Sandia National Laboratories

Exceptional service in the national interest

REVOLUTIONIZING KNOWLEDGE MANAGEMENT WITH PROJECT I.R.I.S.

Intelligent Retrieval and Information System

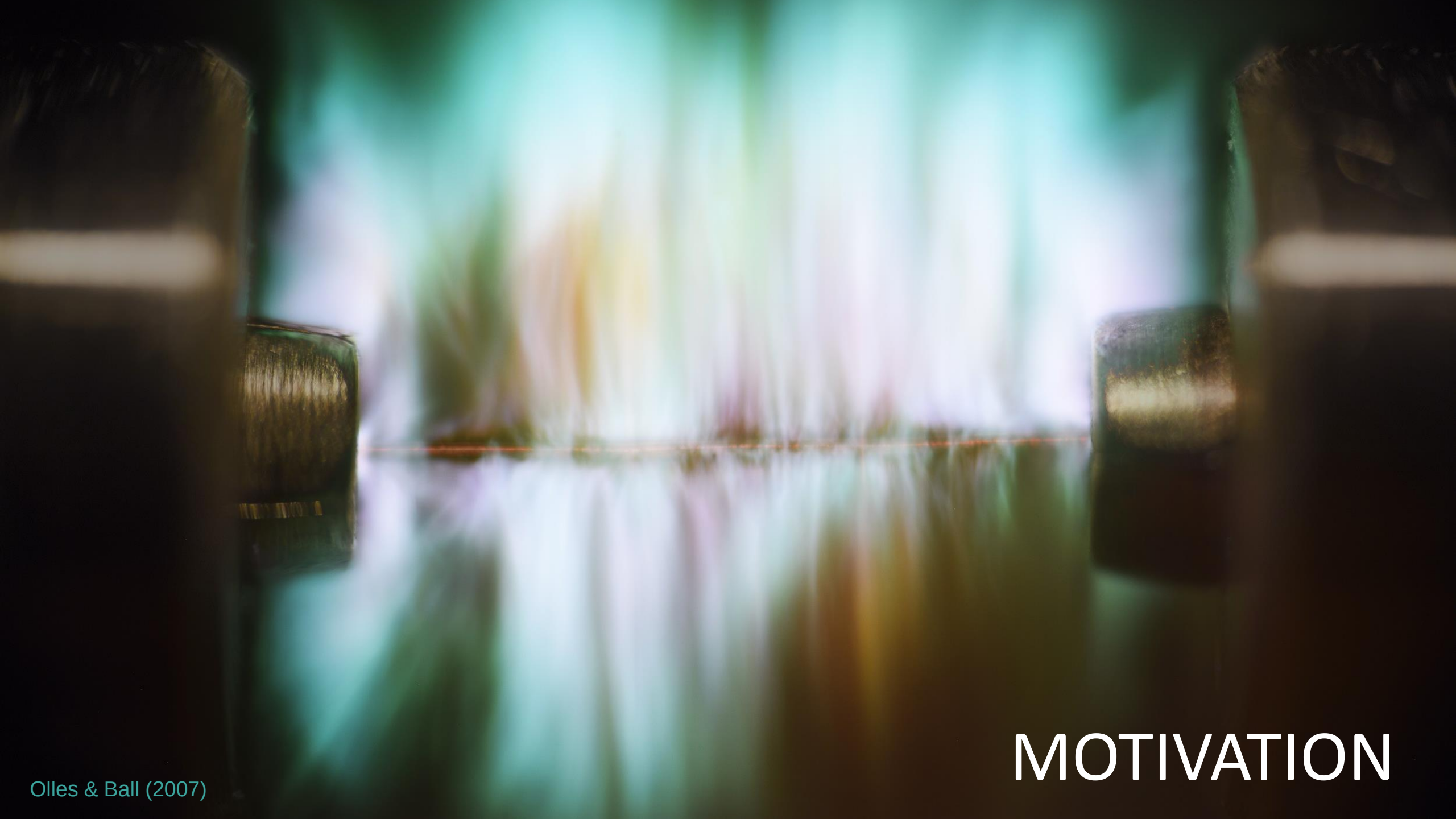
Christopher J. Garasi, Ph.D

In collaboration with David Dumas, Aniket Pant, David Damm, Aaron Moreno, Kevin Pedretti, Angel Beltre, Brendon Mizener, Ed Carroll, Jaimie Murdock, Karen Holt, Charity Karcher, Kate Faford-Johnson, Robert Abbott, Alexander Nemecek, Kevin Zhang, Justin Liang, Thomas Chung, Ryan Roche



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2024-14216C

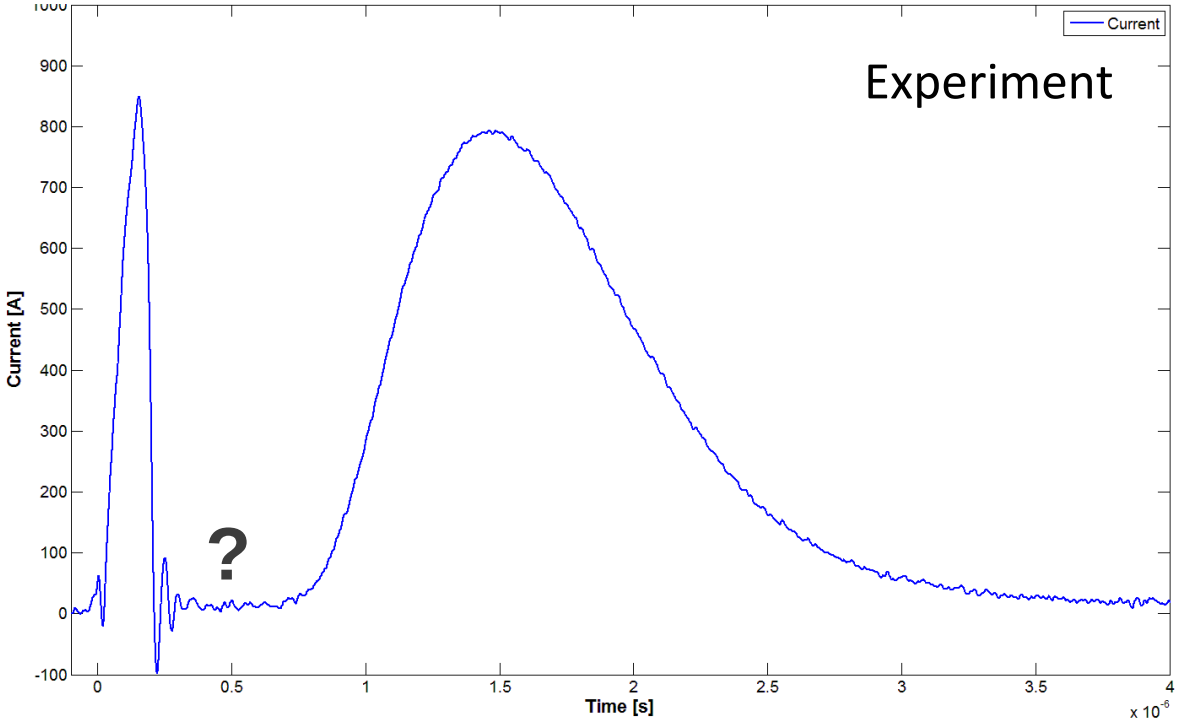


MOTIVATION

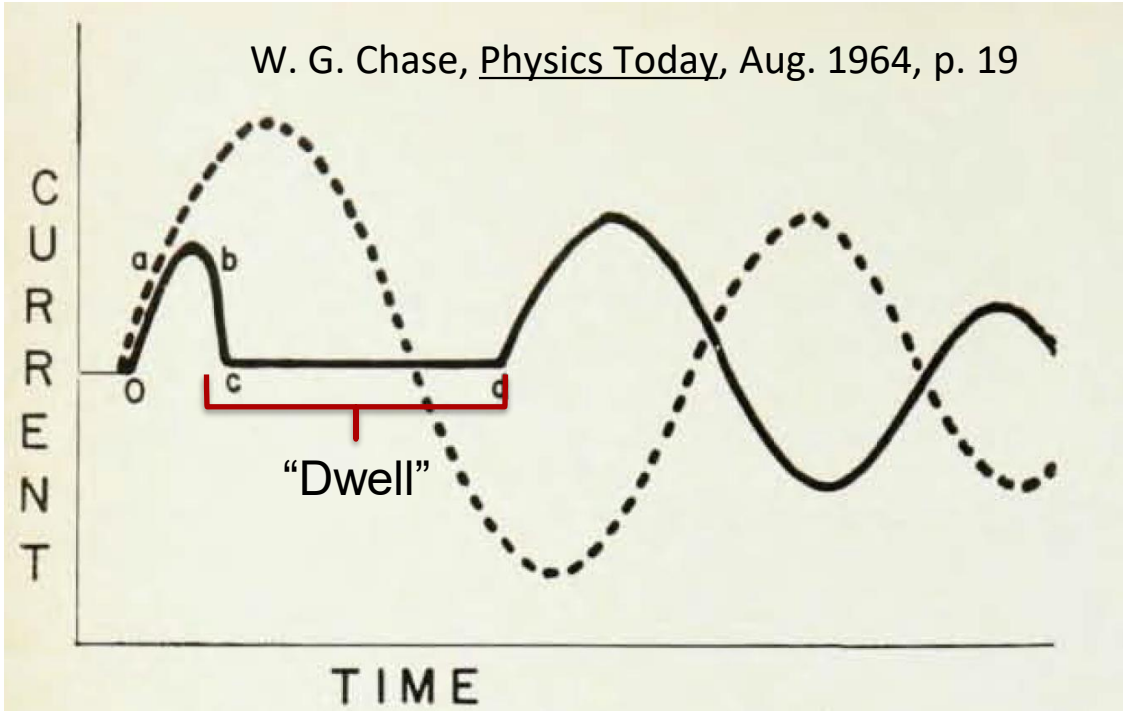
Olles & Ball (2007)

THE ‘DARK’ PAUSE

2014



1964



How long would it take to re-discover a phenomenon that was explained over 50 years ago?

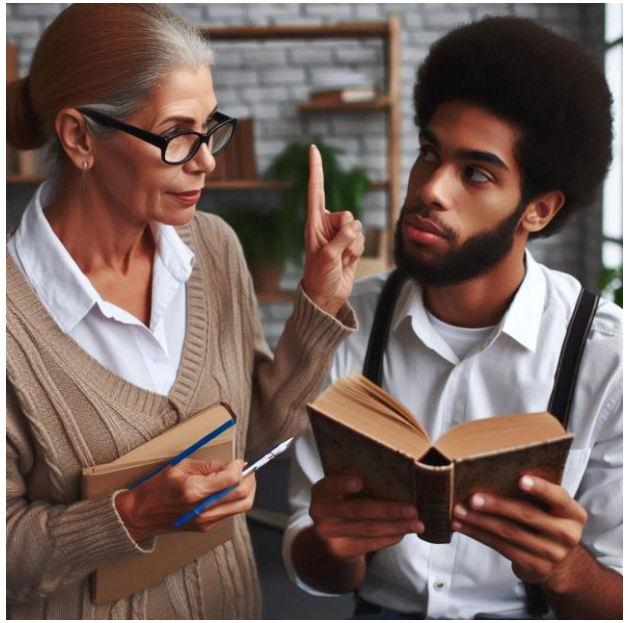
KNOWLEDGE MANAGEMENT CHALLENGES

Knowledge gathered over decades will ultimately be lost if not appropriately stored and can be easily retrieved

- Books & Papers written (information overload; how do you keep up? How do you retain the information?)
- New staff onboarded to decades of information and experience
- Significant cultural and generational differences in learning styles
- Contemporary academic experiences outpace ‘traditional’ laboratory methods
- Staff turnover on an approximately 5-year timeframe

COMMUNITY RESPONSES

- “Let’s host a 2-day seminar”
- “I read the book, you read the book”
- Hard-drives with ‘valuable’ information from retired senior staff
- SME’s manually creating knowledge graphs/maps



PROJECT IRIS GOALS

Create a **‘Virtual SME’**

- Knowledge base creation, information mining, & visualization
- Search & retrieval workflows
- Providing the user with specialized knowledge, guidance, and support
- Simple Q&A
- Not so simple **multi-hop reasoning**
- Use case(s):
  - Electrically exploding wires (UUR)
  - Other application areas are interested

IRIS is a risk-management strategy



TYPES OF QUESTIONS FOR A VSME

A list of questions that progresses from single-hop to increasingly complex multi-hop questions regarding electrically exploded wires

What is an electrically exploded wire?

What applications benefit most from the use of electrically exploded wires?

What are the key materials used in the construction of electrically exploded wires?

Can you explain the mechanism behind the dark pause phenomenon?

What safety precautions are necessary when working with electrically exploded wires?

Describe a notable experiment that demonstrated the effectiveness of electrically exploded wires.

What are the current research trends in the field of electrically exploded wires and their applications?

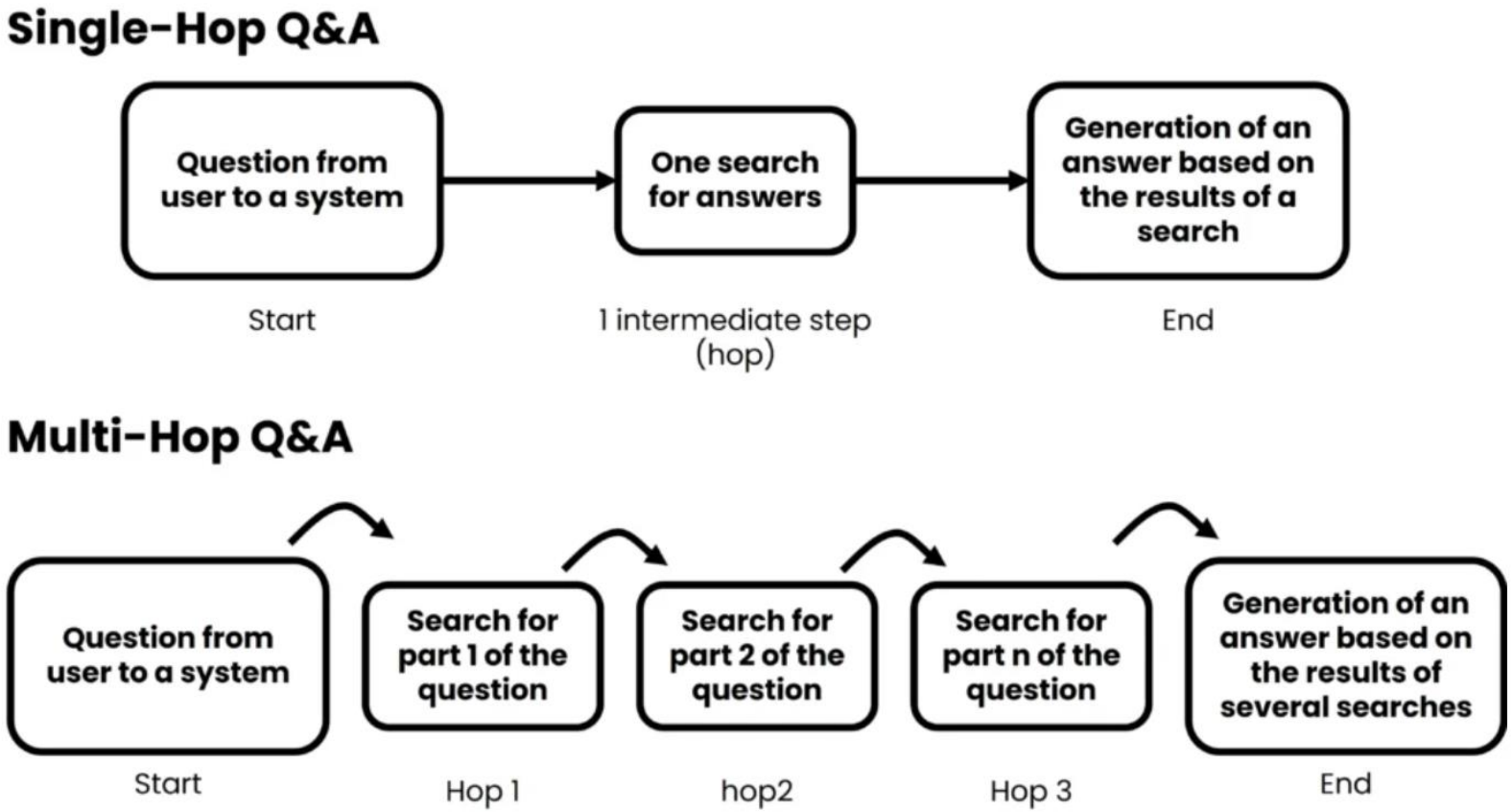
MULTI-HOP REASONING

Multi-hop reasoning involves connecting multiple pieces of information through several intermediate steps to reach a conclusion, making it essential for complex problem-solving where answers are not immediately evident.

Key Characteristics of Multi-hop Reasoning

1. Intermediate Steps
2. Logical Connections
3. Complex Queries: cannot be resolved with a single piece of evidence.
4. Contextual Understanding: Required to navigate through the intermediate steps effectively.

https://www.moveworks.com/us/en/resources/ai-terms-glossary/multi-hop-reasoning



https://ambersearch.de/what-is-multi-hop-qa/

TECHNICAL APPROACH

LARGE LANGUAGE MODELS

What is the “dark pause” in the context of electrically exploding wires?

ChatGPT: In the context of electrically exploding wires, a "dark pause" refers to a phase in the wire explosion process where the **luminosity of the exploding wire diminishes or disappears altogether**. When a wire is subjected to a high-current electrical pulse, it rapidly heats up due to the current passing through it. This intense heating causes the wire to vaporize, creating a bright luminous plasma. Initially, this plasma emits intense light, but as the wire continues to vaporize and disintegrate, **the luminosity may decrease, leading to a period where the plasma appears dimmer or even ceases to emit visible light**. This dark pause is a significant feature in the study of wire explosions and is often associated with specific physical phenomena occurring during the process, such as changes in the wire's material properties or the dynamics of the plasma.



‘Hallucination’: When a LLM is not trained in a specific subject area, resulting in word predictions which appear factual but are not.

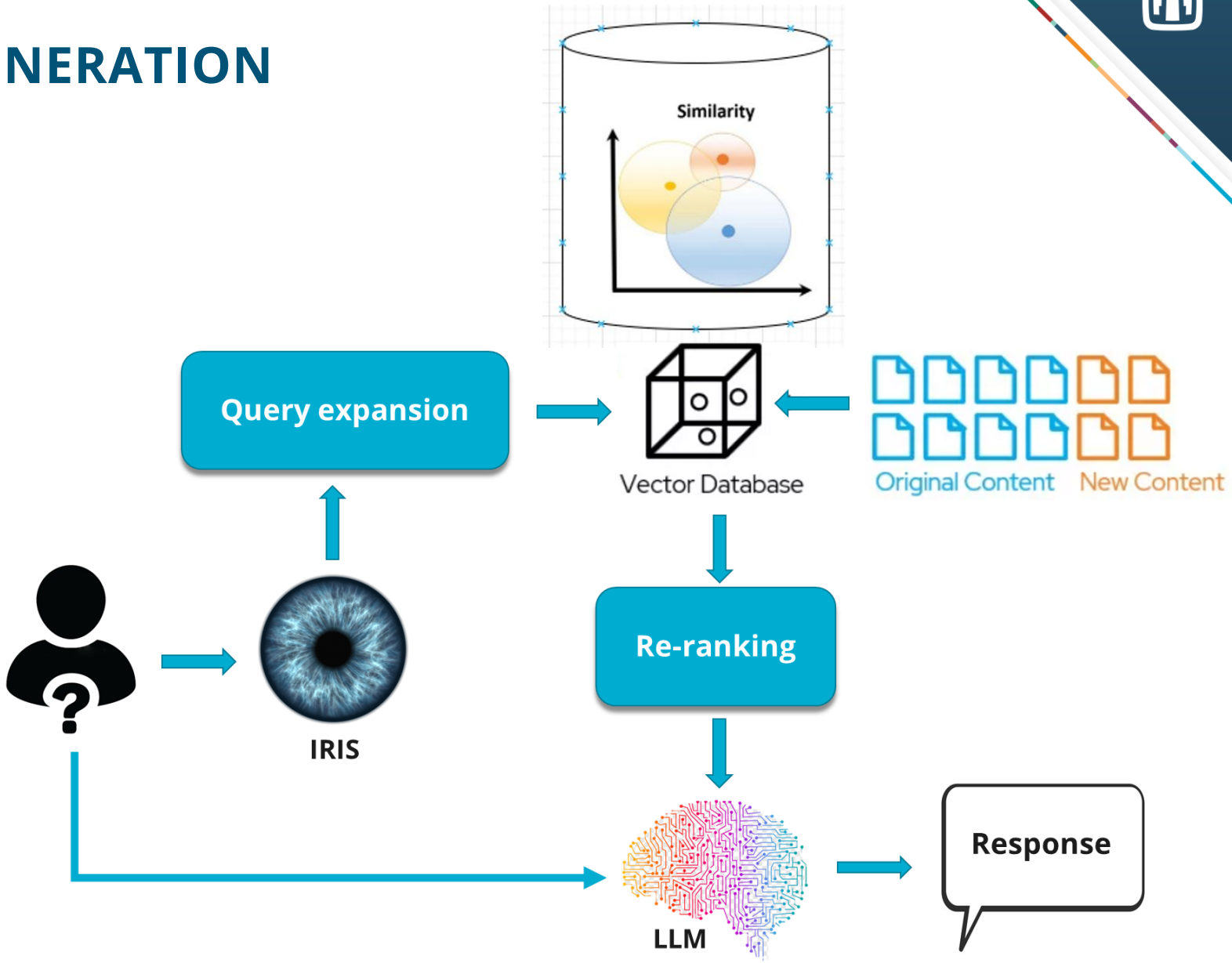
RETRIEVAL AUGMENTED GENERATION

Supplement the gaps in LLM training by using a knowledge base which is populated with technical documentation.

Search the database by creating a vector representation of the query and chunks of the documents, resulting in rapid similarity comparisons and extractions.

Feed the extracted chunks with the query to the LLM for an augmented answer.

'RAG generates an answer to the query & the **references** which the chunks were extracted from



WHAT IS A DARK PAUSE? (IRIS RAG)

IRIS: A dark pause is a phenomenon observed in exploding wire experiments, where the electrical current flowing through a metal wire is halted for a short period of time. This occurs when the metal wire starts to **vaporize due to Joule heating**, increasing its electrical resistivity. The dark pause is followed by a second current surge, which happens a few microseconds later. The phenomenon was first observed optically by Smith and later studied by Wrana using oscillographic techniques. The duration of the dark pause decreases with increasing applied voltage, and it is **attributed to the high electrical resistance of vaporized segments of the wire**. The dark pause is an important feature of the wire explosion, particularly in experiments with low current rates and large wire masses.

Reference	Chunk
2022-Bilbao	more recent works [8]. Previous works by the authors concerned with measuring resistivity limits of metal gas used voltage and current signals during the dark pause [19,20]. On the contrary, other works where the discharge channel of the plasma is imaged [10,21] consider that the dark pause is a mix of liquid droplets surrounded by metal vapor. Initial state of the wire, until the gas expansion is measurable, and presence or absence of liquid droplets inside metal gas are not important for these resistivity limits, but interesting by itself due to the distinct phases of the metal coexisting in the wire during the whole dark pause phenomenon.Exploding wire phenomena, as in the cases of interest to this study, starts with the solid metal wire becoming liquid and later vapor because of the same reason: Joule heating of the wire metal due to the electrical current flowing through it. When the metal starts to vaporize, electrical current begins to halt due to the larger resistivity of gas
1961-Nash	example
2017-Bilbao	example

WHAT IS THE THEORY WHICH EXPLAINS THE DARK PAUSE? (RAG)

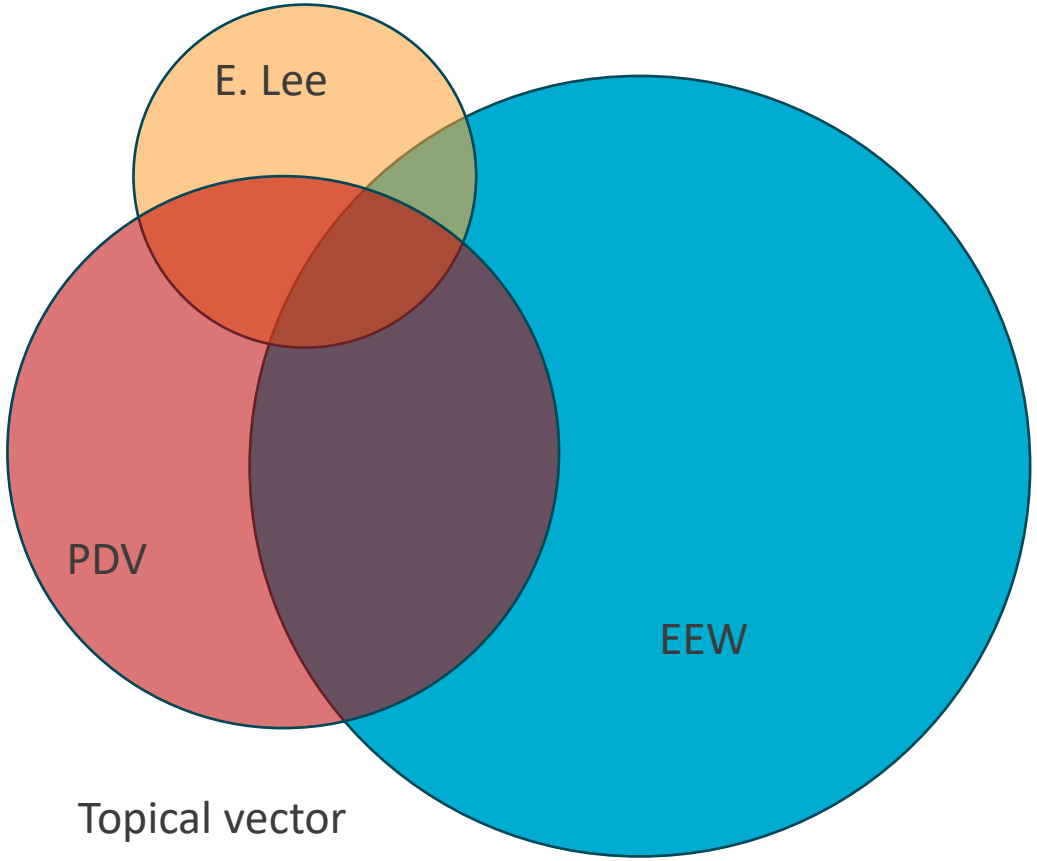
IRIS: The theory that explains the dark pause phenomenon is based on the transition of a solid metal wire into a liquid and later vapor due to Joule heating caused by the electrical current flowing through it. When the metal starts to vaporize, the **electrical current halts due to the larger resistivity of gas**. This effect, known as the "dark pause" or current **"dwell time,"** interrupts the condenser discharge between the initial pulse and the second current surge, which occurs some microseconds later. The duration of the dark pause decreases with increasing applied voltage. **The phenomenon is attributed to the high electrical resistance of vaporized segments of the wire**.

In recent works, the dark pause is considered a **mix of liquid droplets surrounded by metal vapor**…….

Reference
2022-Bilbao
1961-Nash
1962-Bennet
2017-Bilbao

WHAT DID ELIZABETH LEE WRITE ABOUT EBW INITIATION AND PDV?

Results were mostly generic statements **‘diluted’** by references associated with the topic, but NOT at the intersection of the required areas to get the correct answer.



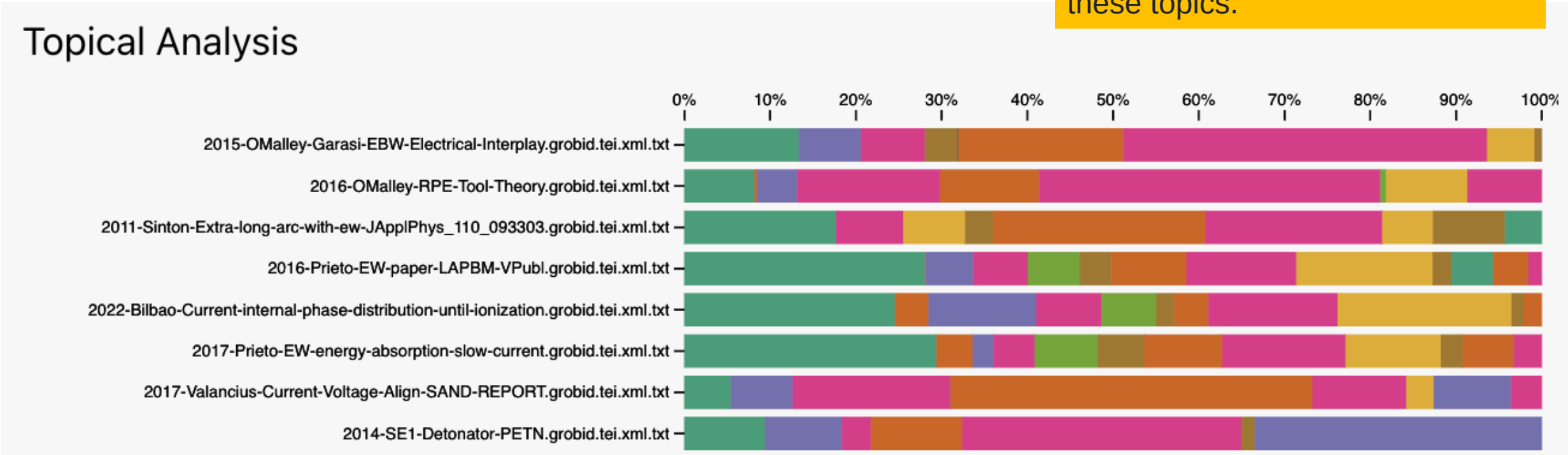
Reference
2014-E-Lee
1963-Leopold
1968-Tucker
2010-Hrousis
2017-Lee
2018-Rae

‘RAG works well for simple Q&A queries. It does not perform well for more ‘complex’ queries

TOPIC MODELING

Topic	Words
Topic 0	tungsten, core, region, breakdown, products, corona, air, copper, structure, dense
Topic 1	solution, cylindrical, evolution, array, water, distribution, equation, piston, boundary, velocity
Topic 2	pulse, circuit, pause, copper, inductance, diameter, restrike, air, channel, oscillograms
Topic 3	petn, detonator, detonation, bridgewire, ebw, explosive, burst, detonators, initiation, threshold
Topic 4	point, critical, metals, heating, vapor, volume, state, melting, transition, heat
Topic 5	material, rate, diameter, heating, deposited, core, formation, similar, along, observed

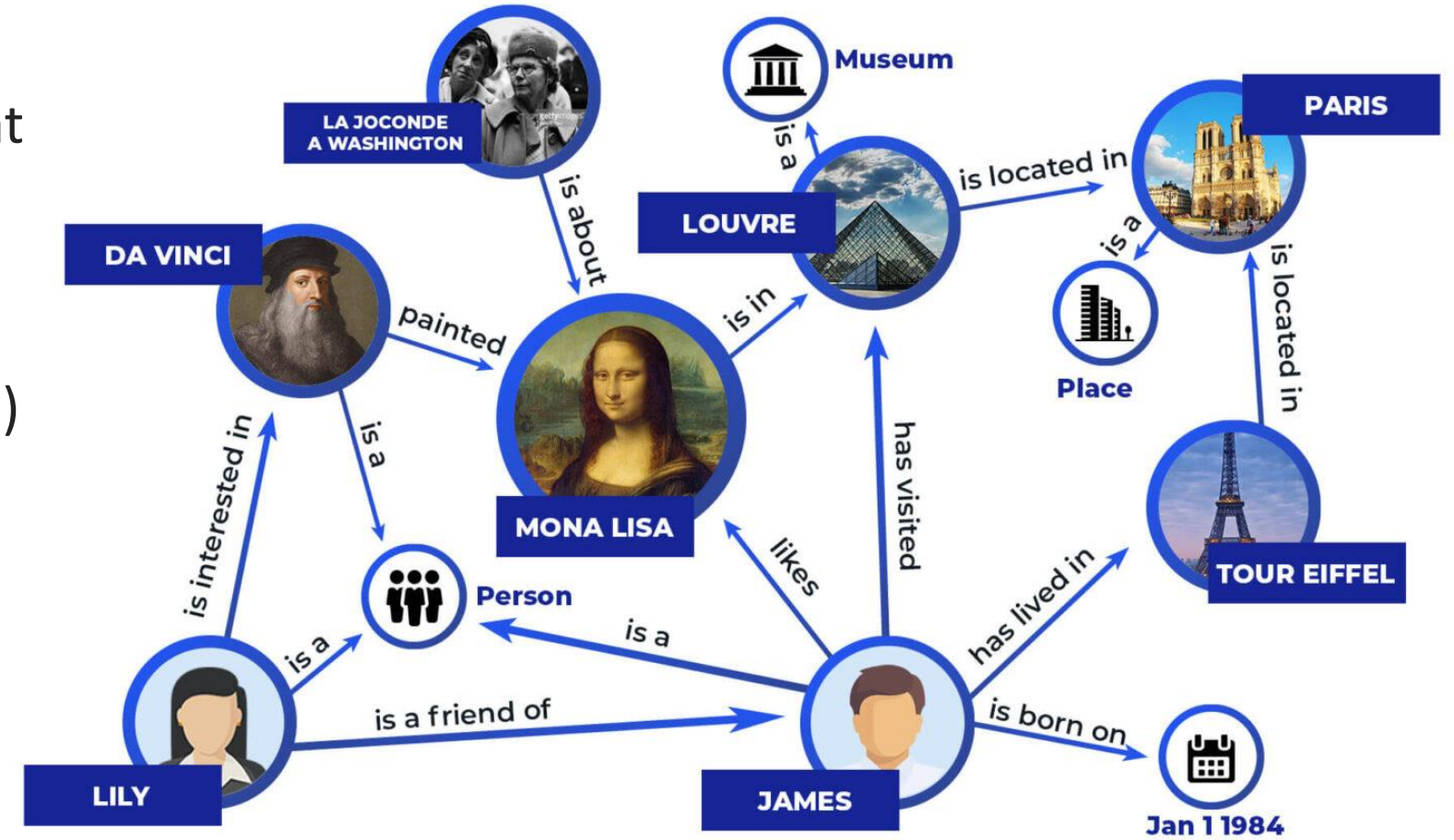
Topic Modeling (LDA, BERTopic): Identifies topics or themes within a collection of documents. Topics are distributions over words; each document is a distribution over these topics.



KNOWLEDGE GRAPHS

https://www.atulhost.com/what-is-knowledge-graph

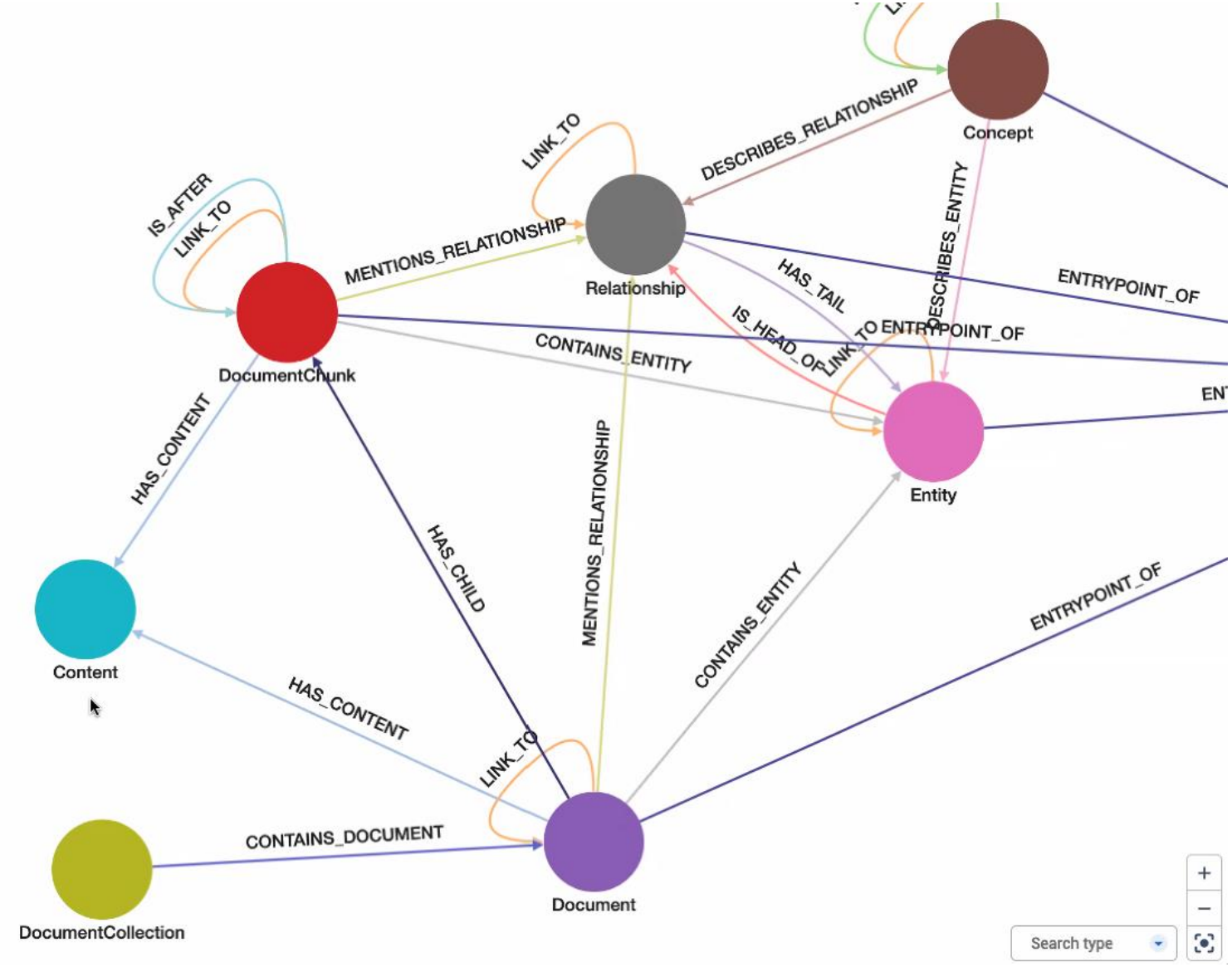
"Knowledge graphs are great at representing and storing heterogeneous and interconnected information (via nodes and relationships) in a structured manner, effortlessly capturing complex relationships and attributes across diverse data types. "



Connectivity results in rapid information gathering over multiple hops

https://medium.com/neo4j/enhancing-the-accuracy-of-rag-applications-with-knowledge-graphs-ad5e2ffab663

KNOWLEDGE GRAPH STRUCTURED DATA LAYOUT



Snapshot from TigerGraph demonstration ("GRAG" workflow)

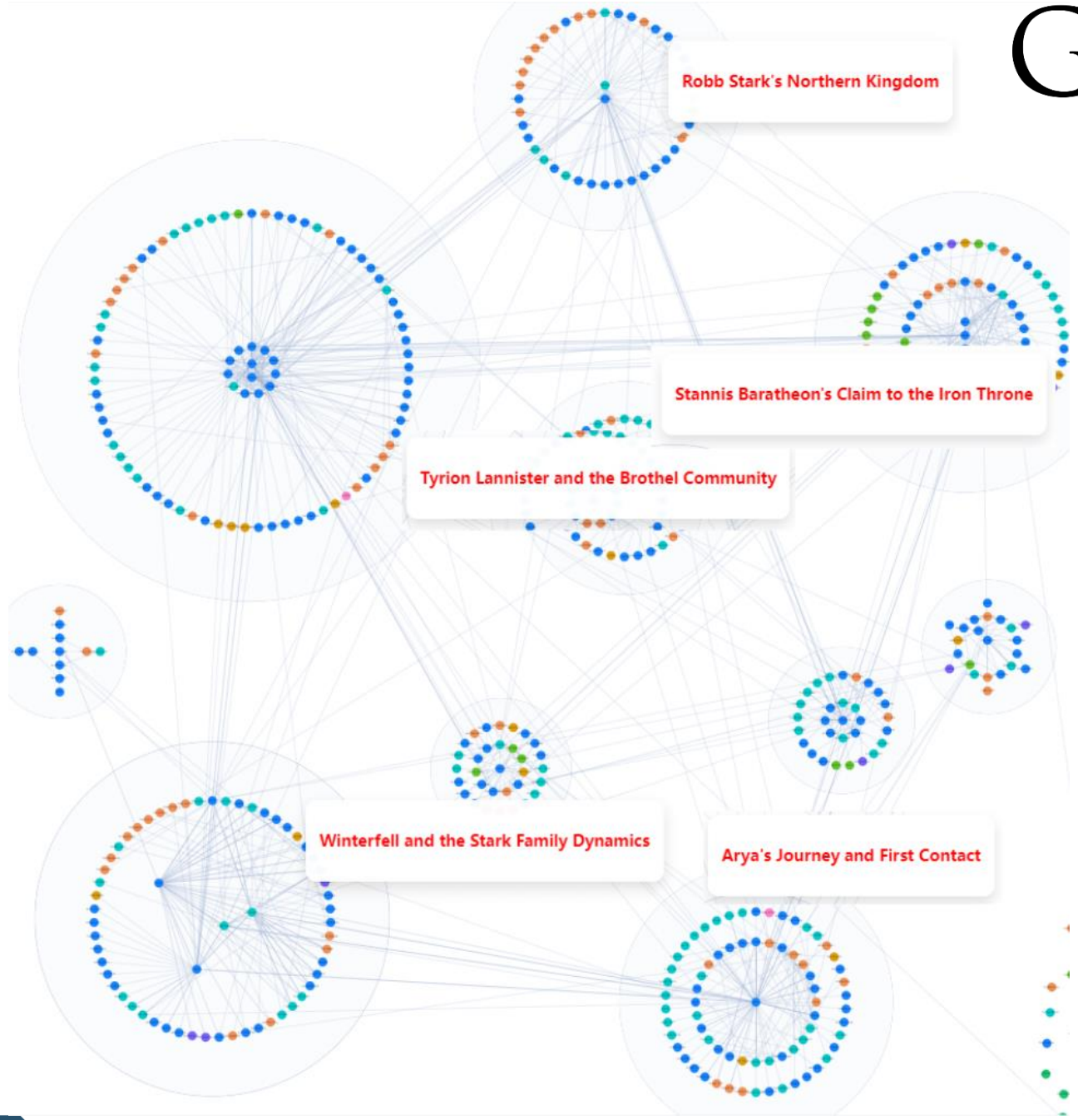
Knowledge Graph Content

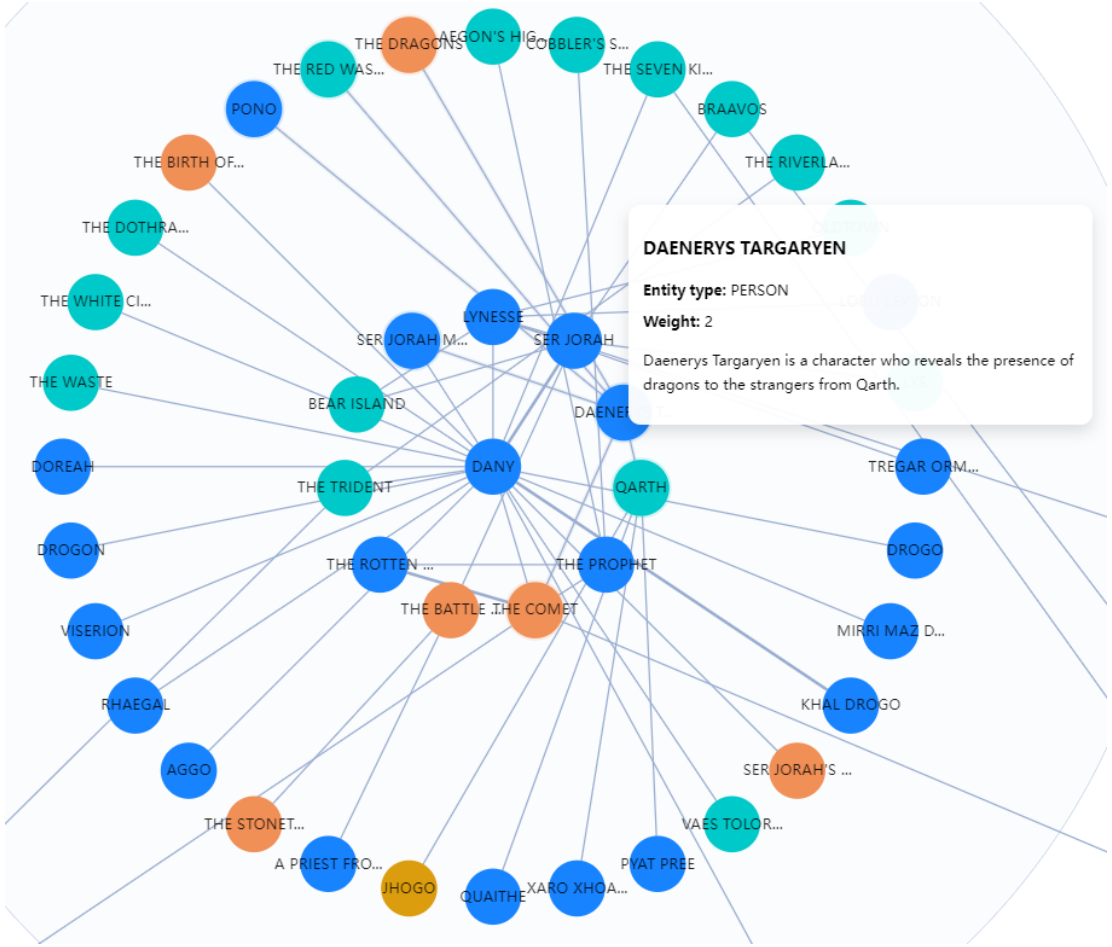
- Documents / chunks
- Relationships
- Topical concepts
- Bibliometric information
- Images (encoded)

"Knowledge Graphs Achieve Superior Reasoning versus Vector Search alone for Retrieval Augmentation"

https://ai.plainenglish.io/knowledge-graphs-achieve-superior-reasoning-versus-vector-search-for-retrieval-augmentation-ec0b37b12c49

GAME OF THRONES





https://medium.com/@infiniflowai/how-our-graphrag-reveals-the-hidden-relationships-of-jon-snow-and-the-mother-of-dragons-bd89084f64ec

GRAG VS 'RAG'

What's the relationship between Daenerys and Jon Snow?

GraphRAG

The relationship between Daenerys Targaryen and Jon Snow is complex, multifaceted, primarily rooted in their shared lineage and their roles within their respective communities. Here are the key points from the knowledge base that highlight their relationship

- **Shared Lineage**: Both Daenerys Targaryen and Jon Snow are members of House Targaryen. Jon Snow, whose true name is Aegon Targaryen, is the song of Rhaegar Targaryen and Lyanna Stark, making him Daenarys's nephew. This familiar connection adds a layer of personal and historical significance to their interactions.
- **Leadership Roles**: Both characters hold significant leadership roles within their respective communities:

…….. *truncated*

RAG*

The relationship between Daenerys Targaryen & Jon Snow involves a number of mutual interactions

- Speaking with each other
- Flying on dragons
- Traveling on boats
- Fighting the undead
- Manipulating one another
- Spending quality alone time
- Crying
- Fighting each other

GRAG is a superior approach when navigating through complex linkages of information

https://medium.com/@infiniflowai/how-our-graphrag-reveals-the-hidden-relationships-of-jon-snow-and-the-mother-of-dragons-bd89084f64ec

*GuEstimation of RAG-like responses

LLM FINE TUNING

https://www.youtube.com/watch?v=ahnGLM-RC1Y)

Context optimization

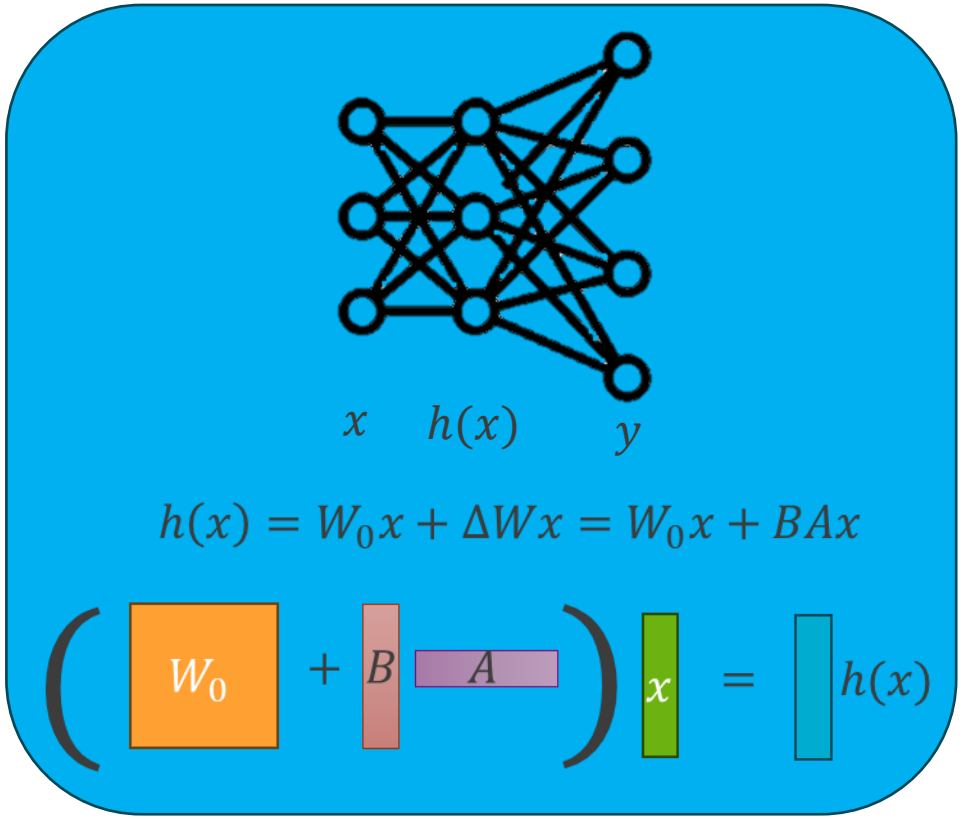
RAG+

RAG + Expert Modules

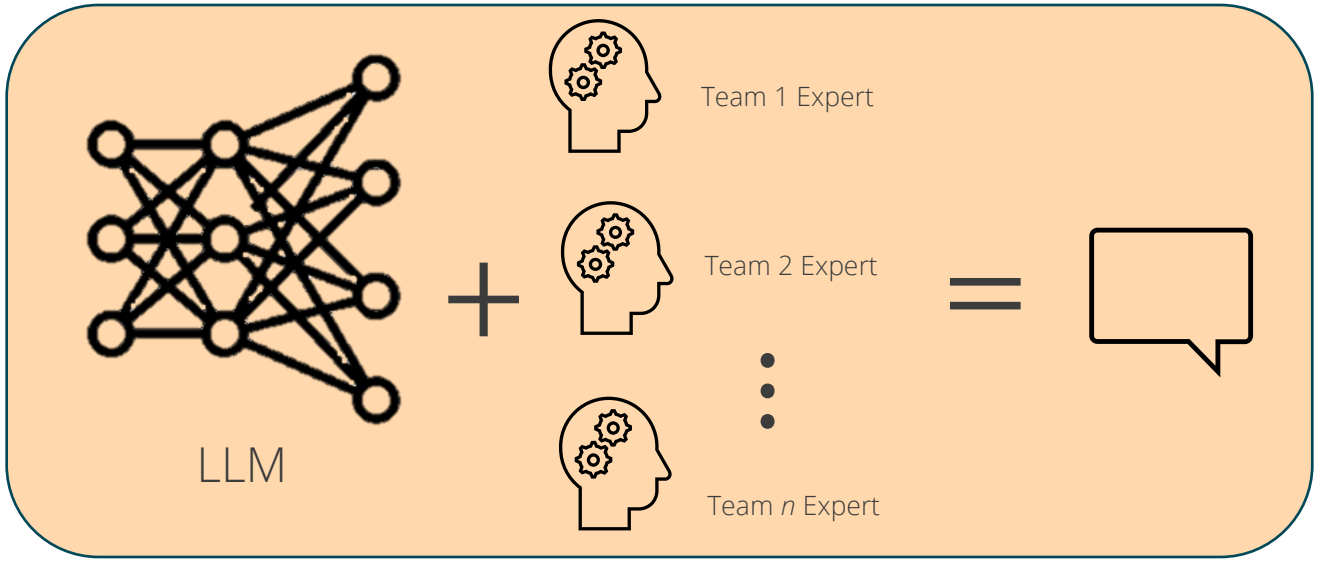
Prompt engineering

Fine-tuning

LLM optimization

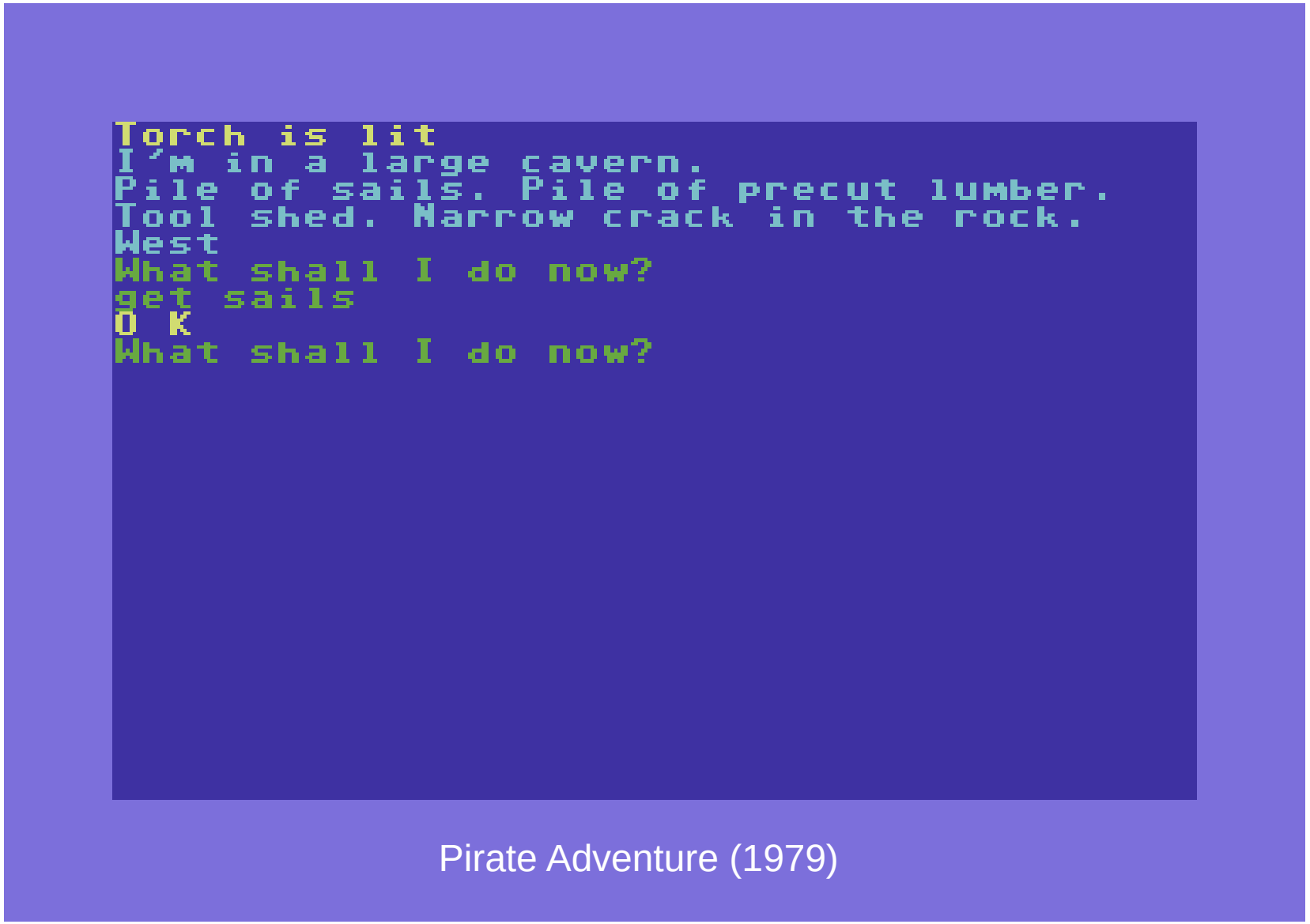


Create an ensemble of small 'expert modules' using parameter efficient fine-tuning (PEFT) – **Poster (Alexander Nemecek)**



HUMAN-CENTRIC UI DESIGN

MUST WE REGRESS TO THE 1970'S?



Pirate Adventure (1979)

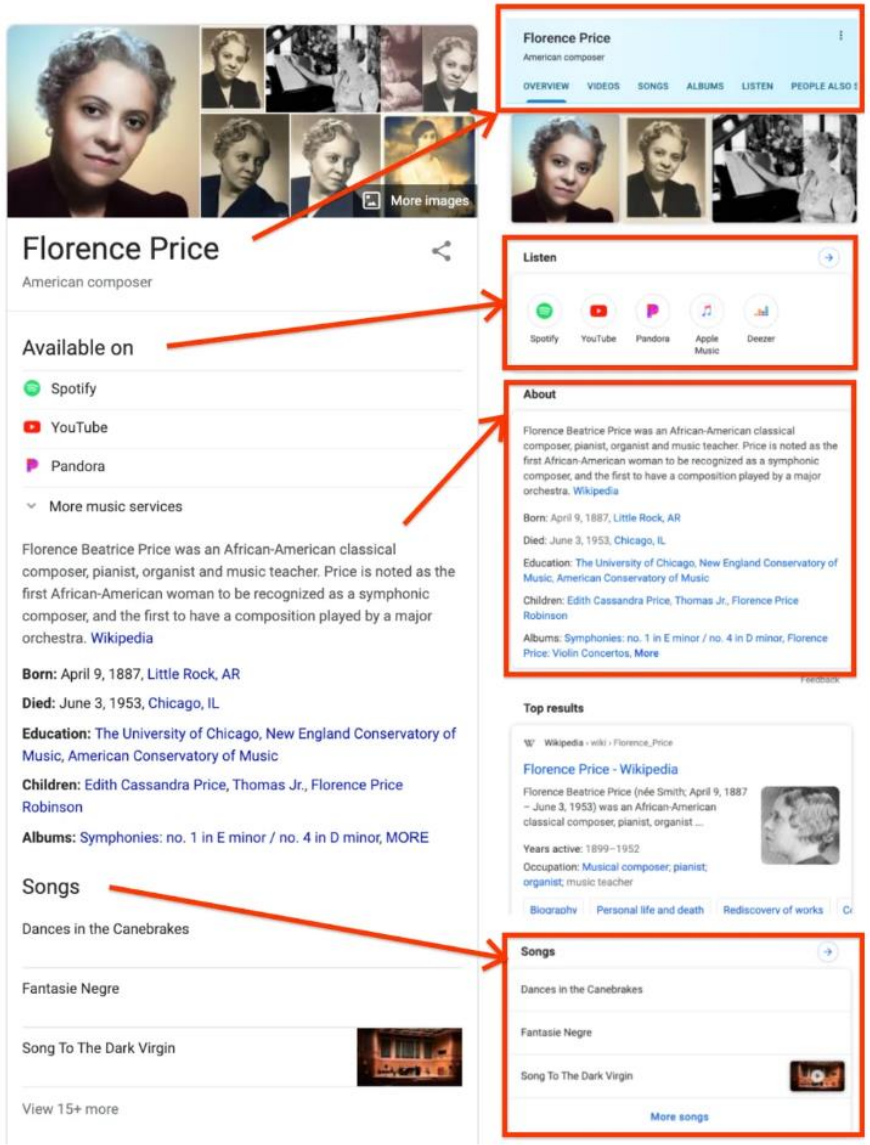
AVAILABLE DATATYPES

What are electrically exploding wires used for?

Information should we could display based on a user query

- Single paragraph response
- Summary of the topic (do you want to learn more?)
- RAG references used to generate the response
- Images from the topic
- List of related authors
- Topical model
- Topical visuals
- Related references/Historical information

INTERESTING UI/UX DESIGNS



Google Knowledge Panel

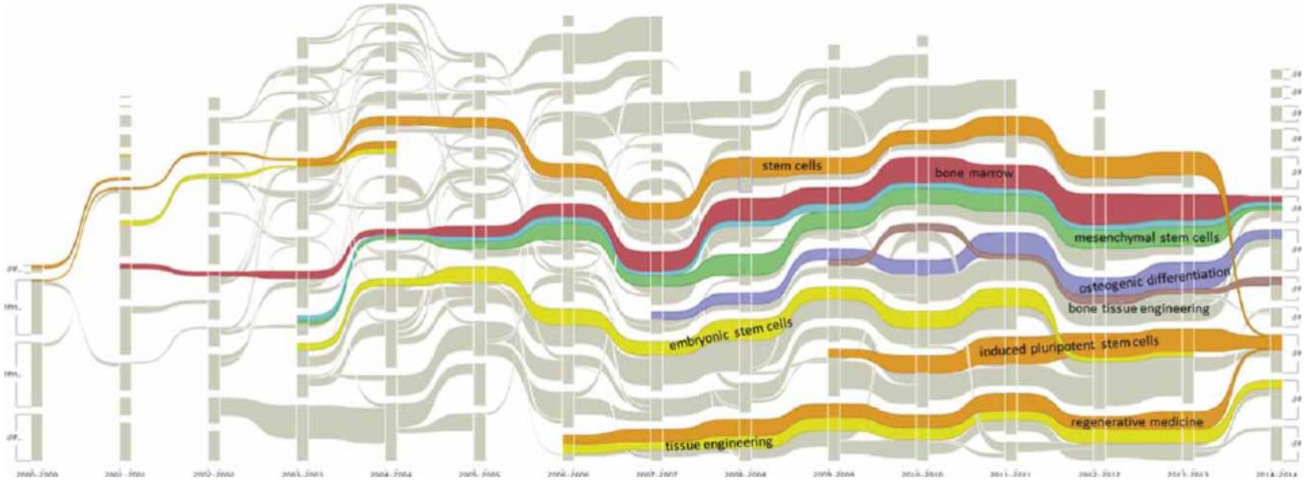


Figure 11. Alluvial flow of terms is shown. Each year top 100 noun phrases formed a Pathfinder network.

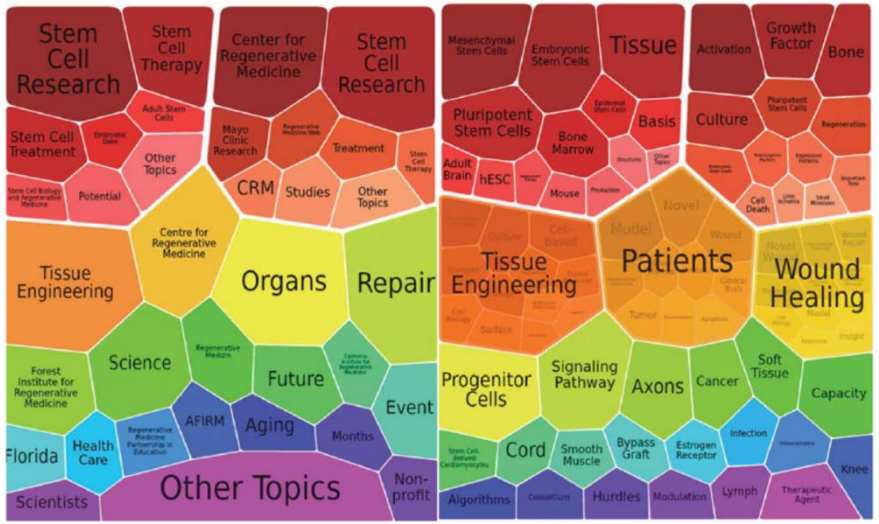
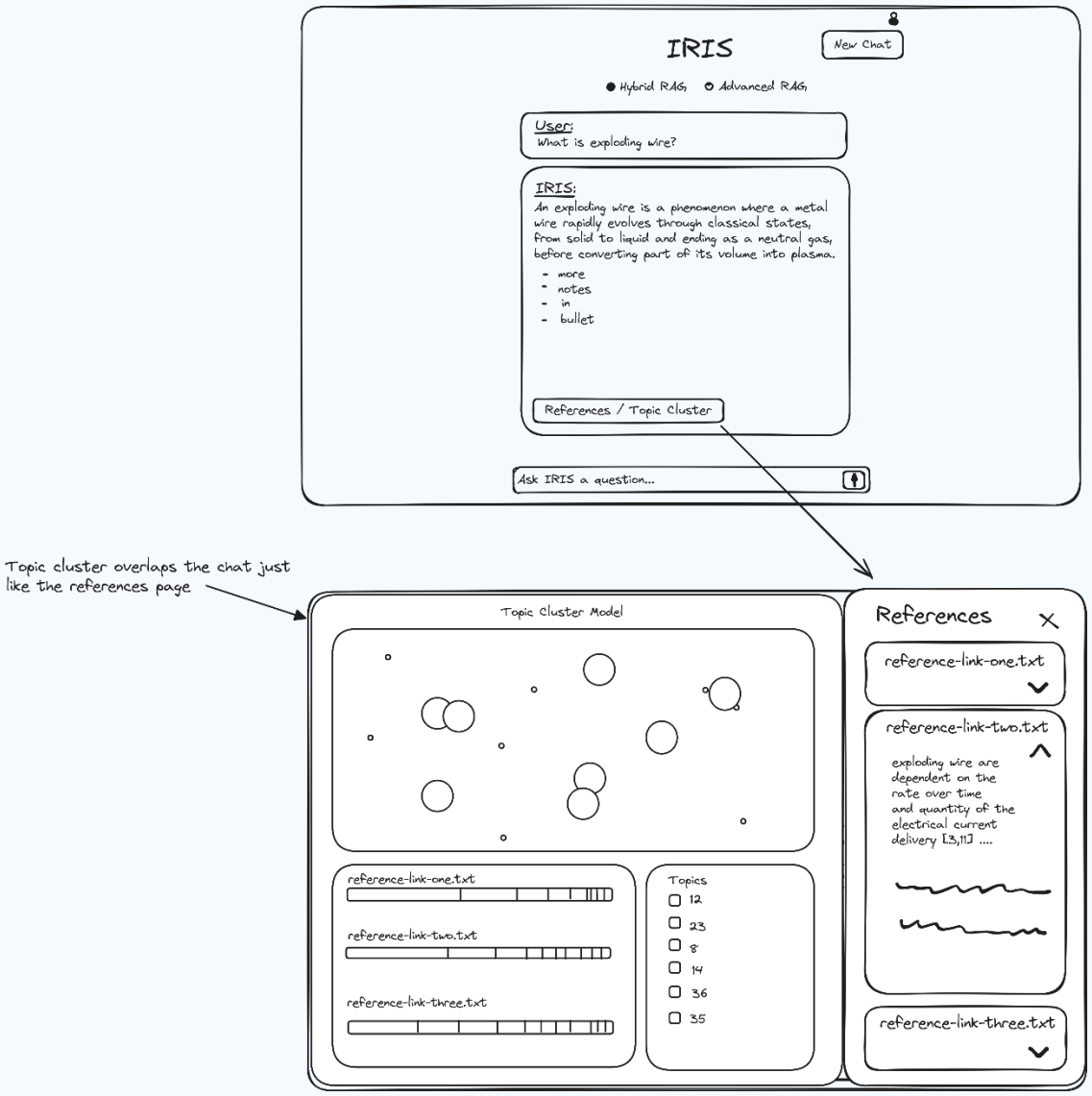


Figure 1. A lightweight survey of major topics on the Internet on regenerative medicine is shown. The visualizations were generated by the Carrot system based on first 100 results of search on regenerative medicine. Left: search results from the web. Right: search results from the PubMed.

IRIS 'KNOWLEDGE PANEL' WIREFRAMING (TRADITIONAL)



Example of 'traditional' wireframing exercise, taking an already existing capability and attempting to expand text to multi-visual output

- Sometimes tedious
- Time-consuming
- 'Slow' design iterations
- Eventually you have to write the code
- 'Rinse & Repeat' many times

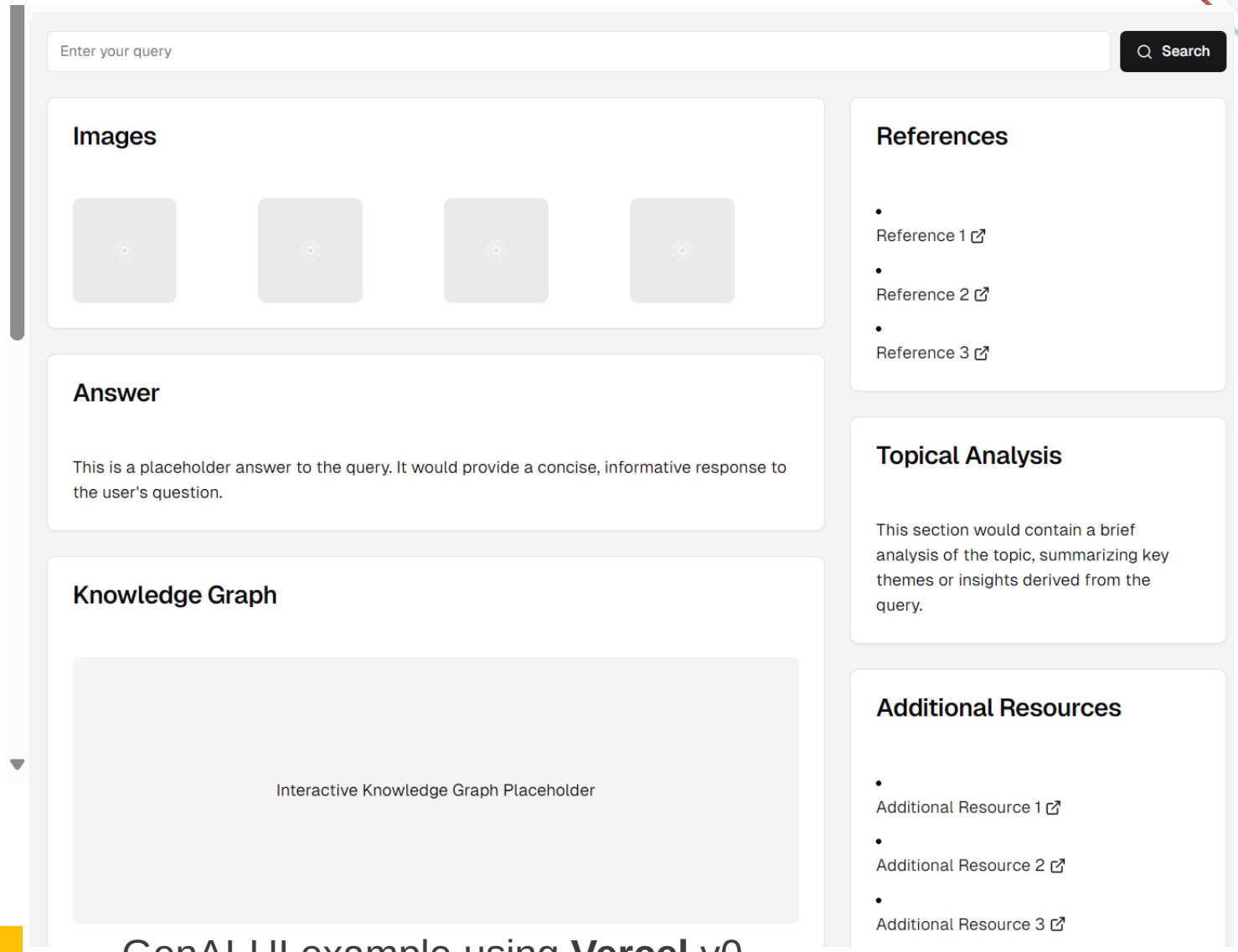
IRIS ‘KNOWLEDGE PANEL’ WIREFRAMING (GENAI – UI)

"Design a layout for a knowledge panel that will display information retrieved based on user queries. The panel should include the following components:

1.Query Input Area: A clear and accessible space for users to enter their queries.
2.Images Section: A visually appealing area to showcase relevant images associated with the query. This section should allow for multiple images to be displayed in a grid format.
3.Text-Based Answer: A prominent section that provides a concise, informative answer to the user's query. This should be easy to read and highlight key points.
4.References Section: A list of references used to generate the response, formatted for clarity and accessibility. This could include links to articles, studies, or other resources.
5.Knowledge Graph: An interactive visual representation of interlinked related information, allowing users to explore connections between concepts.
6.Topical Analysis: A brief analysis of the topic, summarizing key themes or insights derived from the query.
7.Additional References: A section for supplementary resources or further reading, providing users with more options to explore the topic.

Please ensure the layout is user-friendly, visually appealing, and responsive across different devices. Consider the use of colors, fonts, and spacing to enhance readability and engagement."

A LLM-based workflow can generate the UI



GenAI-UI example using **Vercel** v0

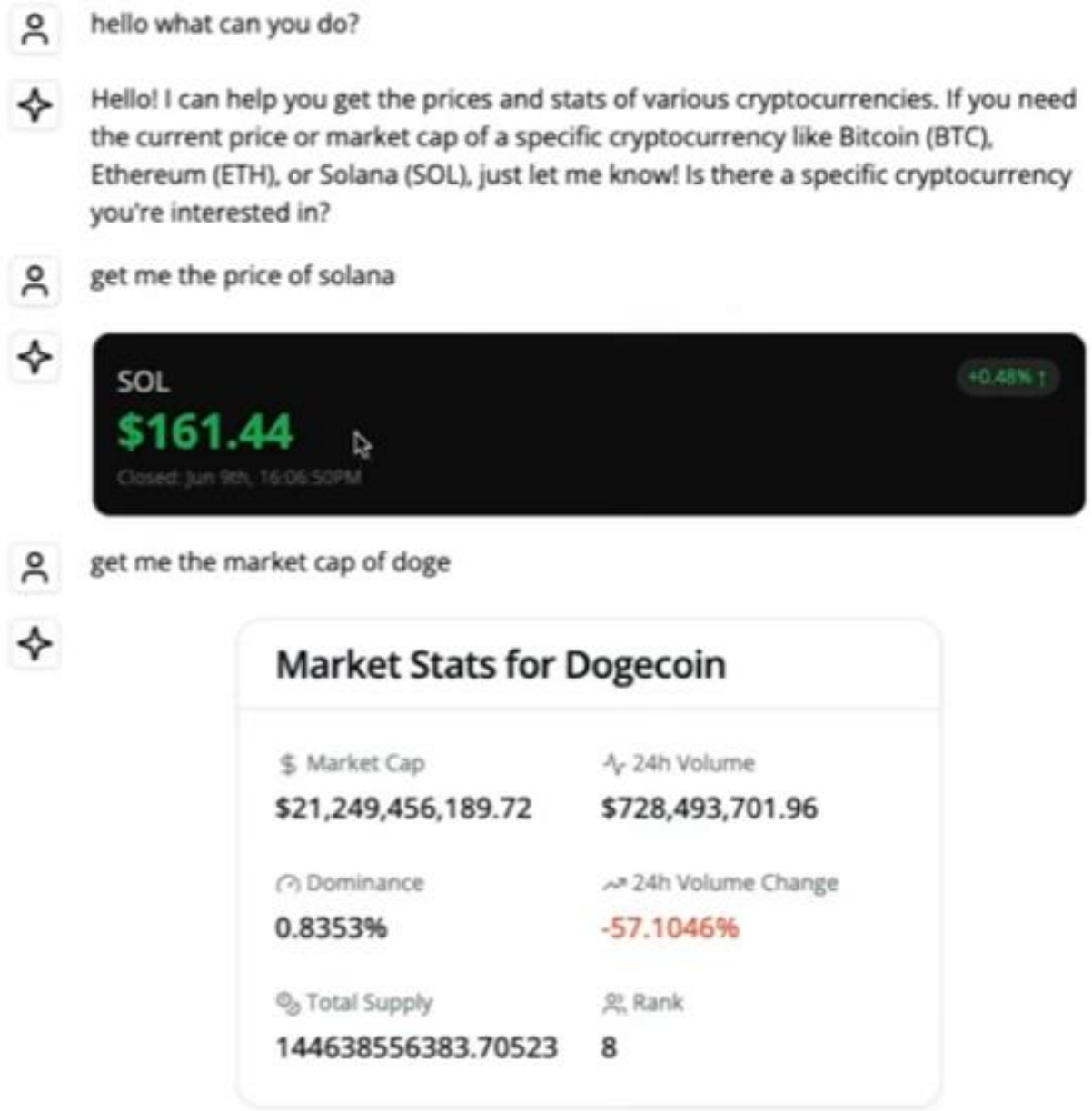
GENAI UI FUNCTION CALLING

The user will generate functions to provide responses to different types of user input.

Function calling: The LLM will review the user prompt and then decide which function to execute, resulting in different user experiences and workflows. Combining this approach with functions/visuals from Vercel v0 results in dynamically varying output based on user queries.

The LLM can also be used as part of the workflow to provide further analysis of images, text, etc.

A LLM can dynamically change the visual output based on the query history



https://www.youtube.com/watch?v=UIMG1kFiWa4

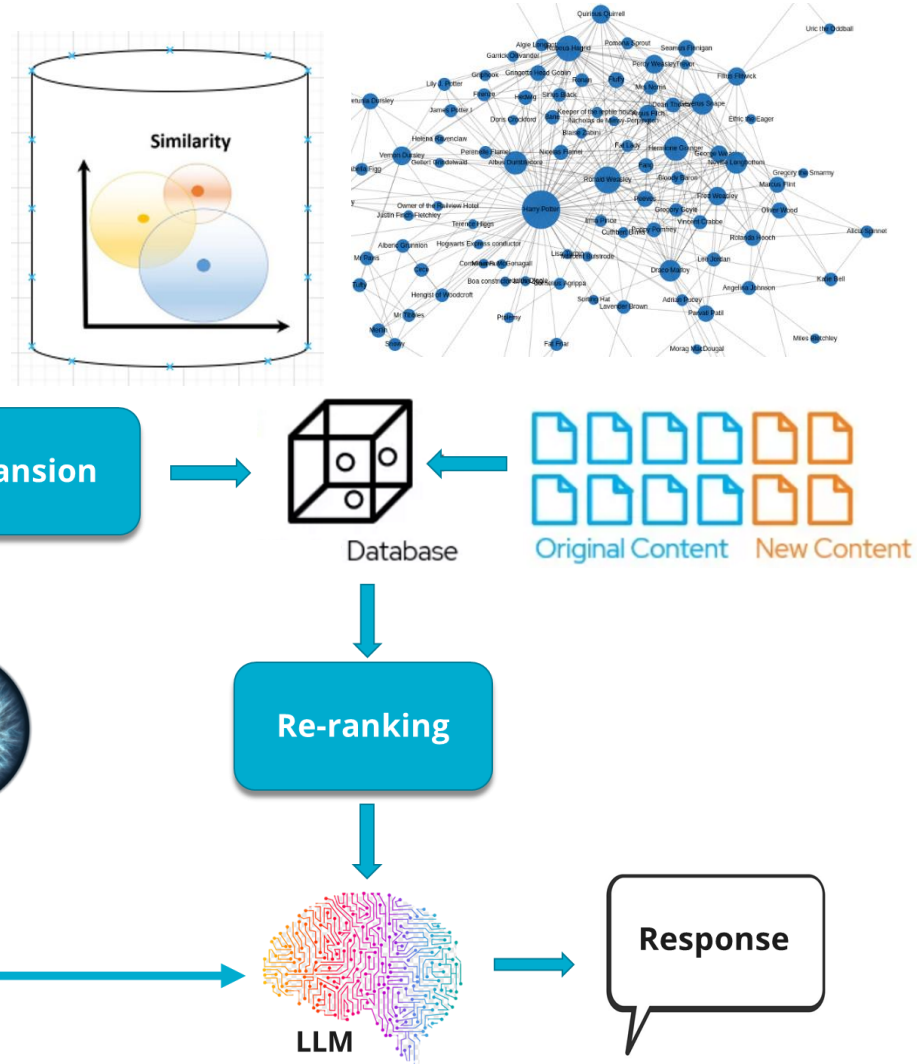
TRUST & RELIABILITY

HOW DO WE ESTABLISH TRUST & VERIFIABILITY?

Prospective users are very concerned about **trust**

- GRAG Critical workflow elements
  - Chunking approach
  - Embedding model
  - Agents
  - LLM
    - Multi-modal extraction
    - Agentic chunking
    - Creates the knowledge graphs
    - LLM used to assemble the final information
    - LLM tuned to create an embedding model
    - Function calling
  - ‘Guardrails’
  - Knowledge Panel (UI/UX)

We must establish trust across the entire GRAG ecosystem



HOW CAN WE TRUST A LLM?

Caveat emptor

Ensuring Trustworthiness in GRAG & LLMs

- Regular benchmarking against established metrics
- Awareness of potential biases in training data
- Human review in critical decision-making processes
- Expert validation to ensure accuracy
- Cross-referencing LLM outputs with trusted databased and literature
- Defining clear accountability for decisions based on LLM/RAG outputs
- Establishing guidelines for responsible use of LLMs in research and operational settings
- Educating users on the limitations and appropriate use cases
- RAG results must be interpretable & understandable

Having LLMs verifying other LLMs is akin to saying ‘it’s LLMs all the way down’



FUTURE DIRECTIONS

How do we capture and retrieve information which took more than 75 years to create?

How do we rapidly onboard future generations using information which took decades to create?

How can we take complex scientific information and distill it into digestible pieces?



Unlocking Decades of Wisdom: Distilling Complex Information for Next-Gen Talent

IRIS FY25 RESEARCH AREAS

- Definition of a **virtual SME**
  - Types of questions, reasoning & research requests
- RAG Workflow
  - Reference research and extraction workflow
  - PDF2Text/Multi-modal strategy assessments
  - Chunking strategy assessments
  - Embedding model assessments
  - LLM deployment strategies & assessments
  - Knowledge graph creation via LLM
  - **Graphical RAG** prototyping & deployment
- **Topical analysis** investigation & visualization
- **UI/UX elements**
  - Knowledge panel wireframing/deployment
  - GenAI – UI Investigation
  - User feedback
- **Fine-tuning** investigation ('expert' modules)



IRIS Deployment

Overarching Goal: Create a virtual SME with 'knowledge' based on corpus of documents

IN THE NOT-TOO-DISTANT FUTURE

AN INFORMATION RETRIEVAL TRANSFORMATION WILL REQUIRE

LLMs

- ‘trawling’ through extensive archives of information, extracting text, capturing topical information and relationships, including citations
- creating knowledge graphs
- executing a RAG + KG workflow

An enterprise-wide RAG + KG deployment

Altering data management at Sandia

We’re going to need a **LOT OF GPUs** (‘a bigger boat’)



LLMs ‘trawling’ for documents, data, figures

A FUTURE VISION & FOOD FOR THOUGHT

Can GenAI become a means for ?

- Technical Onboarding & Hands-on Training
- Mentorship & Support
- Continuous Education
- Skill Development
- Progress Reviews
- Useful Performance Feedback & Goal Setting



From <https://ai.plainenglish.io/enriching-language-models-with-knowledge-graphs-for-powerful-question-answering-3c08b3b8020f>

REFERENCES

- GraphRAG, https://github.com/microsoft/graphrag
- GoT knowledge graph: https://medium.com/@infiniflowai/how-our-graphrag-reveals-the-hidden-relationships-of-jon-snow-and-the-mother-of-dragons-bd89084f64ec
- Multi-hop Q&A: https://ambersearch.de/what-is-multi-hop-qa/
- Multi-hop Q&A: https://www.moveworks.com/us/en/resources/ai-terms-glossary/multi-hop-reasoning
- Vercel AI SDK Video: https://www.youtube.com/watch?v=UDm-hvwpzBI (build a chatbot in 40 lines of code)
- Alluvial diagram: Emerging Trends and New Developments in Regenerative Medicine: A Scientometric Update (2000-2014)
- Carrot system image: Emerging Trends and New Developments in Regenerative Medicine: A Scientometric Update (2000-2014)
- GenAI UI development using Vercel v0 video: https://www.youtube.com/watch?v=61st-JJSvYA
- Vercel v0, Build an AI application in 5 minutes: https://www.youtube.com/watch?v=-mDcL314lFI
- Vercel SDK with function calling: https://www.youtube.com/watch?v=UIMG1kFiWa4
- LLM function calling: https://www.youtube.com/watch?v=i0tsVzRbsNU