filter (e.g. tex		V F ∧ X
	Code	Message	File	Source
	rustc(Click for full co	use of moved value: 'player_config' value used here after move	src/main.rs [Ln 197, Col 31]	rustc
	rustc[Click for full co	use of moved value: 'obstacle_config' value used here after move	src/main.rs [Ln 273, Col 31]	rustc
	rustc Click for full co	no field 'positions_to_render' on type '&impl UIElement' unknown field	src/main.rs [Ln 310, Col 33]	rustc
	rustc(Click for full co	no field 'positions_to_render' on type '&impl UIElement' unknown field	src/main.rs [Ln 311, Col 22]	rustc
	rustc(Click for full co	no method named `clone` found for struct `ScreenConfig` in the current scope items from tra	src/main.rs [Ln 332, Col 55]	rustc
	rustc(Click for full co	no method named 'clone' found for struct 'BackgroundConfig' in the current scope items fro	src/main.rs [Ln 332, Col 82]	rustc
	rustc(Click for full co	no method named 'clone' found for struct 'ScreenConfig' in the current scope items from tra	src/main.rs [Ln 333, Col 47]	rustc
	rustc Click for full co	no method named 'clone' found for struct 'PlayerConfig' in the current scope items from trai	src/main.rs [Ln 333, Col 70]	rustc
	rustc Click for full co	no method named 'clone' found for mutable reference '&mut Obstacle' in the current scope i	src/main.rs [Ln 370, Col 45]	rustc
	ruste(Click for full co	no method named 'clone' found for struct 'ScreenConfig' in the current scope items from tra	src/main.rs [Ln 377, Col 36]	rustc
	rustc(Click for full co	no method named `clone` found for struct `ObstaclesConfig` in the current scope items from	src/main.rs [Ln 378, Col 39]	rustc
	rustc(Click for full co	unused import: 'ExecutableCommand' '#[warn(unused_imports)]' on by default	src/main.rs [Ln 6, Col 5]	rustc

```
"namespace",
'isport",
'preprocessor',
'error'

}

*deror'

*arror'

*arror
```

```
EXPLORER
                                                        @ ui_element.rs X
中
       V TERMINAL_JUMPER_RUST [ □ □ 0 0 src > 0 ul_element.rs
         W B STC
             background.rs
             @ config.rs
             @ game.rs
             @ main.rs
             @ obstacle.rs
                                                                   pub trait UIElement {
             @ ui_element.rs
                                                                        fn screen_config(&self) -> ScreenConfig;
fn render_character(&self) -> char;
fn positions_to_render(&self) -> &Vec<(usize, usize)>;
            target
             .gitignore
            Cargo.lock
                                                                        fn update(&mut self):
0
            T Cargo.tomi
             Dockerfile
                                                                             let screen_config = self.screen_config().clone();
let render_character = self.render_character();
                                                                                          stdout.execute(cursor::MoveTo(x as u16, y as u16)).unwrap();
                                                                   // Function to check collision between two UIElements
pub fn check_collision(element_0: &dyn UIElement, element_1: &dyn UIElement) → bool {
  for &(x0, y0) in element_0.positions_to_render() {
                                                                              if element_1.positions_to_render().iter().any(|\delta(x1, y1)| \times 1 = x0 \delta \delta y1 = y0) {
```

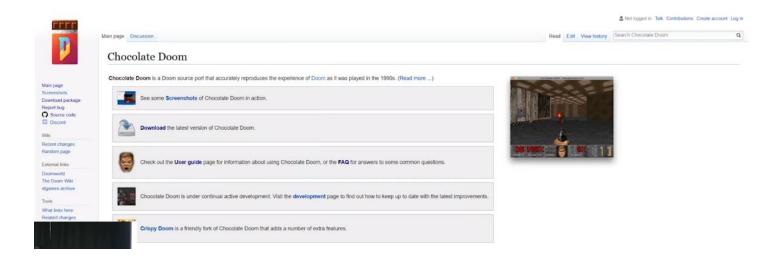
```
⊚ game.rs ×

        TERMINAL JUMPER RUST DE DE
                                                   src > @ game.rs
        ∨ as src
           Ø background.rs
           @ config.rs
           @ game.rs
            @ main.rs
            @ obstacle.rs
            player.rs
                                                                    self.background.render(&mut self.stdout);
self.player.render(&mut self.stdout);
          le target
           .gitignore
           Cargo.lock
0
           T Cargo.toml
           Dockerfile
                                                                       write!(self.stdout, "Score: {}", self.score;.unwrap();
self.stdout.flush().unwrap();
څ
                                                                  stdout.execute(terminal::EnterAlternateScreen).unwrap();
stdout.execute(cursor::Hide).unwrap();
```

### Chocolate Doom

Original 30+ year old Doom Source updated to run on modern OS

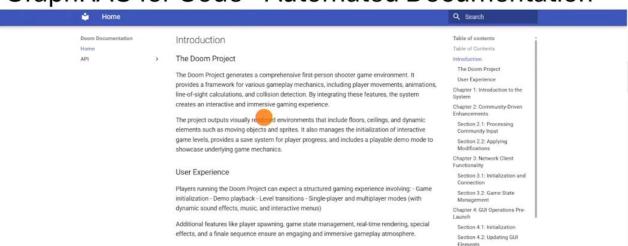
- Doom codebase is
  - · 96K+ lines of code
  - 231 files
  - · 800K+ tokens



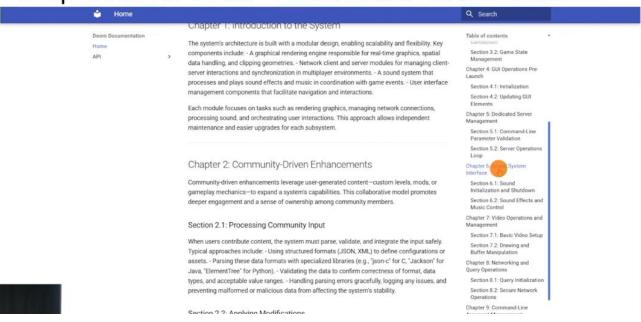
## GraphRAG for Code - Automated Documentation



# GraphRAG for Code - Automated Documentation



## GraphRAG for Code - Automated Documentation



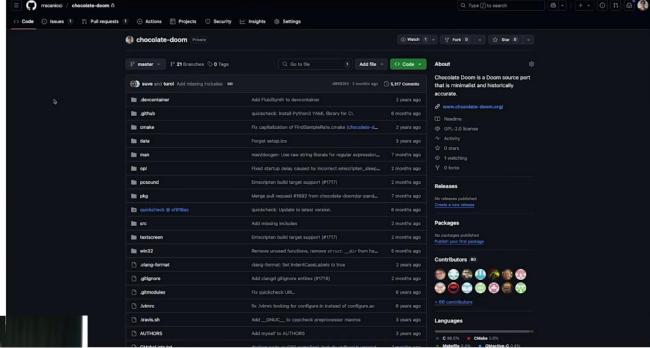
## Code Feature Development

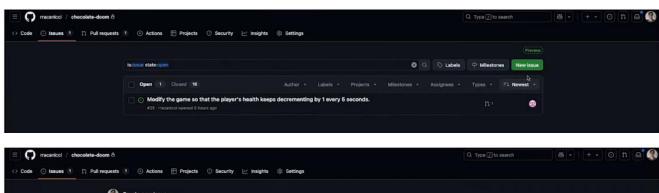
### Feature: Add ability for players to jump in Doom

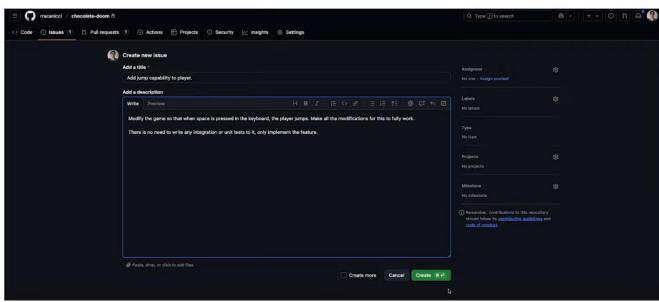
- Existing agents fail on complex tasks
  - Multi-file modifications cause problems
  - · Lack of understanding of how everything fits together

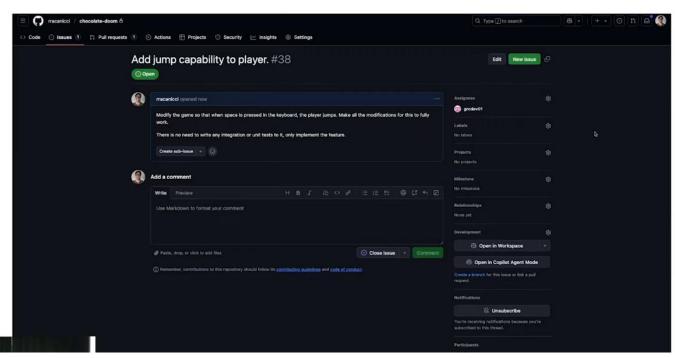
This is where **GraphRAG** can actually help

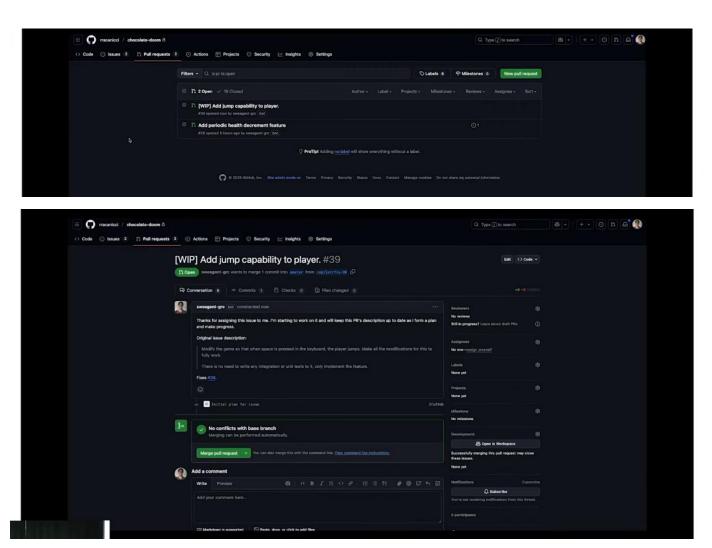


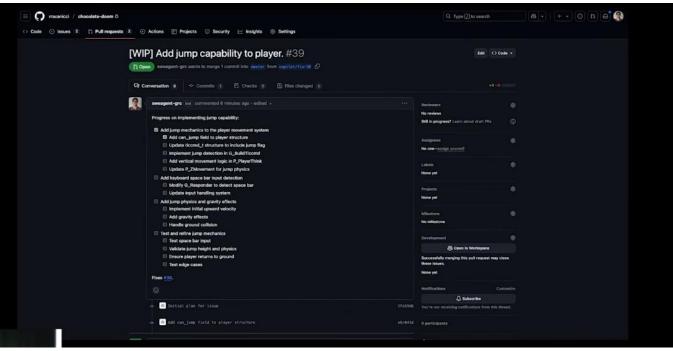


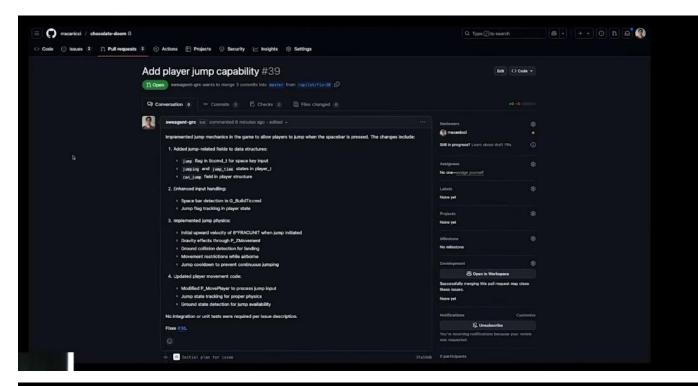


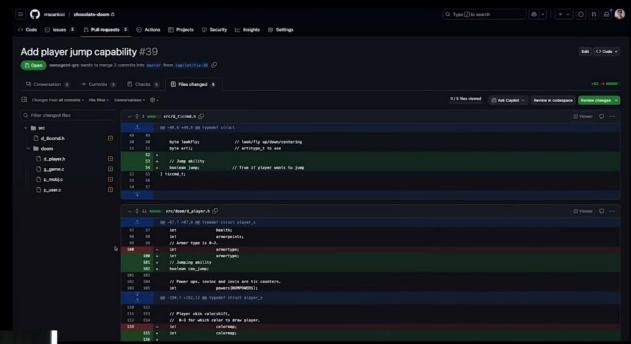


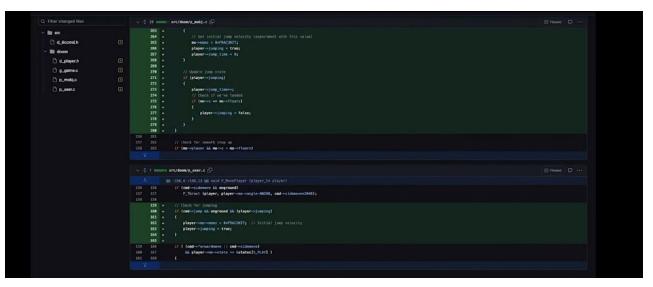








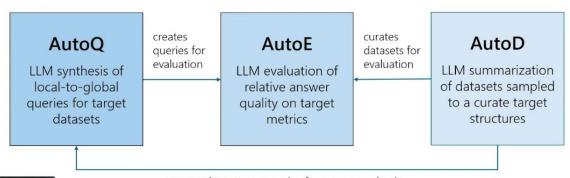






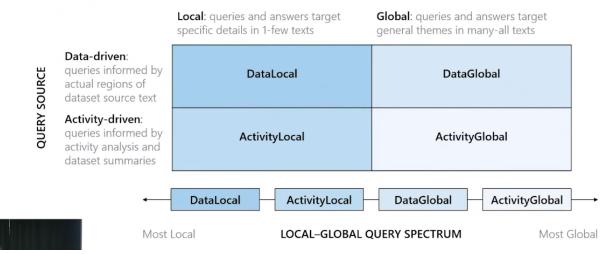
## BenchmarkQED: Automated RAG Benchmarking

Available now at <a href="https://github.com/microsoft/benchmark-qed">https://github.com/microsoft/benchmark-qed</a>



### AutoQ: Synthetic queries with variable data scope

#### **QUERY SCOPE**



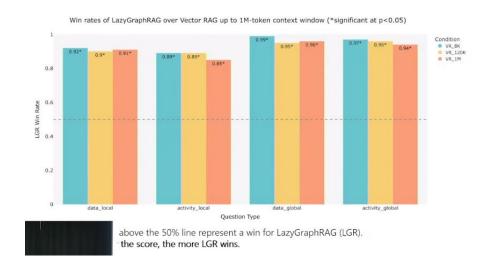
## AutoQ: Synthetic queries with variable data scope

#### **QUERY SCOPE**



## AutoE: Evaluation by query class & quality metric

Composite score of (Comprehensiveness, Diversity, Empowerment, and Relevance)



#### **Evaluation insights**

LazyGraphRAG significantly beat VectorRAG

- long context (1M tokens) makes little difference to VectorRAG performance
- LazyGraphRAG excels for queries spanning the local– global spectrum
- LazyGraphRAG was ~10% the cost of 1M token Vector RAG

### Azure Local

Azure Arc Blog

Caseh this community





ATURE ARC BLOG S MIN READ

Transforming On-Premises Data with RAG Capabilities on Azure Local



Discovering the Advanced Capabilities of RAG on Azure Local

Support for Hybrid Search, and soon Lazy Graph RAG, allowing robust, fast, low-coindexing and providing quality and relevant answers regardless of query type.

Announcing

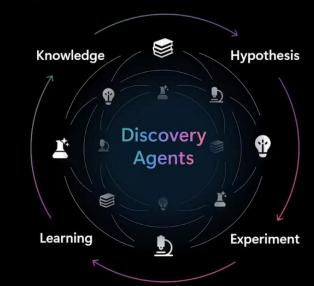
# **Microsoft Discovery**

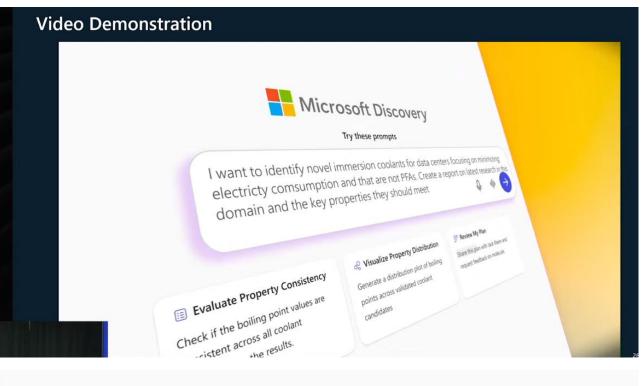
Graph-based scientific co-reasoning

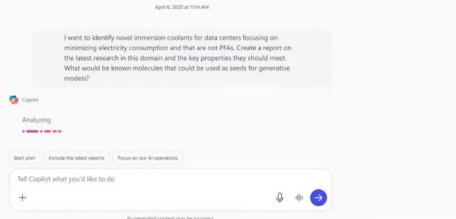
Specialized discovery agents for conducting research

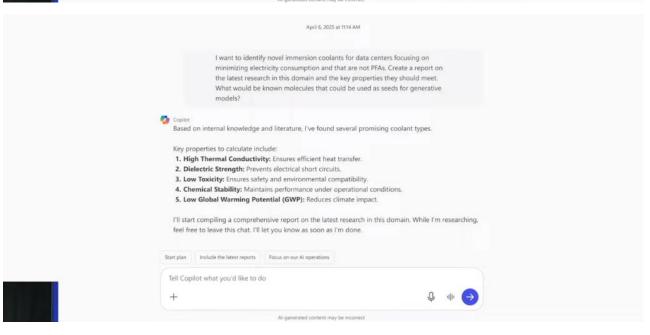
Extensible and enterprise-ready

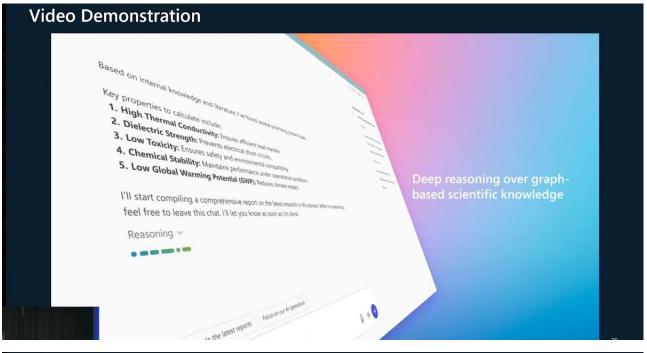
# Transforming the scientific method with Agentic Al

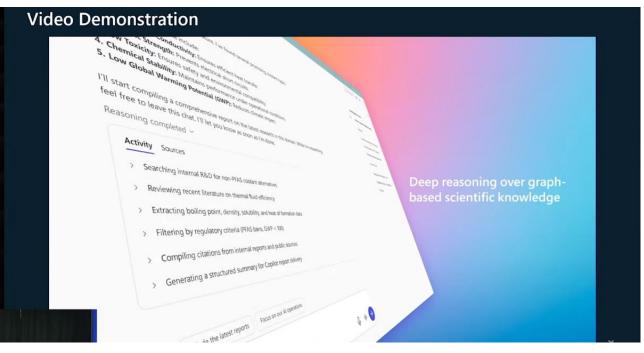


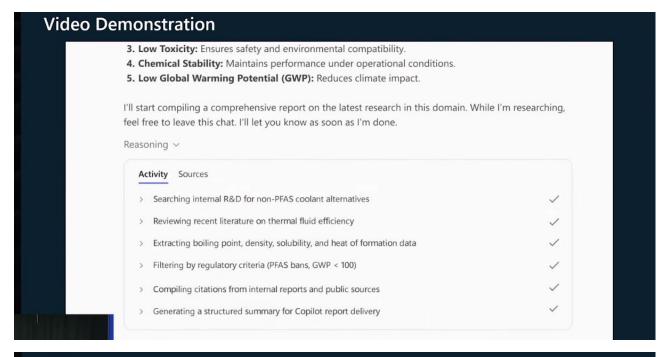












### Takeaways:

- 1. LLM memory with structure is a key enabler
- 2. Agents can massively amplify this power