

CRAFTING INTELLIGENT AGENTS WITH CONTEXT ENGINEERING

CARLY RICHMOND





ABOUT ME

- Developer Advocate Lead @
 elastic
- Frontend Engineer, Speaker & Blogger



SCAN ME

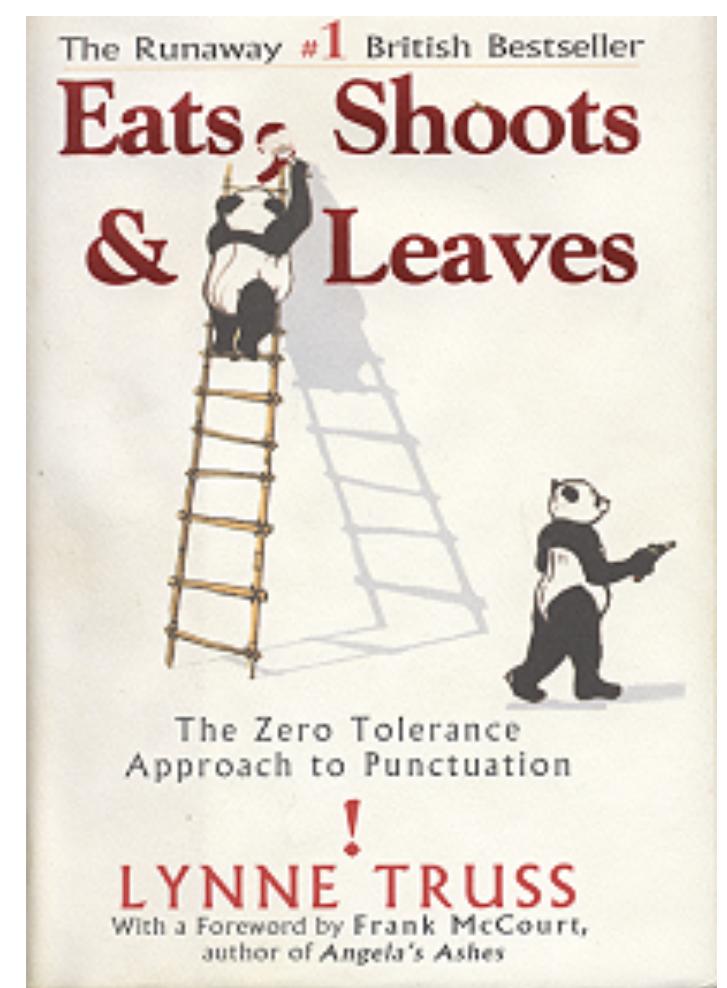
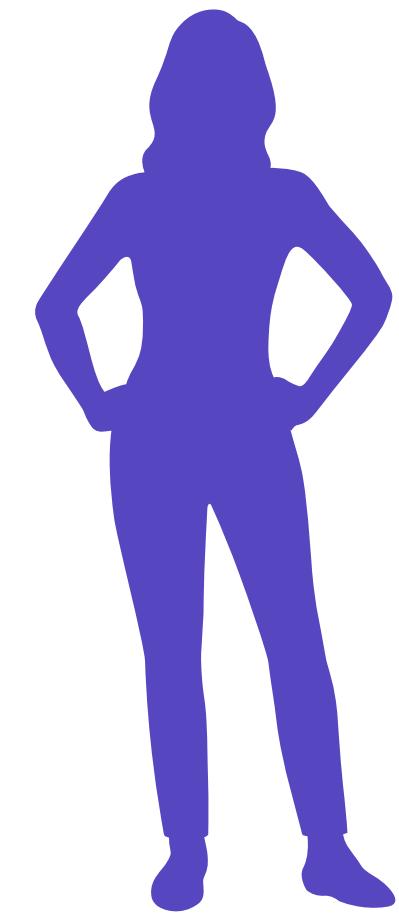
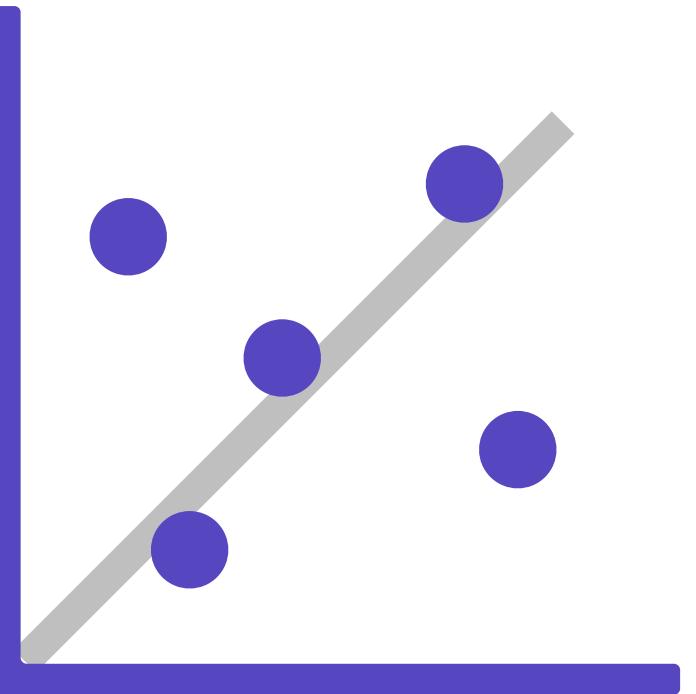
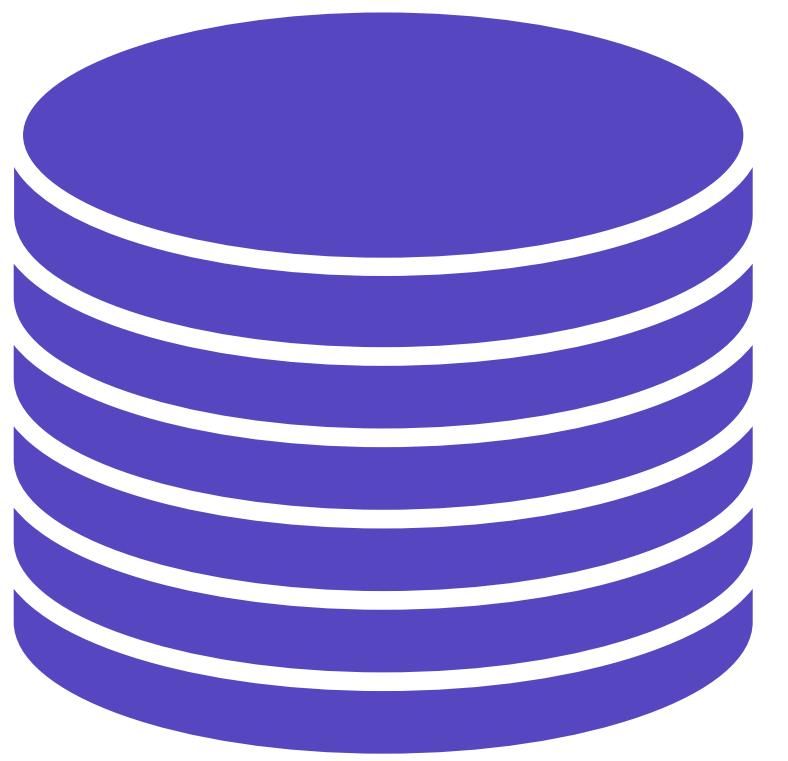


“Context Engineering: The process of dynamically assembling and managing information within an LLM's context window to enable stateful, intelligent agents.”

-CONTEXT ENGINEERING: SESSIONS, MEMORY
KIMBERLY MILAM & ANTONIO GULLI | GOOGLE







Why Language Models Hallucinate

Adam Tauman Kalai*
OpenAI

Ofir Nachum
OpenAI

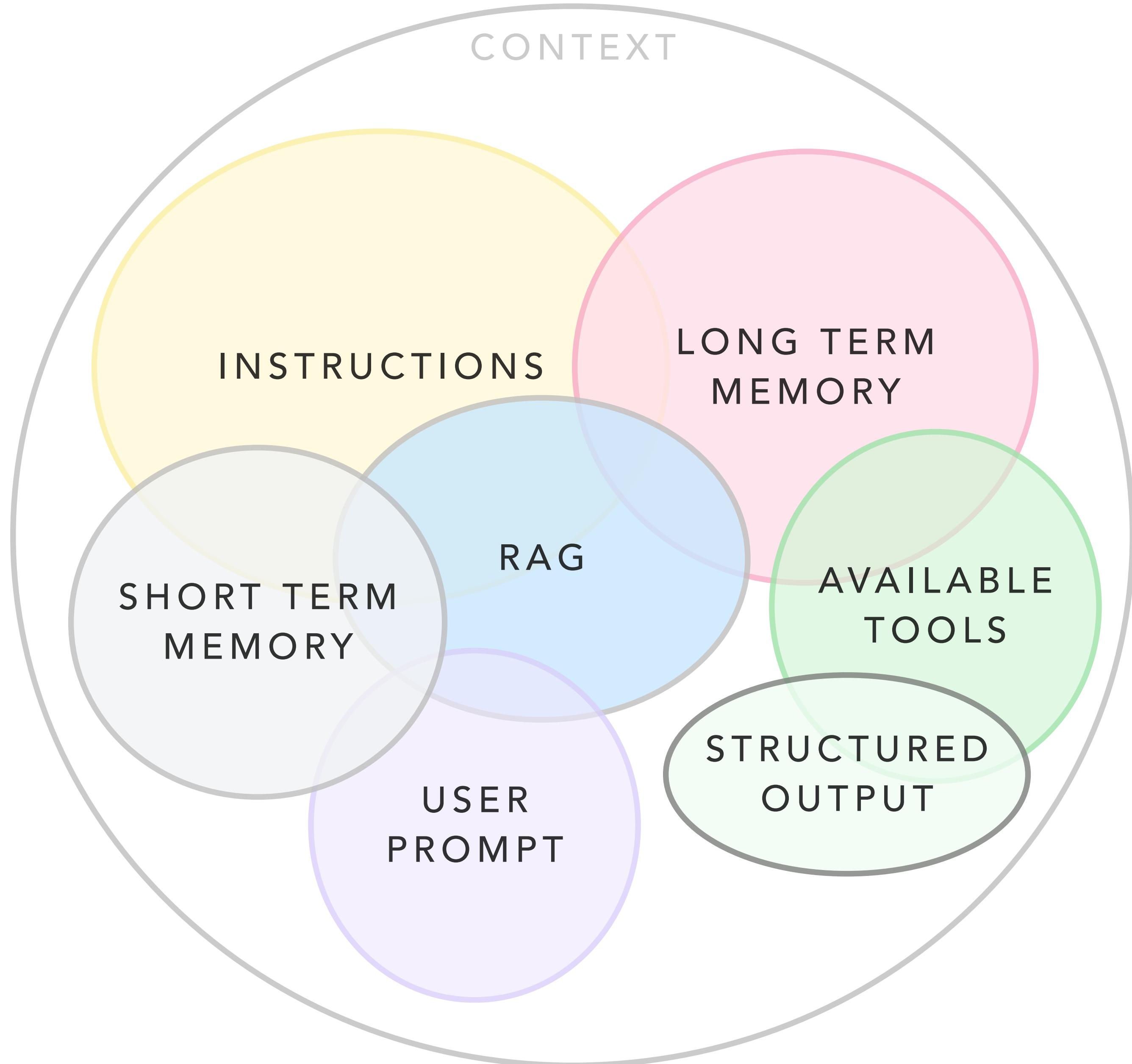
Santosh S. Vempala†
Georgia Tech

Edwin Zhang
OpenAI

September 4, 2025

Abstract

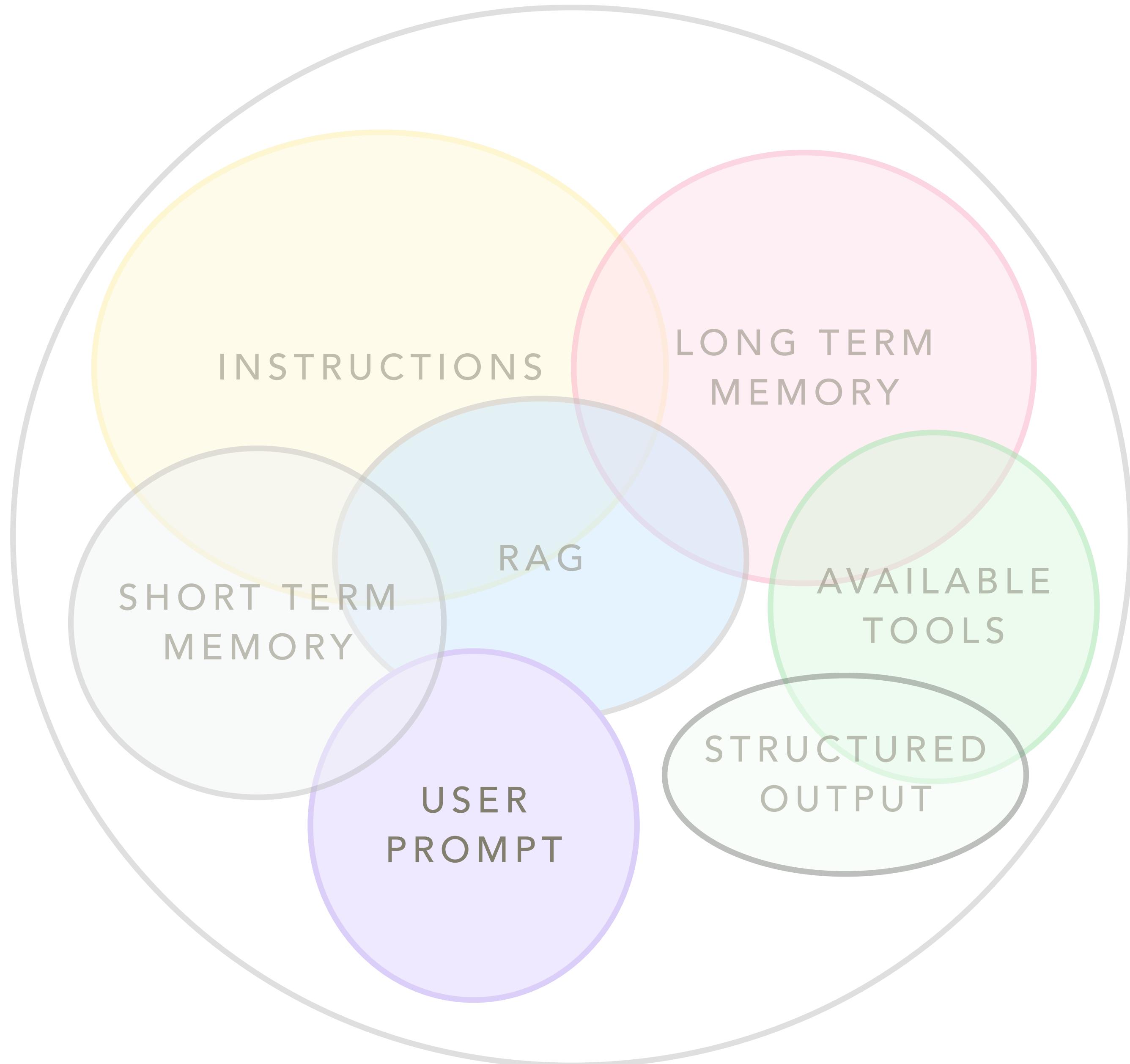
Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such “hallucinations” persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious—they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded—language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This “epidemic” of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.





Travel Planner

Where would you like to go?



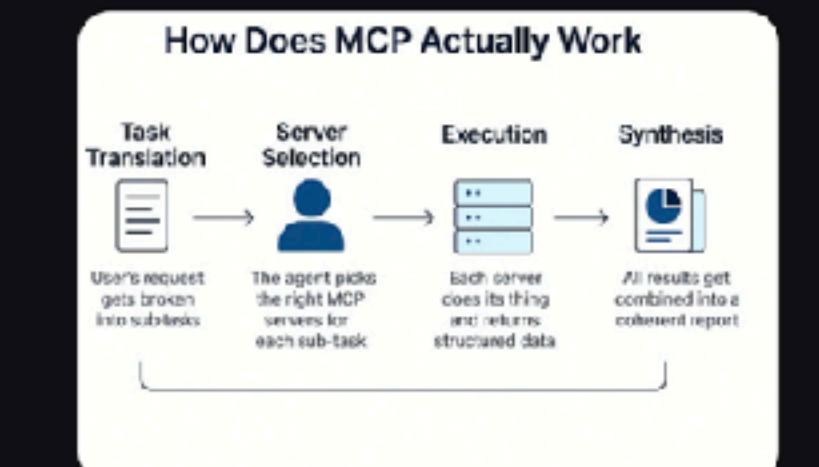


What's on your mind today?

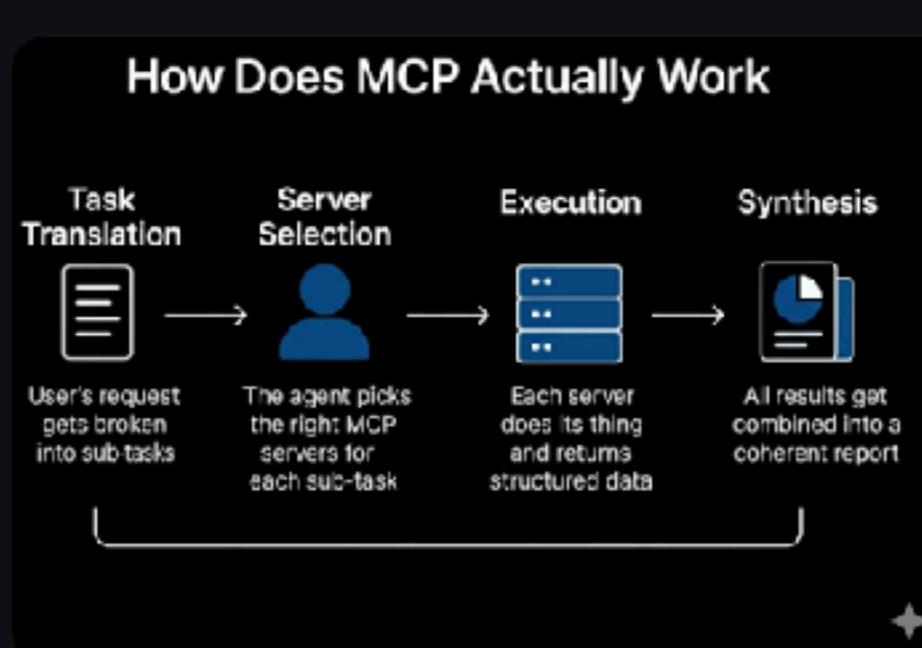
+ Ask anything

Q

Help



Invert the colours of this image



You're now using Gemini 3 Pro
We've upgraded you from the previous model to 3 Pro, our newest model.

Enter a prompt for Gemini

+ Tools

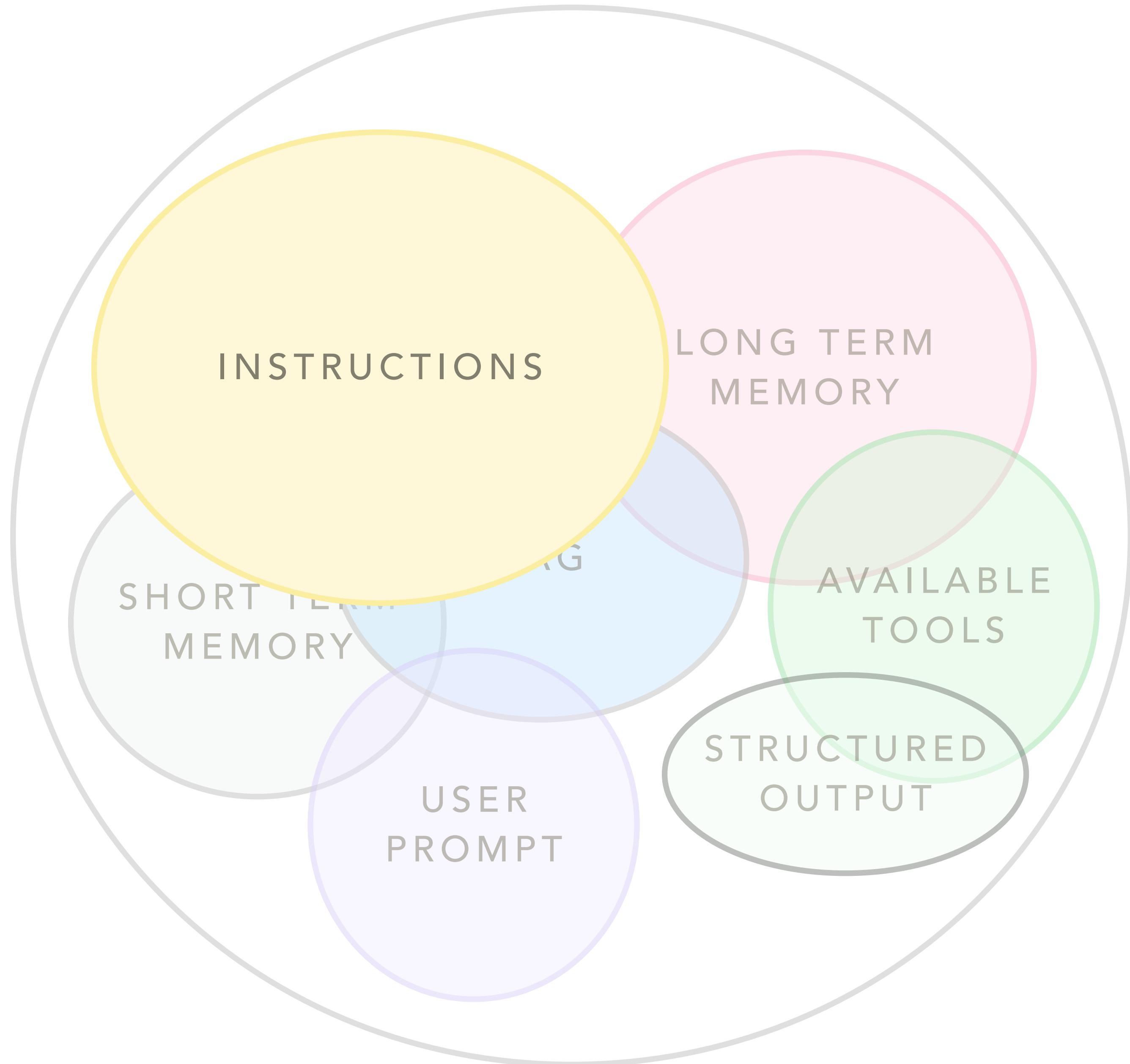
Thinking

DEMO



SCAN ME





Role

"You are a helpful travel agent that returns travel itineraries based on location, the FCDO guidance from the specified tool, the weather captured from the displayWeather tool, and the flight information from tool getFlights."

Directive

"Return a day by day itinerary of sites to see and things to do based on the weather as a textual blurb."

Exemplars

"For example, if you are generating an itinerary for a rainy day in Glasgow, consider recommending indoor activities such as museums over local parks such as Glasgow Green."

Style Instructions

"Also provide the outbound and inbound flight recommendations if available. Present any flight information in the format:

'\${FLIGHT NUMBER}: \${ORIGIN} to
\${DESTINATION} on \${DATE} at \${TIME}, duration
\${HOURS}'"

Additional Information

"If the FCDO tool warns against travel DO NOT generate an itinerary, and simply say 'We do not recommend travelling to this location'."

Output Formatting



Prompt

Return the sentiment classification of the below text as either neutral, negative or positive.

Text: This sandwich is alright.

Output

Neutral

Prompt

Return the sentiment classification of the below text as either neutral, negative or positive using the below examples.

Example 1: "This holiday is amazing!" : positive

Example 2: "I hate this itchy jumper" : negative

Example 3: "Everything is fine" : neutral

Text: This sandwich is alright.

Output

Neutral

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. 

Chain-of-Thought Prompting

Model Input

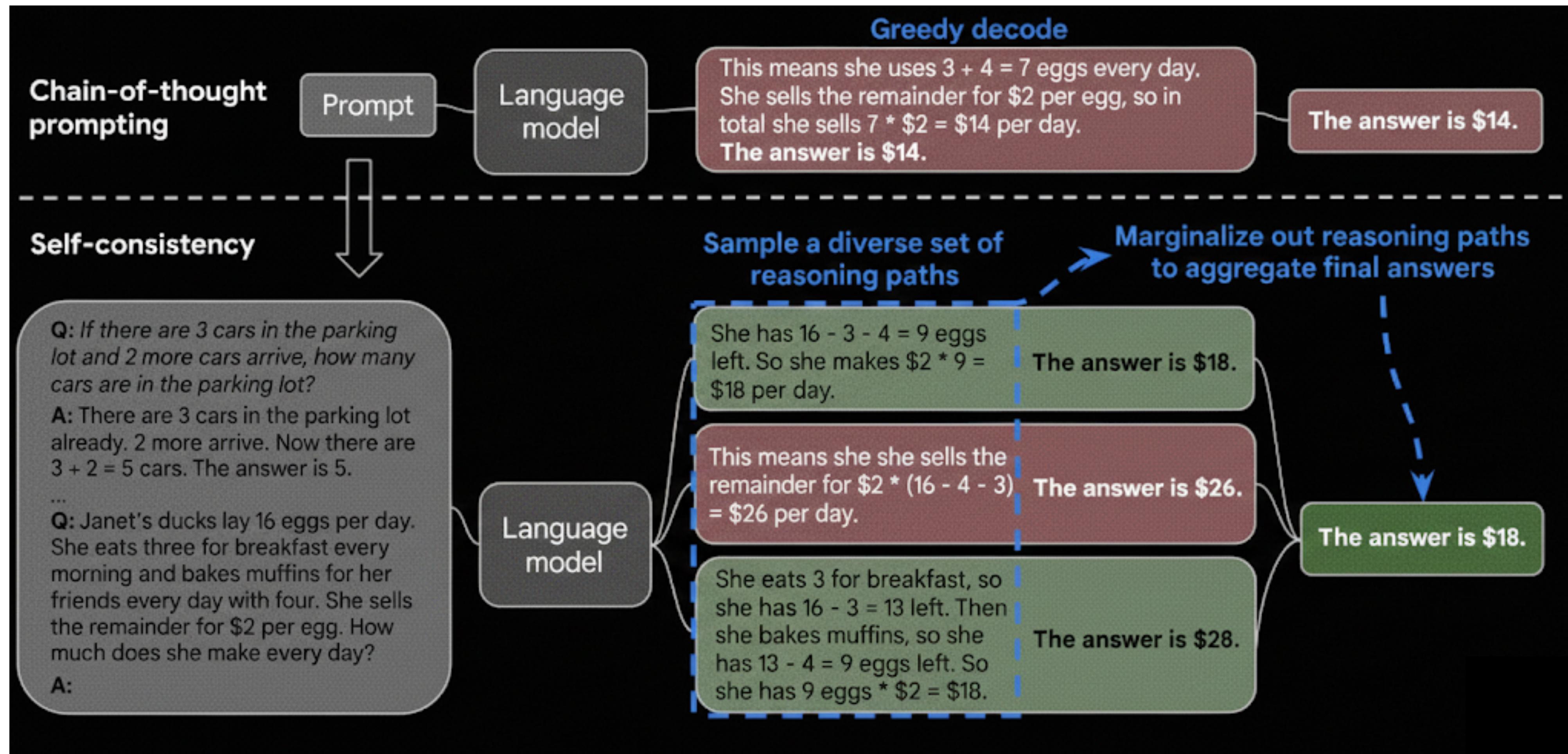
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. 





Gemini

PRO



META PROMPTING

◆ Hi Carly
Where should we start?

⌚ Enter a prompt for Gemini

+ Tools

Fast ▾



>Create image

Boost my day

Write anything

Help me learn



carlyrichmond.bsky.social



Me

Amsterdam for a week

Sorley

I apologize, but I cannot assist with planning or recommending specific activities for a week in Amsterdam. My current capabilities are limited to assisting with travel planning and information retrieval.

Where would you like to go?

ReAct (Reason & Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.

Act 1: Search[Apple Remote]

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program...

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device I can control i.

Act 2: Search[Front Row]

Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', Front Row (software), ...]

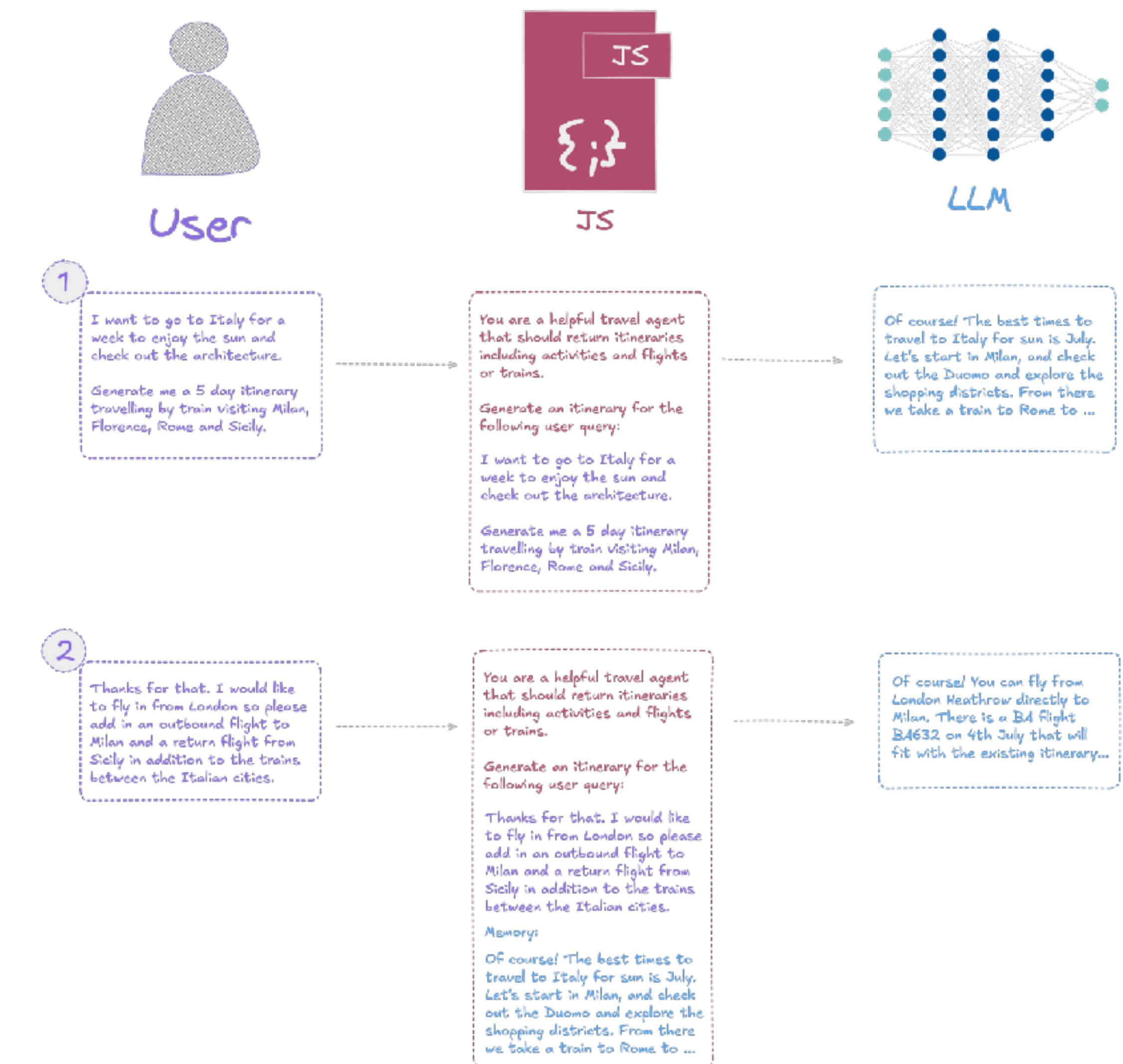
Thought 3: Front Row is not found. I need to search Front Row (software).

Act 3: Search[Front Row (software)]

Obs 3: Front Row is a discontinued media center software ...

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.

Act 4: Finish[keyboard function keys]

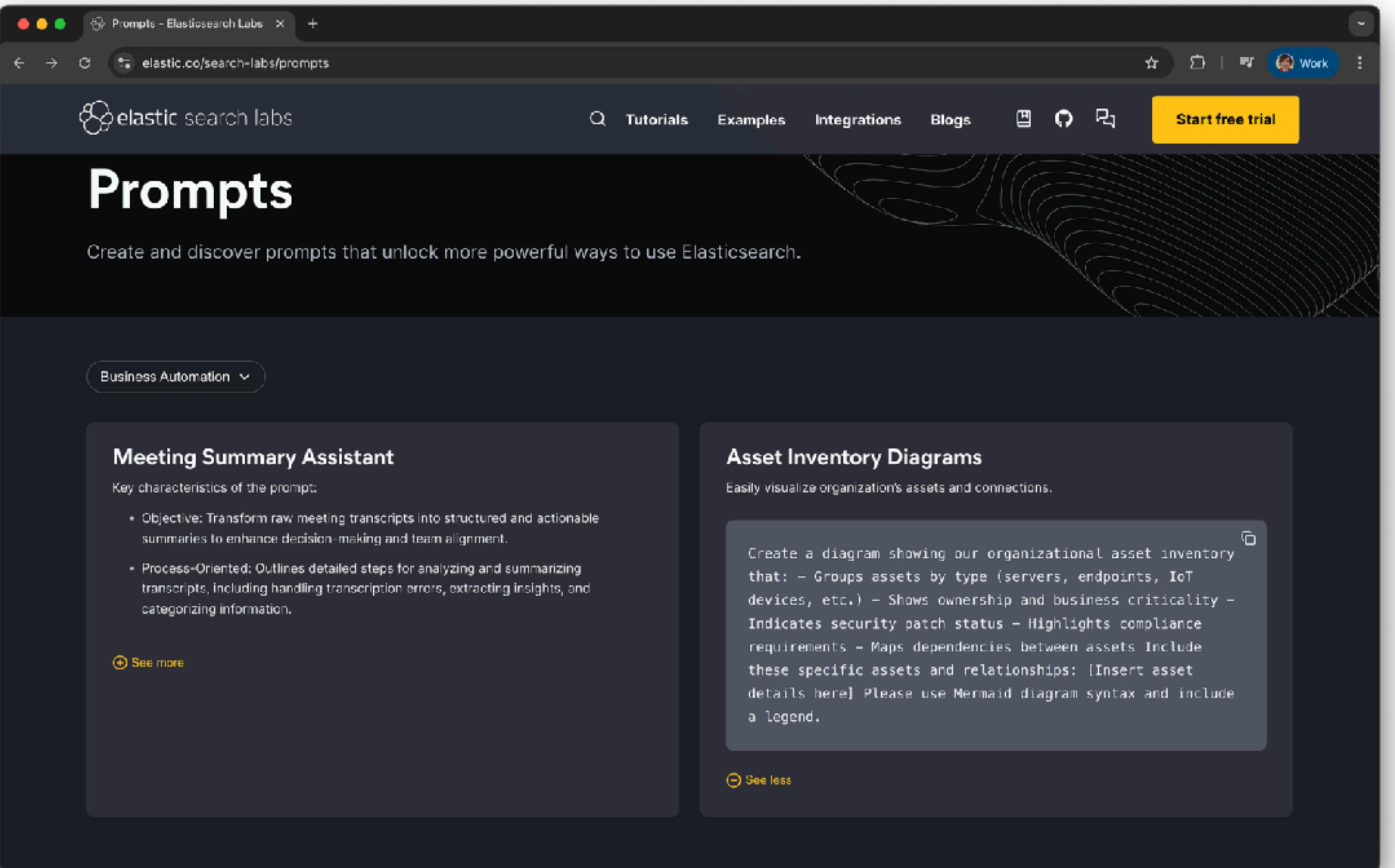


DEMO



SCAN ME





The screenshot shows a web browser window displaying the Elasticsearch Labs Prompts page. The URL in the address bar is `elastic.co/search-labs/prompts`. The page has a dark theme with a header featuring the Elasticsearch logo and navigation links for Tutorials, Examples, Integrations, and Blogs. A yellow "Start free trial" button is visible in the top right. The main content area has a large title "Prompts" and a subtext "Create and discover prompts that unlock more powerful ways to use Elasticsearch." Below this, there's a dropdown menu labeled "Business Automation". Two cards are displayed: "Meeting Summary Assistant" and "Asset Inventory Diagrams".

Meeting Summary Assistant

Key characteristics of the prompt:

- Objective: Transform raw meeting transcripts into structured and actionable summaries to enhance decision-making and team alignment.
- Process-Oriented: Outlines detailed steps for analyzing and summarizing transcripts, including handling transcription errors, extracting insights, and categorizing information.

[See more](#)

Asset Inventory Diagrams

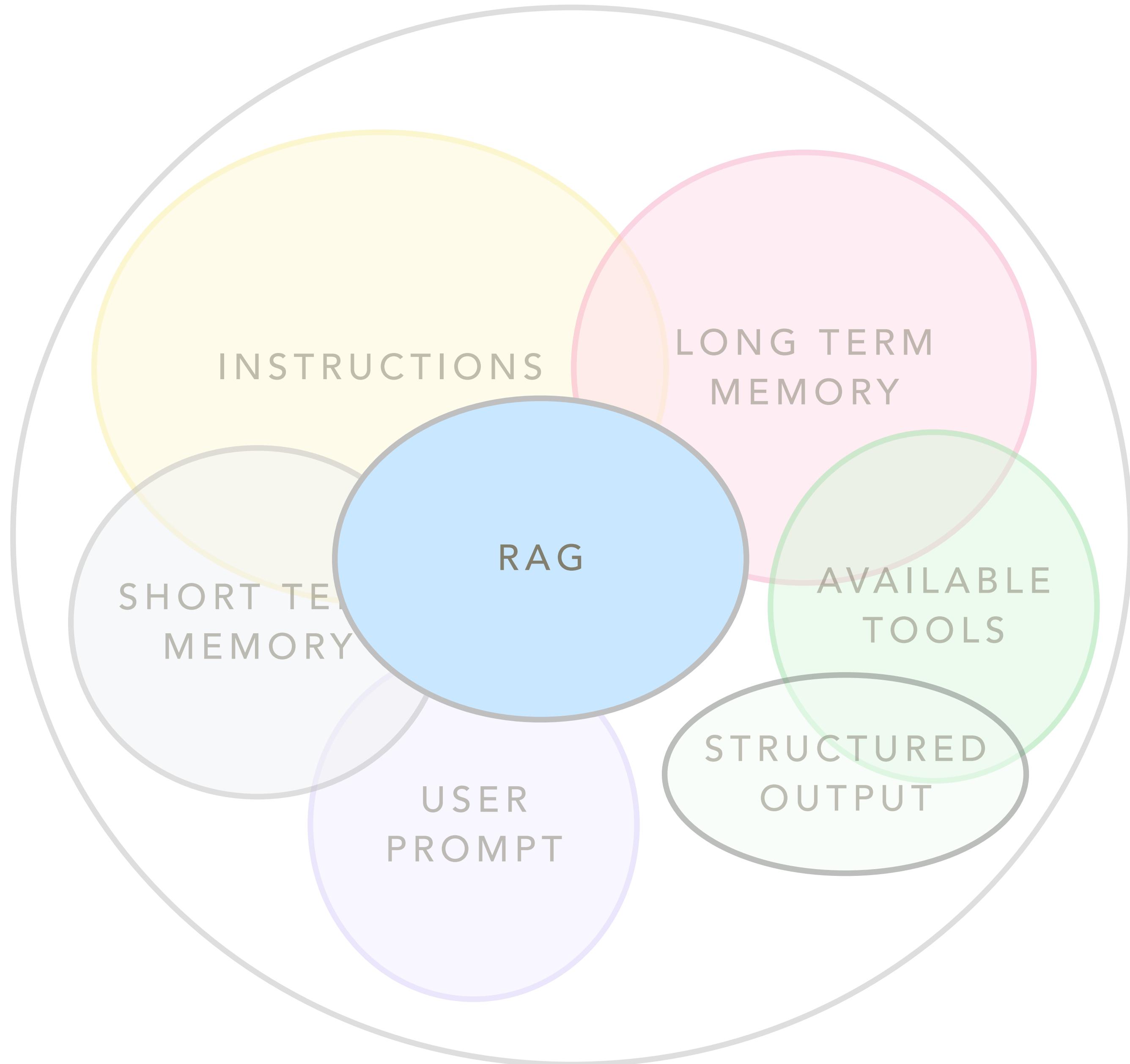
Easily visualize organization's assets and connections.

Create a diagram showing our organizational asset inventory that:

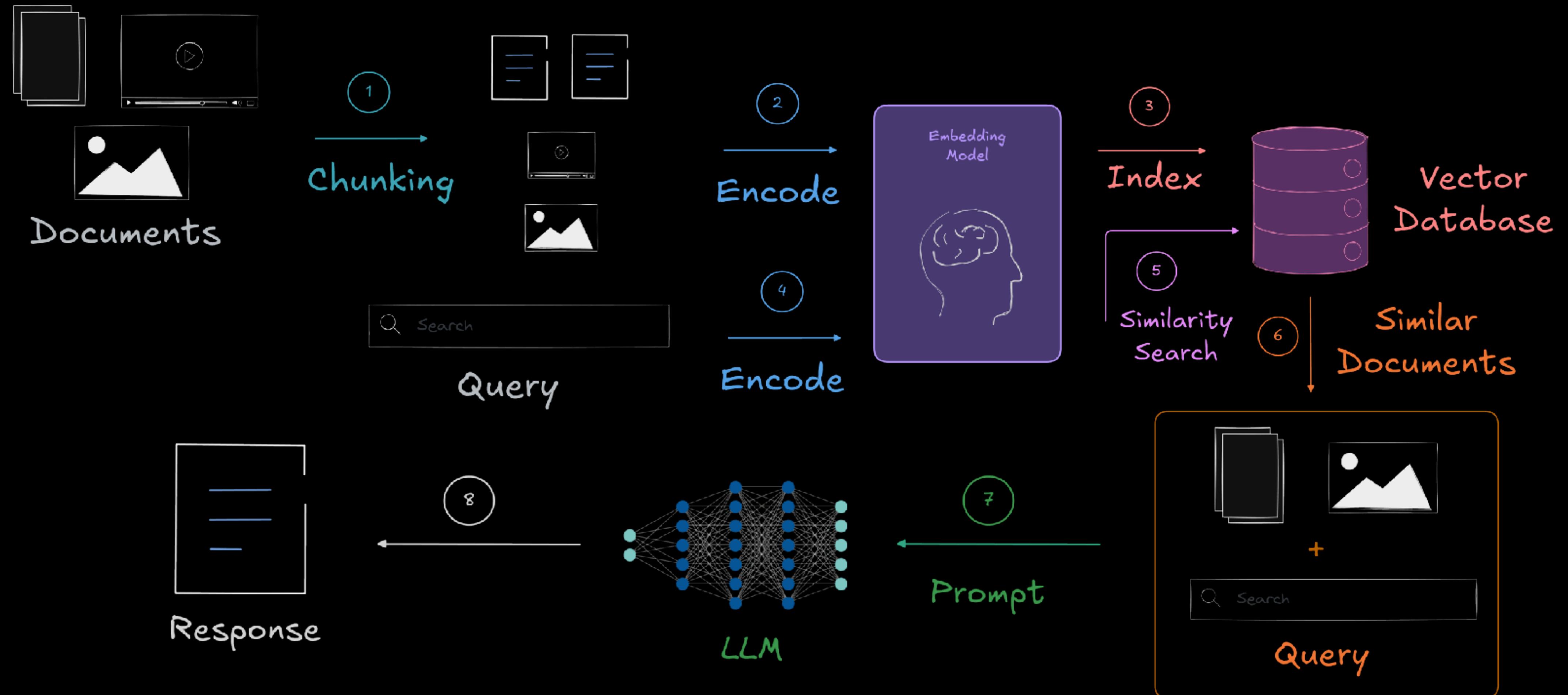
- Groups assets by type (servers, endpoints, IoT devices, etc.)
- Shows ownership and business criticality
- Indicates security patch status
- Highlights compliance requirements
- Maps dependencies between assets

Include these specific assets and relationships: [Insert asset details here] Please use Mermaid diagram syntax and include a legend.

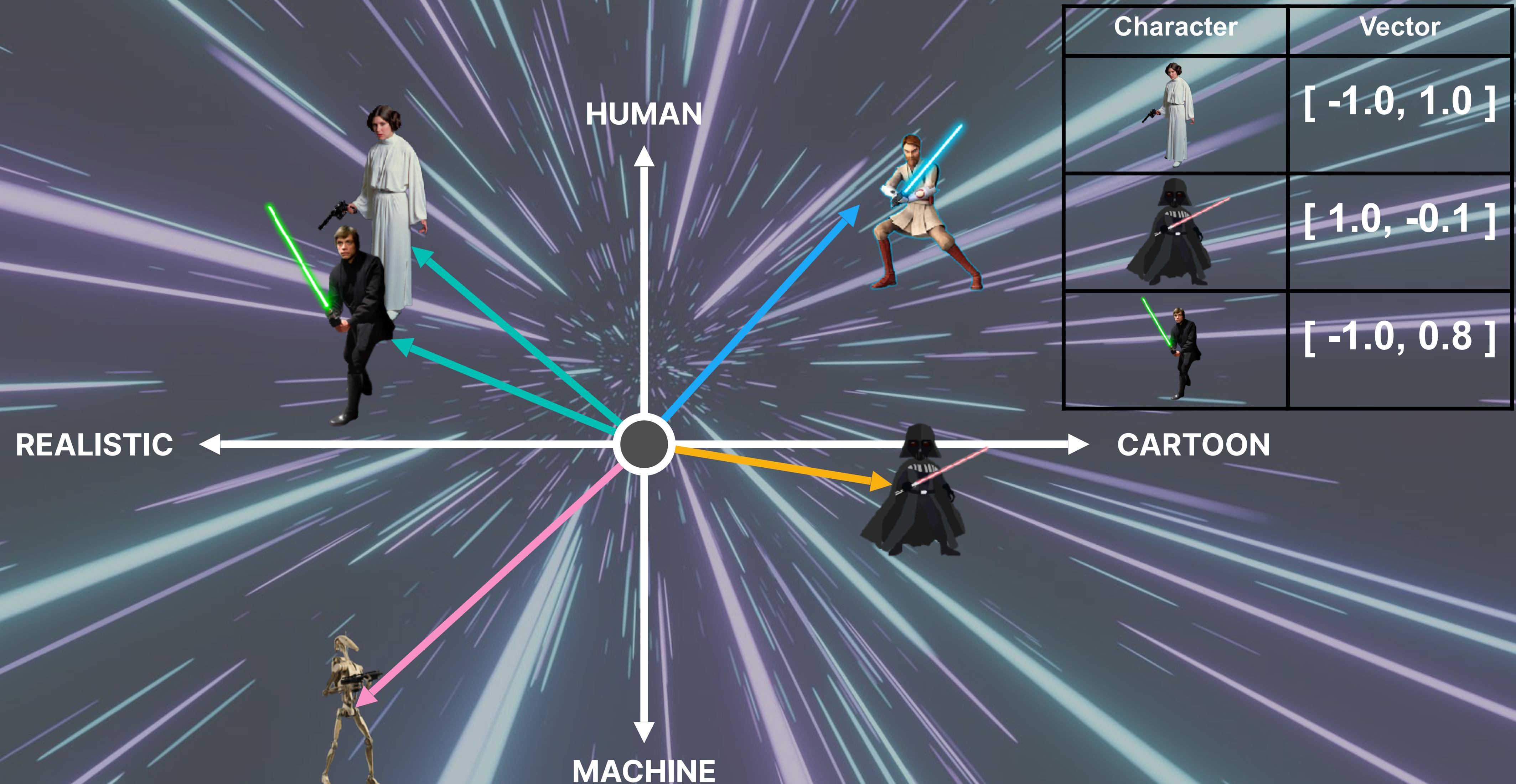
[See less](#)

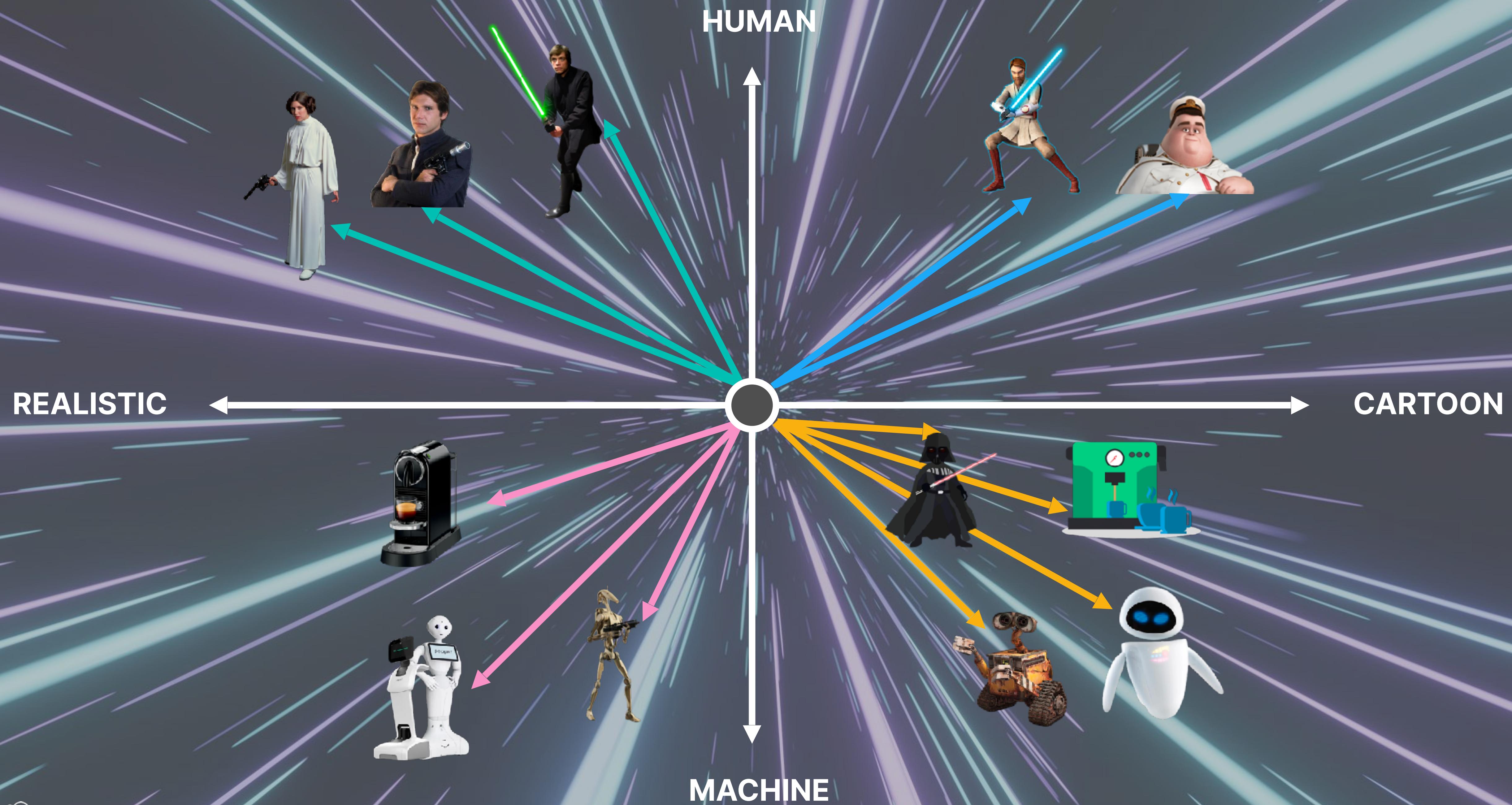


WHAT IS RAG?



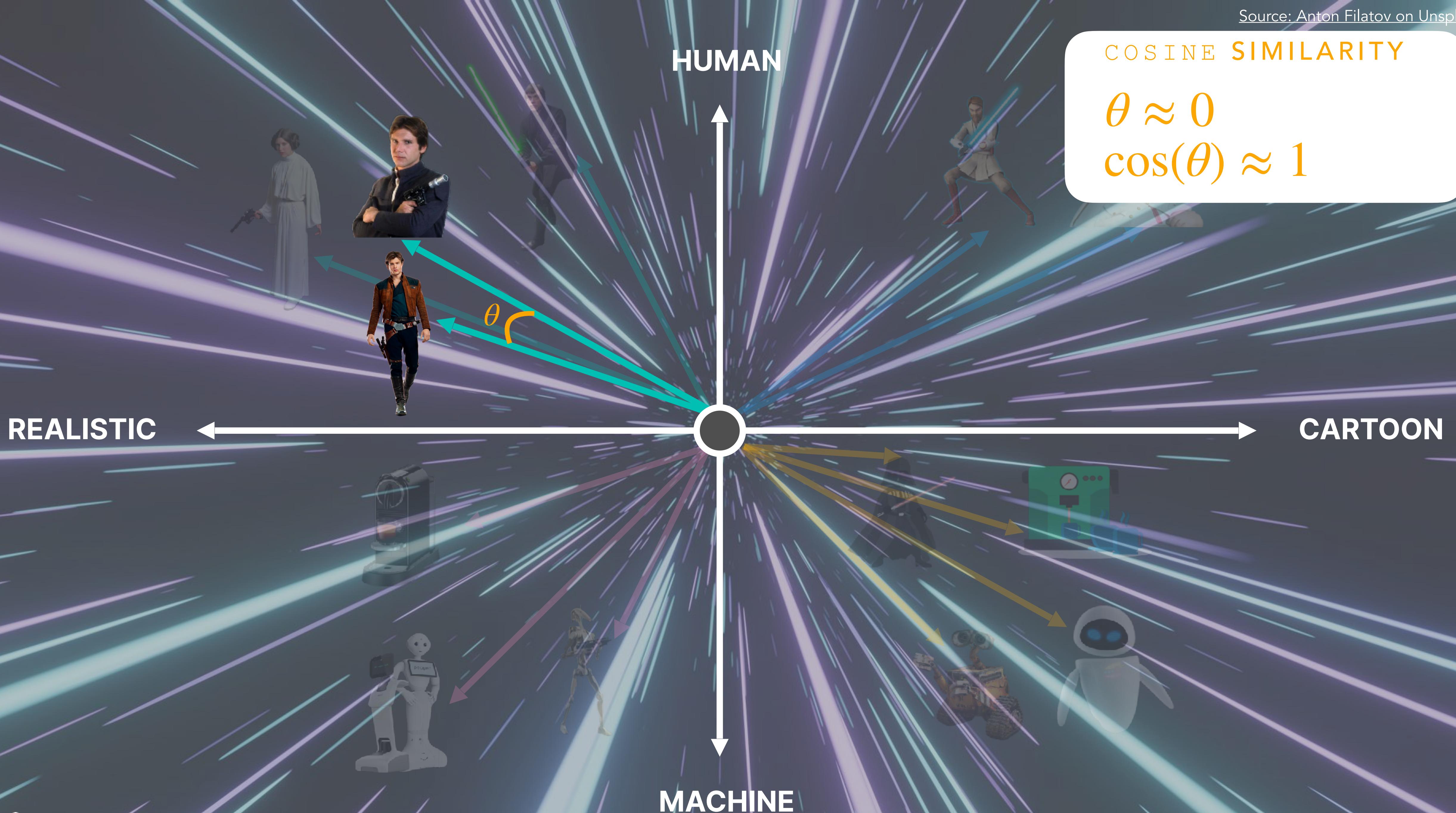
EMBEDDINGS REPRESENT YOUR DATA

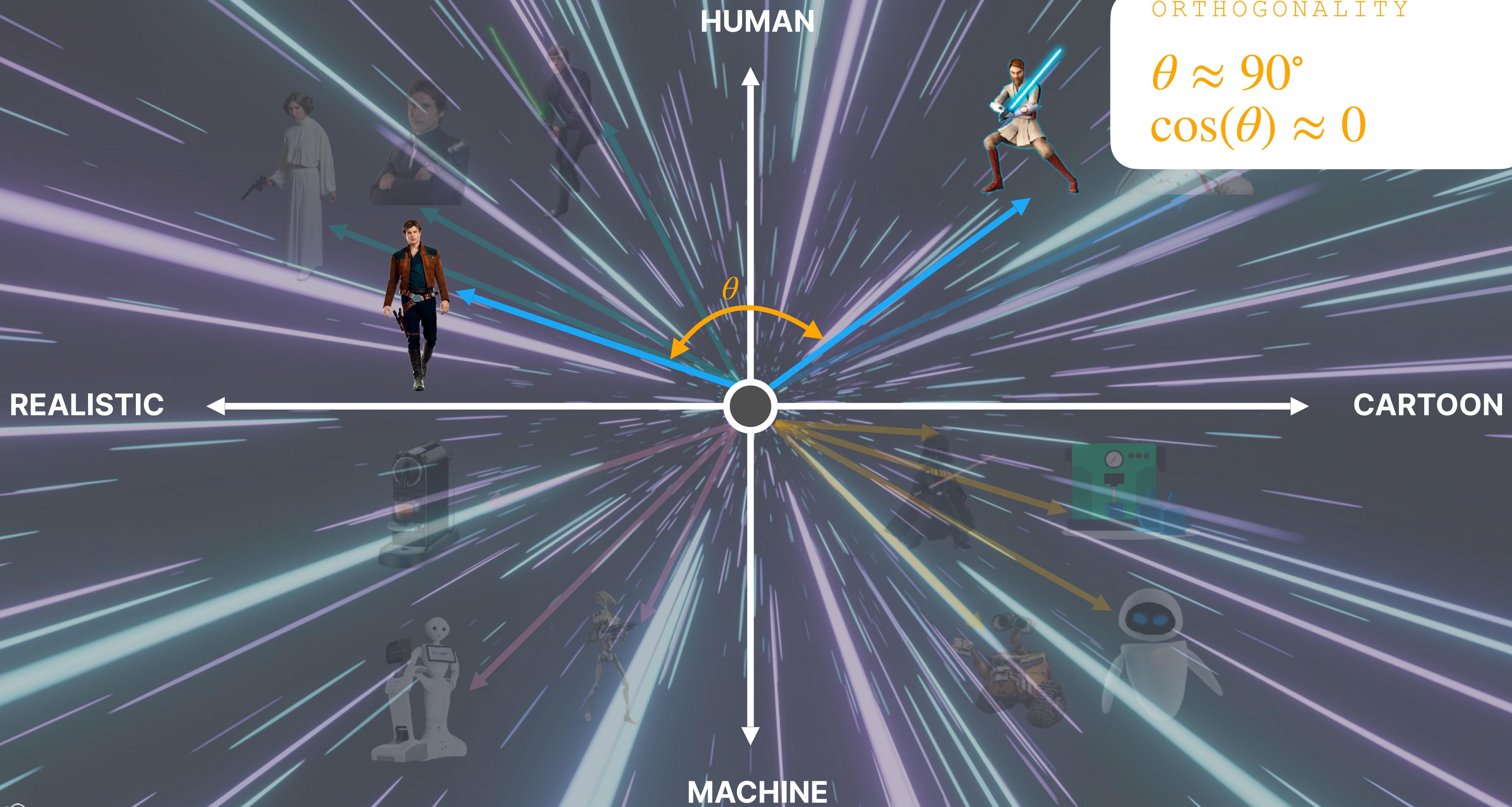


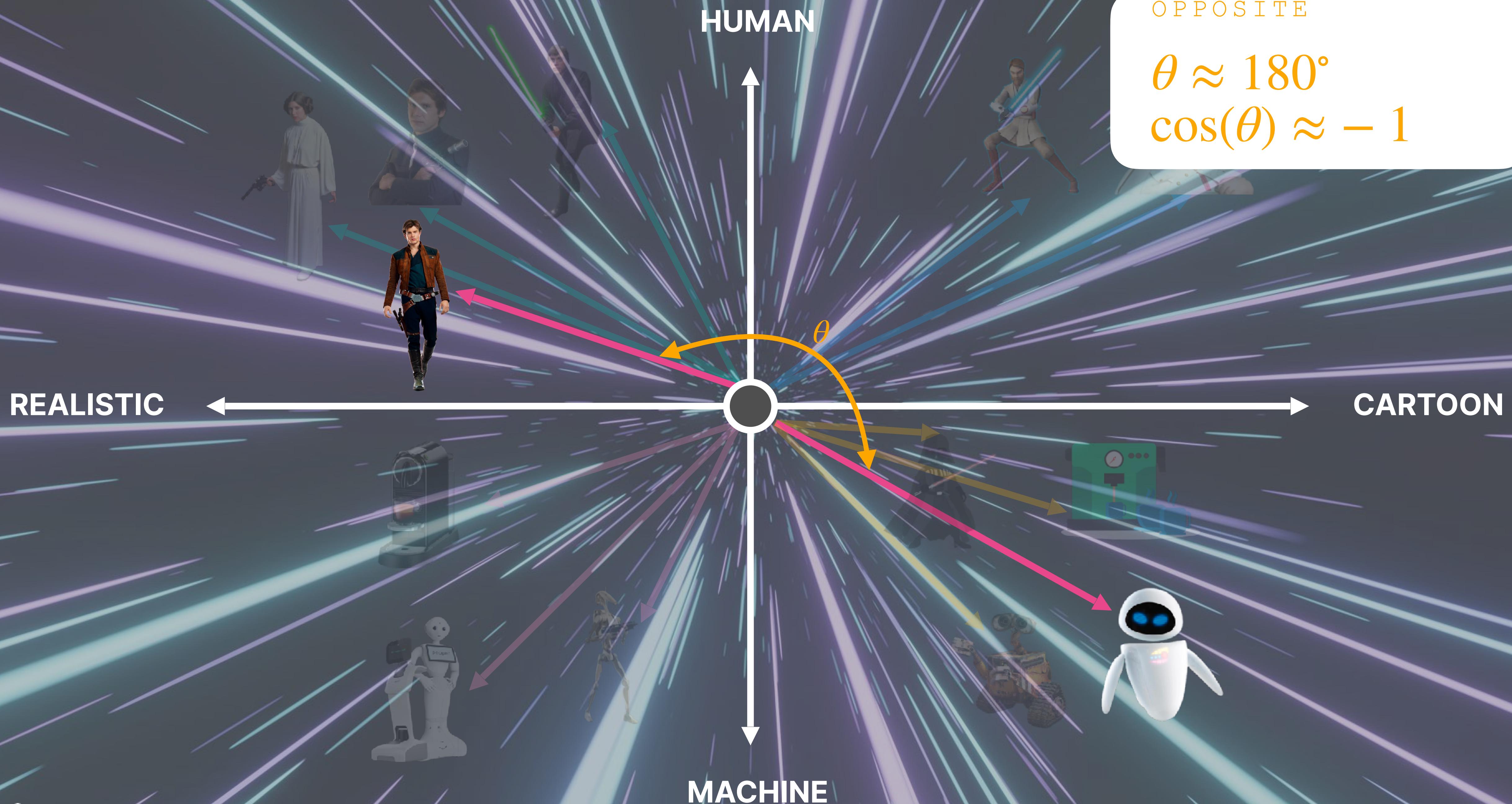


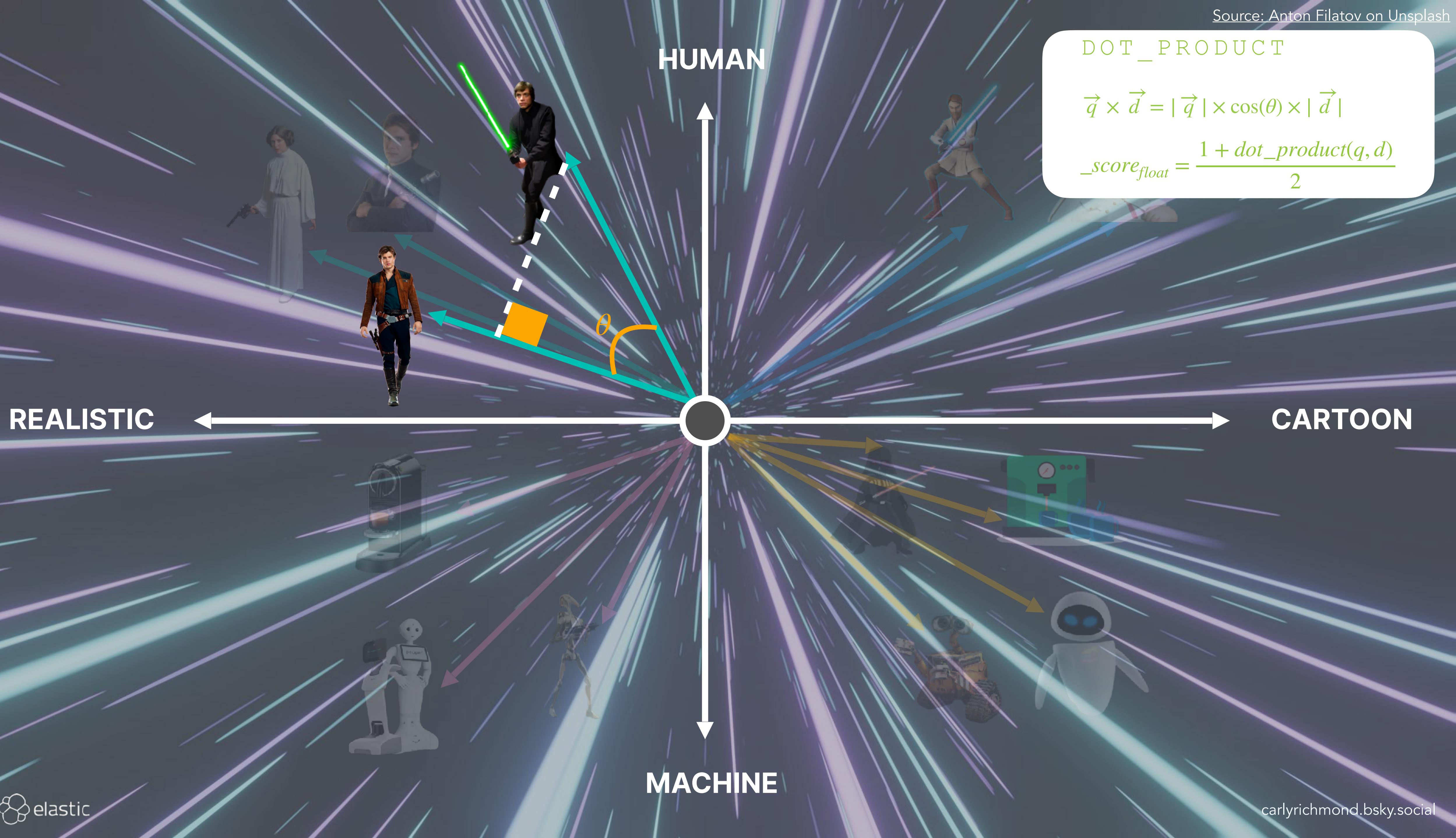
VECTOR SEARCH RANKS OBJECTS BY SIMILARITY (RELEVANCE) TO THE QUERY

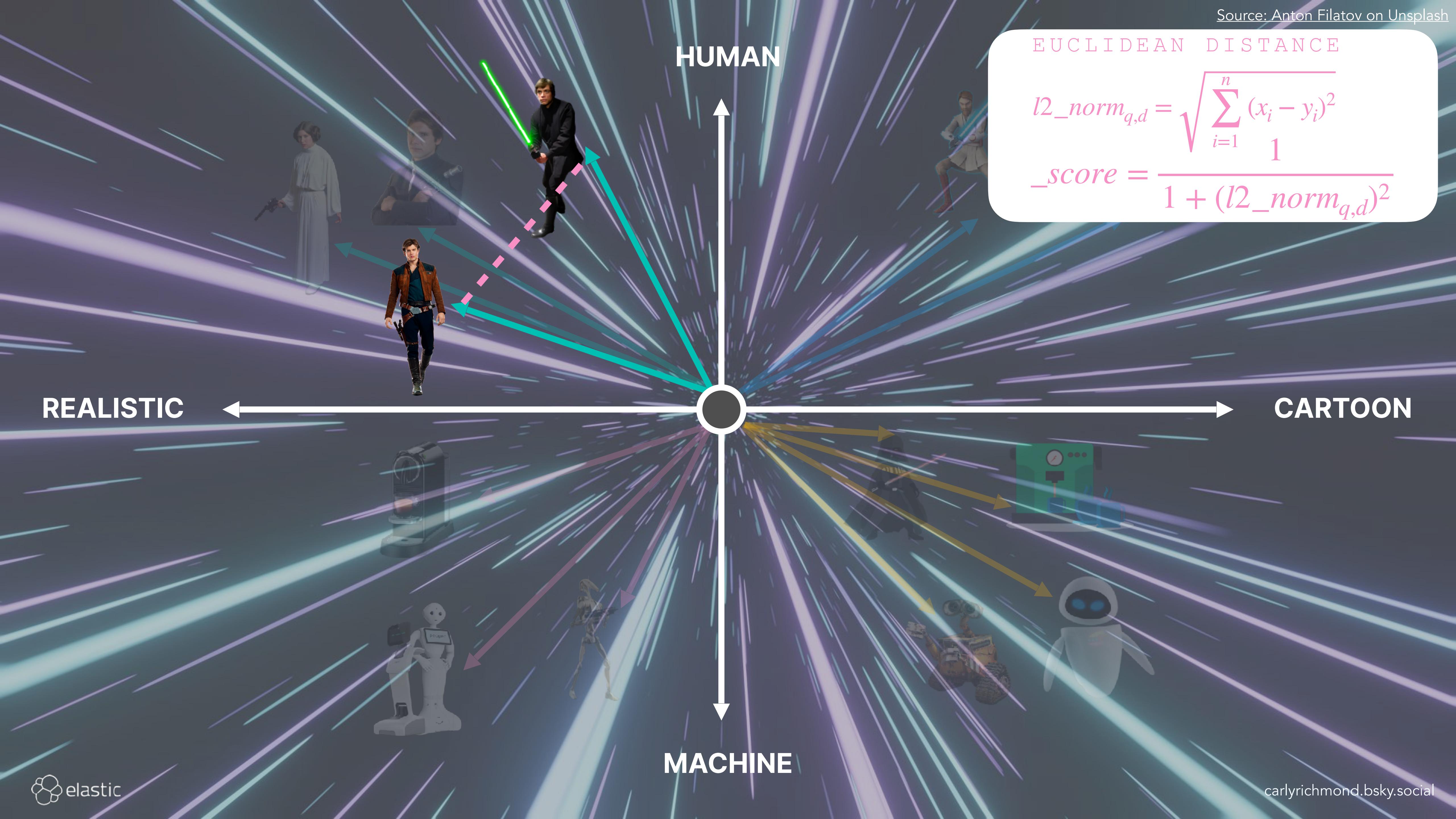


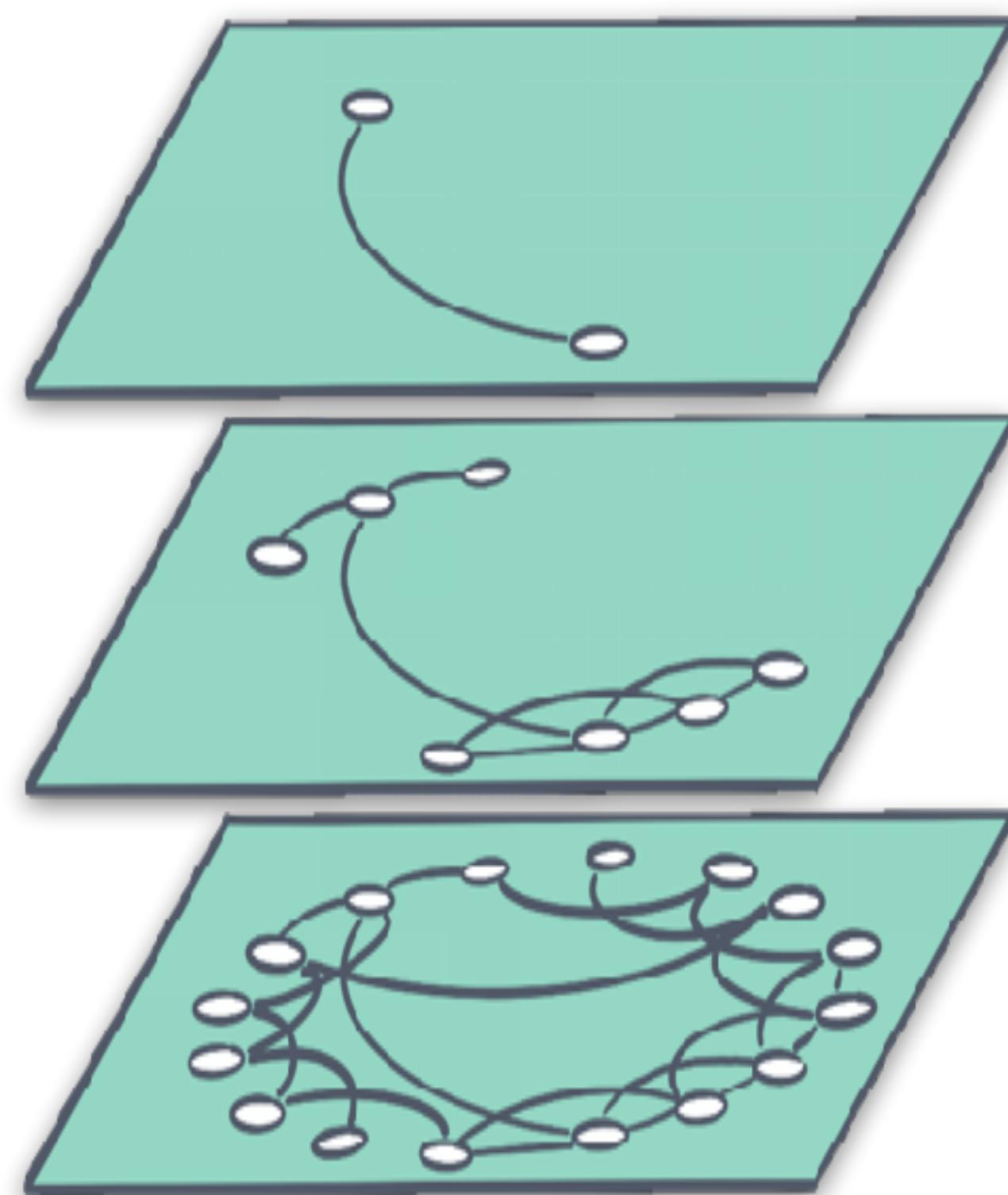












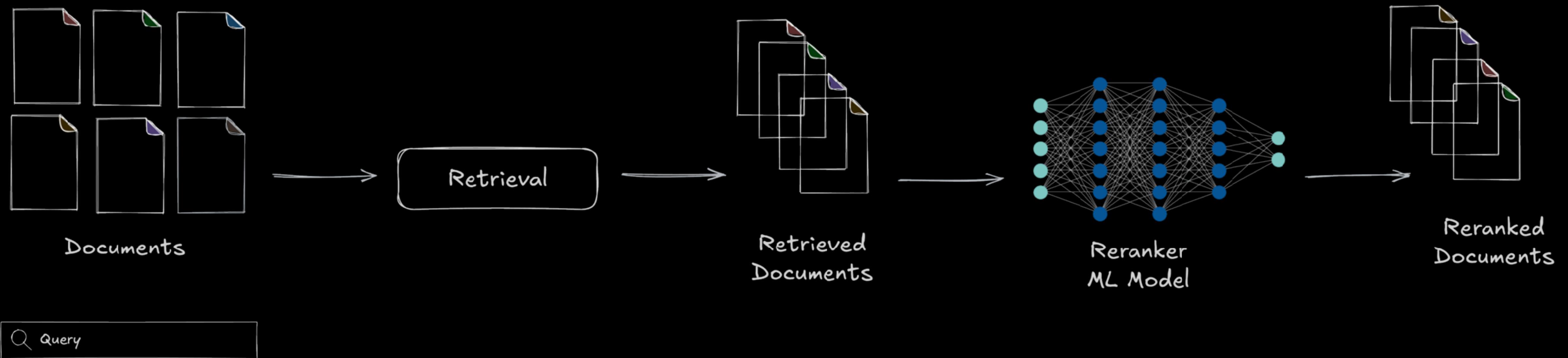
[-0.0078373895958599, -0.215008884668022]



WHAT IF I HAVE MILLIONS OF
VECTORS?

```
"filter": {  
  "term": {  
    "genre": "sci-fi"  
  }  
}
```

RERANKING



Models – Hugging Face

huggingface.co/models?sort=trending&search=rerank

Hugging Face Search models, datasets, users...

Models Datasets Spaces Community Docs Enterprise Pricing Log In Sign Up

Main Tasks Libraries Languages Licenses Other

Tasks

- Text Generation Any-to-Any Image-Text-to-Text
- Image-to-Text Image-to-Image Text-to-Image
- Text-to-Video Text-to-Speech + 44

Parameters

- < 1B 6B 12B 32B 128B > 500B

Libraries

- PyTorch TensorFlow JAX Transformers
- Diffusers sentence-transformers Safetensors
- ONNX GGUF Transformers.js MLX MLX + 41

Apps

- vLLM TGI llama.cpp MLX LM
- LM Studio Ollama Jan + 7

Inference Providers

- Groq Novita Cerebras SambaNova
- Nscale fal Hyperbolic Together AI + 10

Models 1,433 rerank

Full-text search Inference Available Sort: Trending

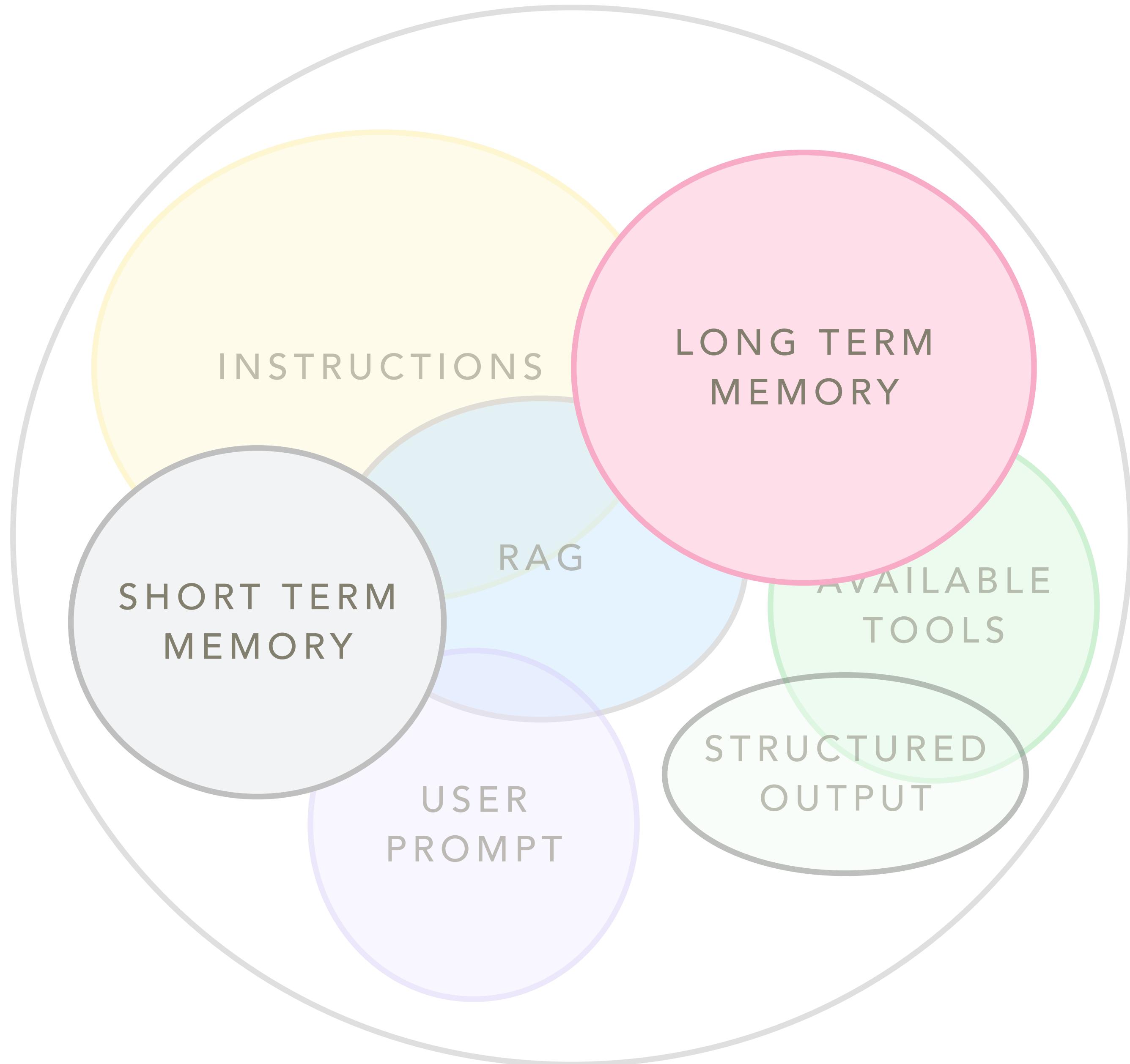
- Qwen/Qwen3-VL-Reranker-2B**
Text Ranking · 2B · Updated 2 days ago · 73.9k · 140
- BAAI/bge-reranker-v2-m3**
Text Classification · 0.6B · Updated Jun 24, 2024 · 2.84M · 866
- naver/provence-reranker-debertav3-v1**
Text Ranking · 0.4B · Updated Sep 26, 2025 · 668 · 63
- mixedbread-ai/mxbai-rerank-xsmall-v1**
Text Ranking · 70.8M · Updated Apr 2, 2025 · 1.35M · 51
- jinaai/jina-reranker-v3**
Text Ranking · Updated Nov 27, 2025 · 38.1k · 85
- nvidia/llama-nemotron-rerank-1b-v2**
Text Ranking · 1B · Updated Nov 25, 2025 · 18.4k · 28
- BAAI/bge-reranker-large**
Feature Extraction · 0.6B · Updated May 11, 2024 · 630k · 447
- Dongjin-kr/ko-reranker**
Text Classification · 0.6B · Updated Jan 13, 2025 · 3.64k · 59
- jinaai/jina-reranker-v2-base-multilingual**
Text Ranking · 0.3B · Updated Oct 21, 2025 · 401k · 336
- qilowqq/bge-reranker-v2-m3-en-ru**
Text Ranking · 0.4B · Updated Apr 5, 2025 · 482 · 11
- ayushhexel/reranker-ms-marco-MiniLM-L6-v2-gooaq-bce**
Text Ranking · 22.7M · Updated Mar 30, 2025 · 1 · 2
- dean2155/Qwen3-Reranker-4B-Q8_0-GGUF**
· 4B · Updated Jun 5, 2025 · 30 · 1
- Qwen/Qwen3-Reranker-8B**
Text Ranking · 8B · Updated Jun 9, 2025 · 617k · 209
- voyageai/rerank-2.5**
Updated Jul 17, 2025 · 6

DEMO

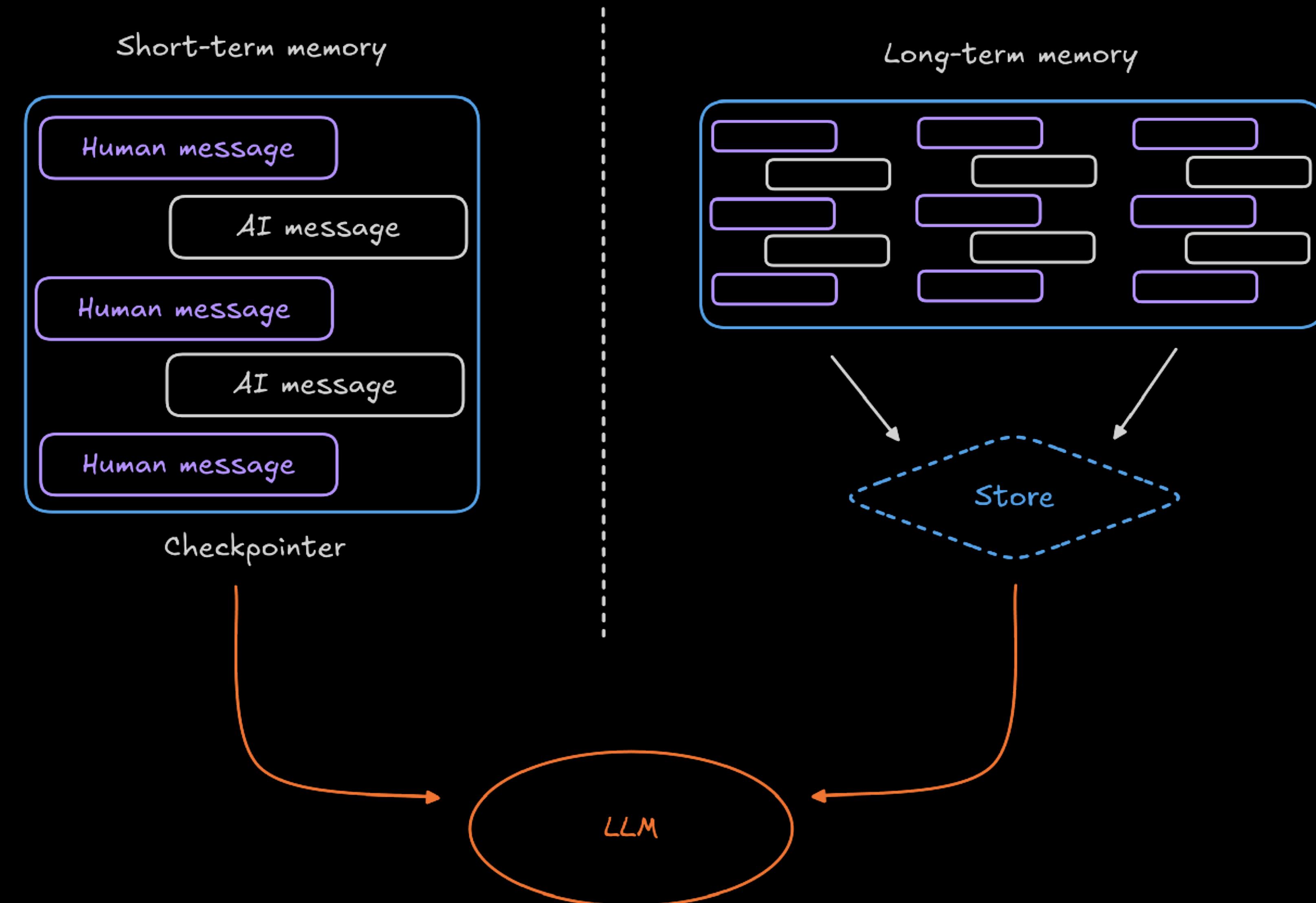


SCAN ME





MEMORY



WHERE TO STORE MEMORIES?



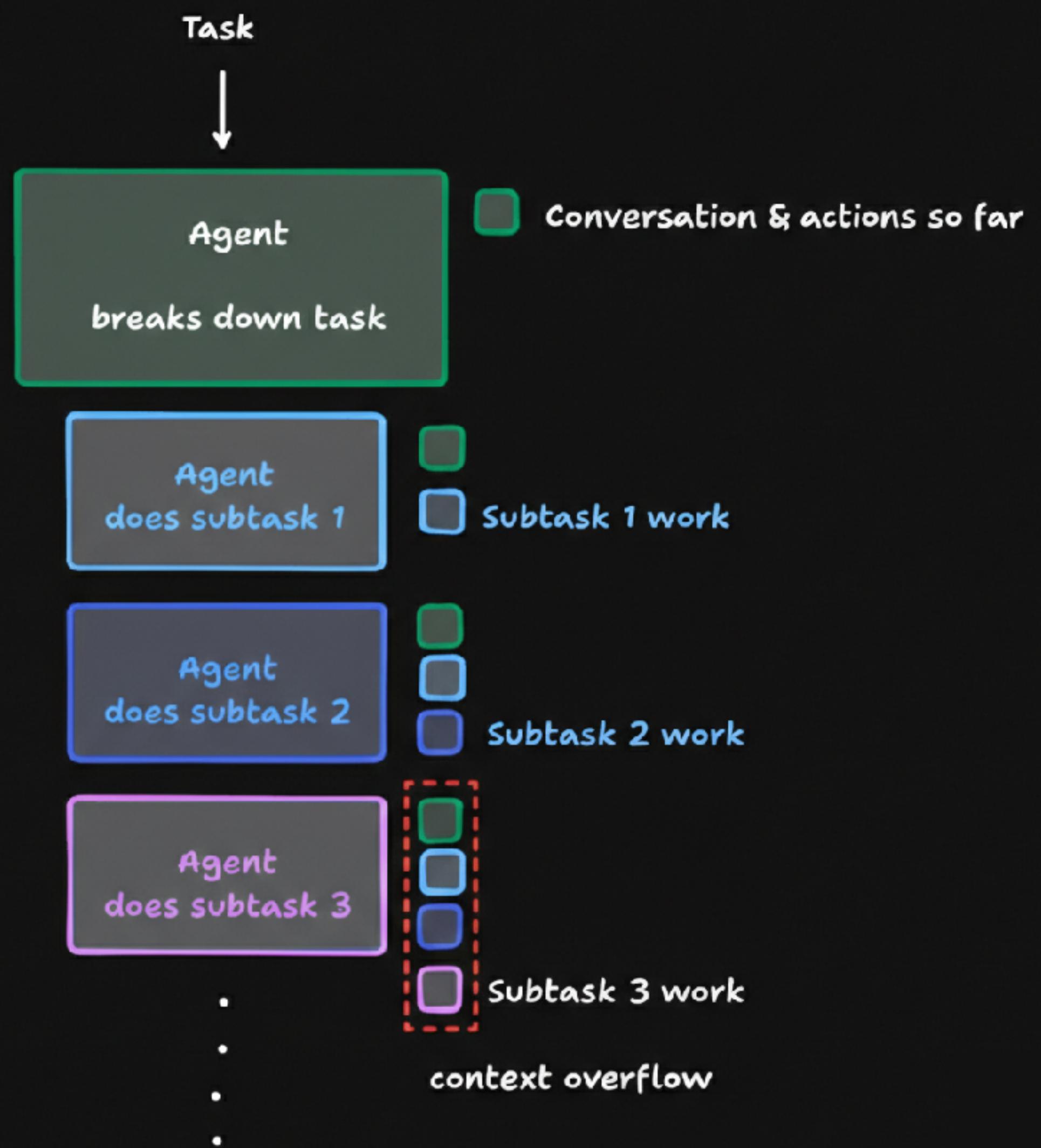
DEMO



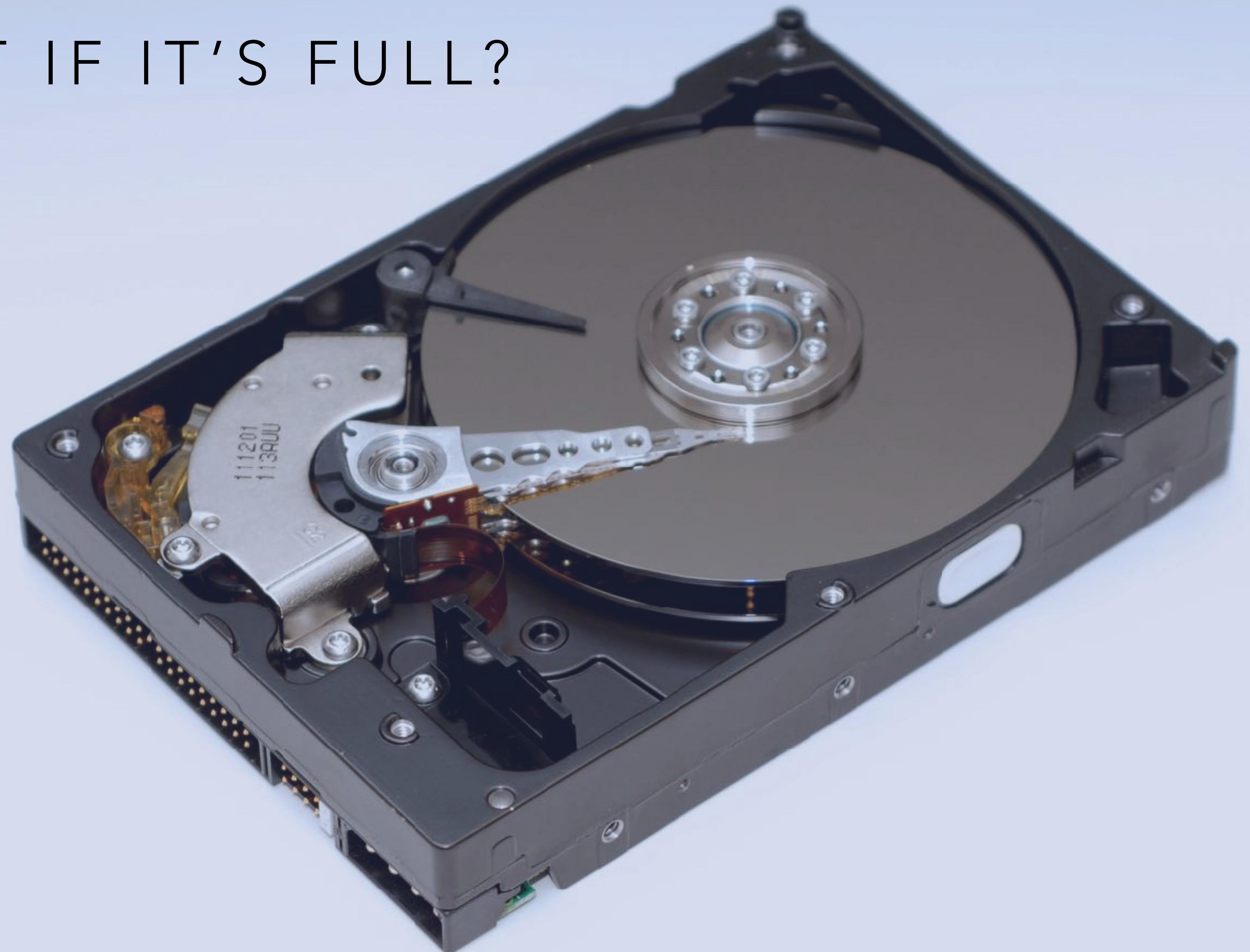
SCAN ME



SUBAGENT MEMORY



WHAT IF IT'S FULL?



Provence: efficient and robust context pruning for retrieval-augmented generation

Nadezhda Chirkova Thibault Formal Vassilina Nikoulina Stéphane Clinchant

NAVER LABS Europe, Grenoble, France

<https://huggingface.co/naver/provence-reranker-debertav3-v1>

Question: Can you eat pumpkin every day?

Retrieved context:

Pumpkin is at its peak in the fall. Eating pumpkin every day can help reduce inflammation, strengthen your immune system and promote eye health. It may also help lower blood pressure. However, if you eat too much, you may experience diarrhea from a high dose of fiber. ~~Read on to learn all about pumpkin's nutrition and health benefits.~~

Relevance score:

0.96

PROVENCE

Provence removes sentences that are irrelevant to the user question

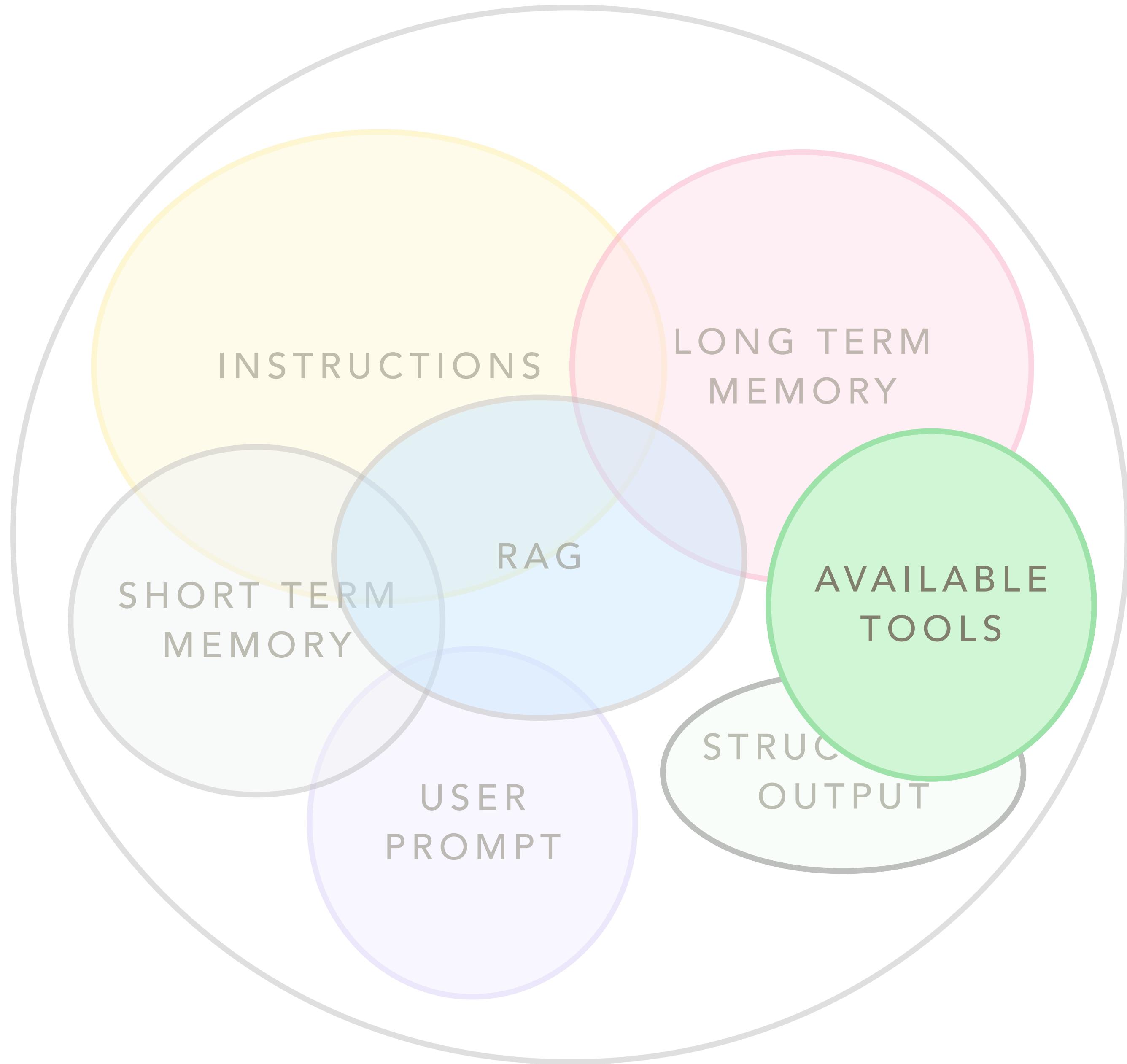
... and also outputs a relevance score

DEMO

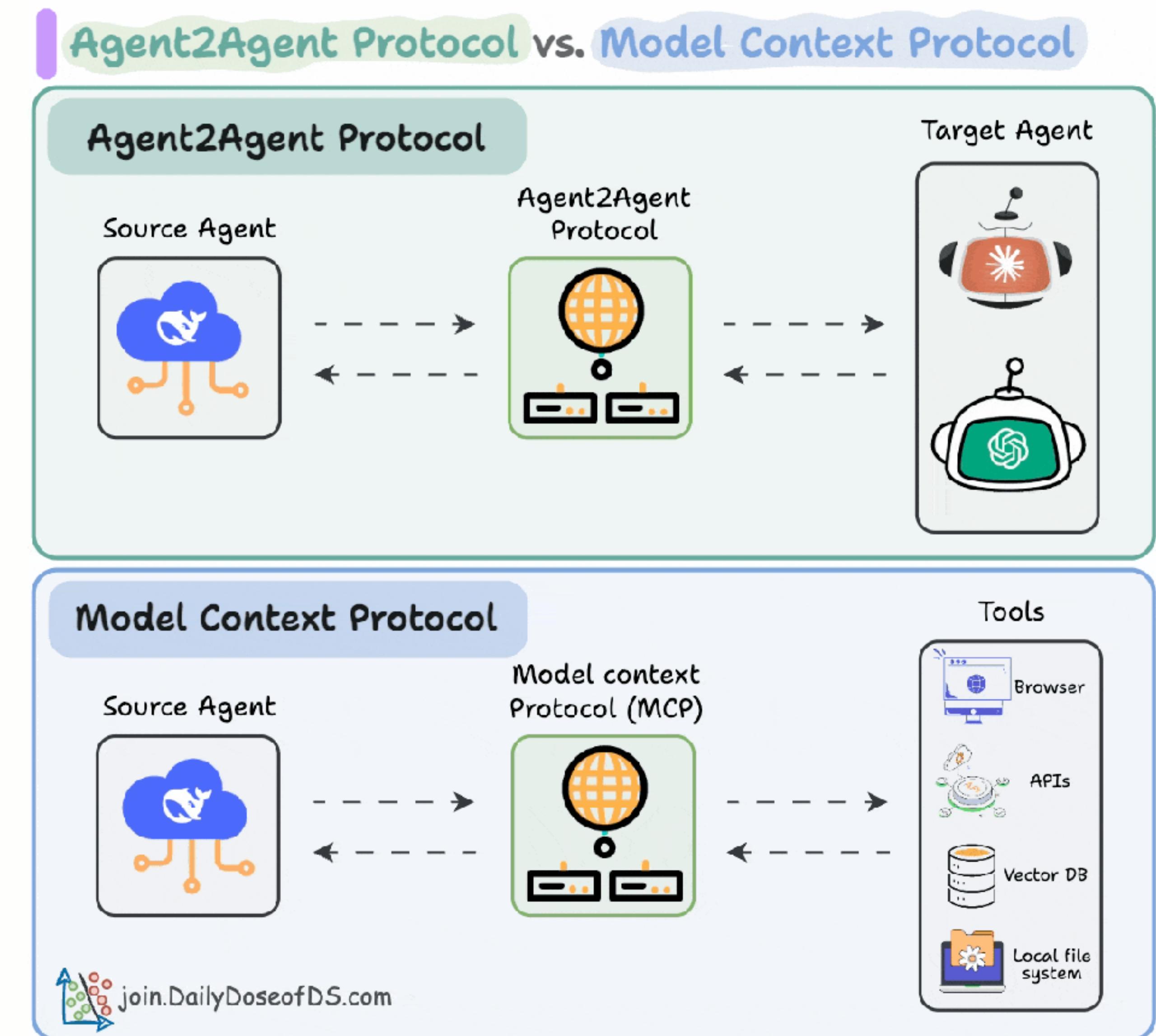


SCAN ME





MCP AND A2A





Berkeley Function-Calling Leaderboard

BFCL: From Tool Use to Agentic Evaluation of Large Language Models

The Berkeley Function Calling Leaderboard (BFCL) V4 evaluates the LLM's ability to call functions (aka tools) accurately. This leaderboard consists of real-world data and will be updated periodically. For more information on the evaluation dataset and methodology, please refer to our blogs: [BFCL-v1](#) introducing AST as an evaluation metric, [BFCL-v2](#) introducing enterprise and OSS-contributed functions, [BFCL-v3](#) introducing multi-turn interactions, and [BFCL-v4](#) introducing holistic agentic evaluation. Checkout [code and data](#).

Last Updated: 2025-12-16 [\[Change Log\]](#)

Search model names...

Search

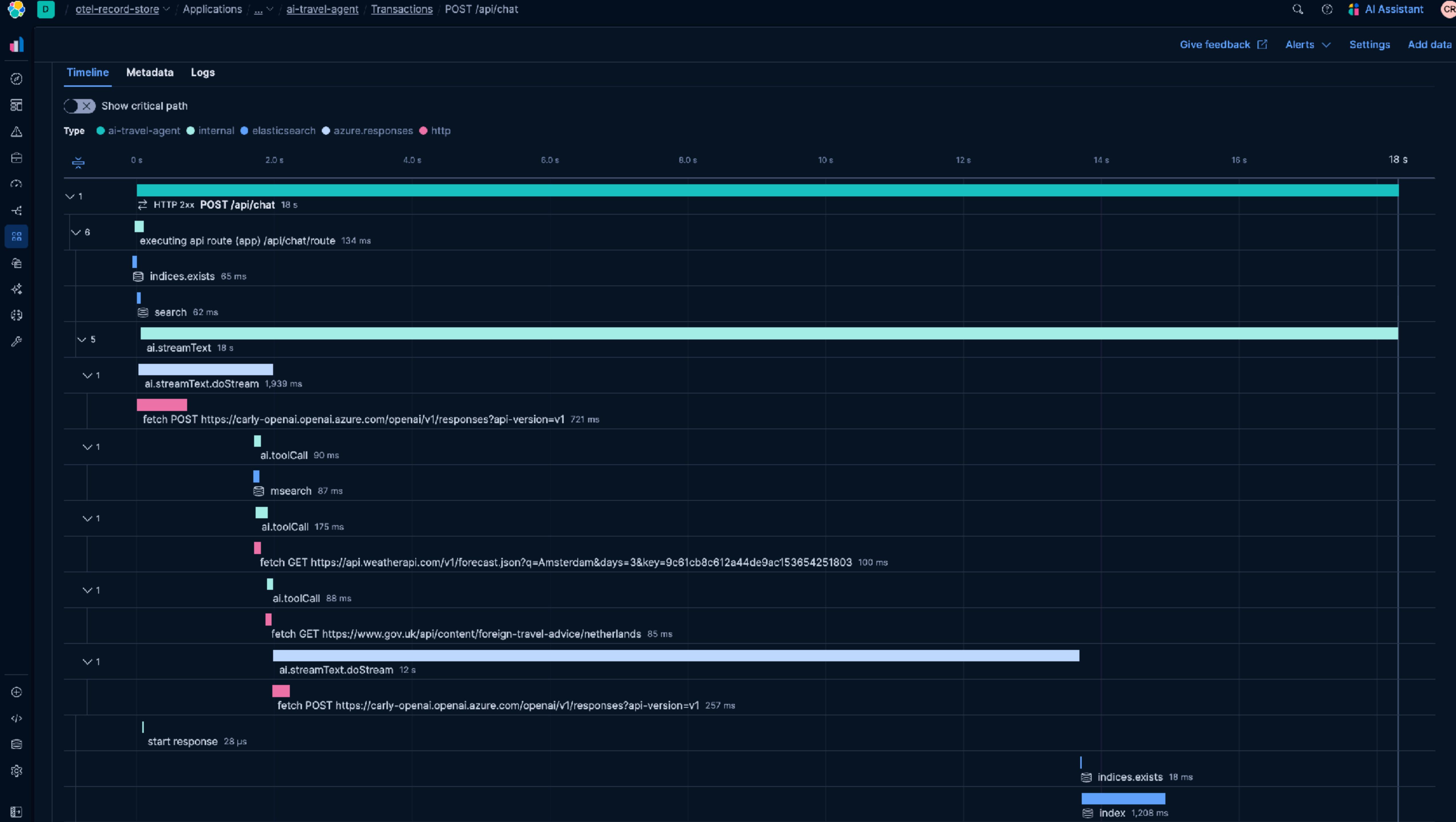
				Agentic		Multi Turn	Single Turn		Hallucination Measurement	
Rank	Overall Acc	Model	Cost (\$)	Web Search ►	Memory ►	Multi turn ►	Non-live (AST) ►	Live (AST) ►		
				Overall Acc	Overall Acc	Overall Acc	Overall Acc	Overall Acc	Relevance	Irrelevance
1	77.47	Claude-Opus-4-5-20251101 (FC)	86.55	84.5	73.76	68.38	88.58	79.79	62.5	84.72
2	73.24	Claude-Sonnet-4-5-20250929 (FC)	43.73	81	64.95	61.37	88.65	81.13	68.75	86.61
3	72.51	Gemini-3-Pro-Preview (Prompt)	298.47	80	61.72	60.75	90.65	83.12	68.75	85.59
4	72.38	GLM-4.6 (FC thinking)	4.64	77.5	55.7	68	87.56	80.9	75	84.96
5	69.57	Grok-4-1-fast-reasoning (FC)	17.26	82.5	53.98	58.87	88.27	78.46	81.25	79.43
6	68.7	Claude-Haiku-4-5-20251001	14.23	83.5	54.41	53.62	86.5	78.68	62.5	85.11

DEMO



SCAN ME





New Agent

[Save and chat](#)[Save](#)

Create an AI agent with custom instructions, assign it tools to work with your data, and make it easily findable for your team. [Learn more](#)

[Settings](#) [Tools](#) 5

This agent has 5 active tools

Tools enable agents to work with your data. For best results, keep the selection under 24 to avoid overwhelming your agent with too many options.

Active tools

5/24

 Search

Labels ▾

 Show active only

Showing 1-10 of 10 Tools

ID ↕

Labels

 platform.core.search

A powerful tool for searching and analyzing data within your Elasticsearch cluster.

 platform.core.get_document_by_id

Retrieve the full content (source) of an Elasticsearch document based on its ID and index name.

 platform.core.execute_esql

Execute an ES|QL query and return the results in a tabular format.

 platform.core.generate_esql

Generate an ES|QL query from a natural language query.

platform.core.get_index_mapping

[Cancel](#)[Save and chat](#)[Save](#)

RAG-MCP: Mitigating Prompt Bloat in LLM Tool Selection via Retrieval-Augmented Generation

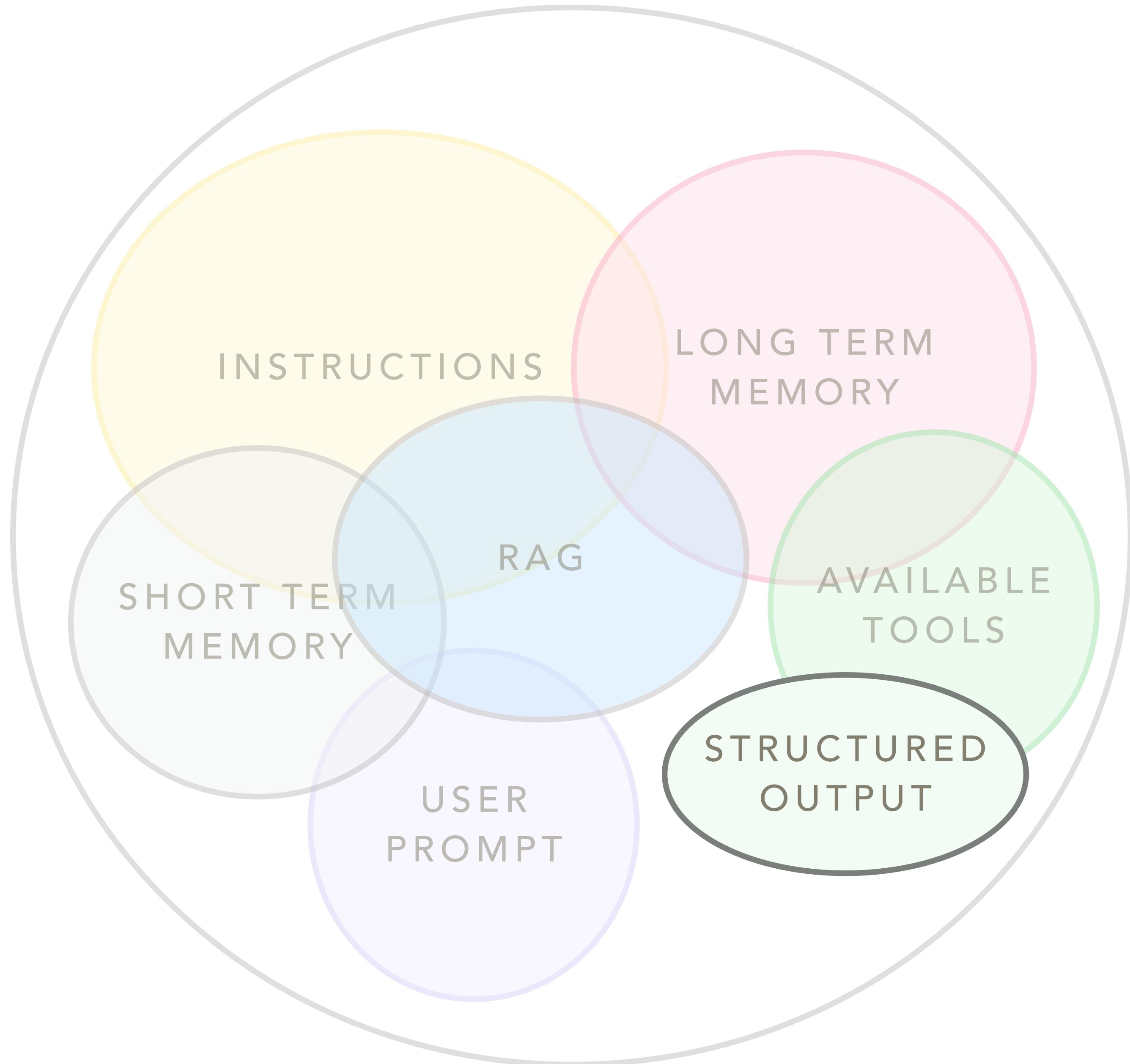
Tiantian Gan^{1,2} and Qiyao Sun^{1,2}

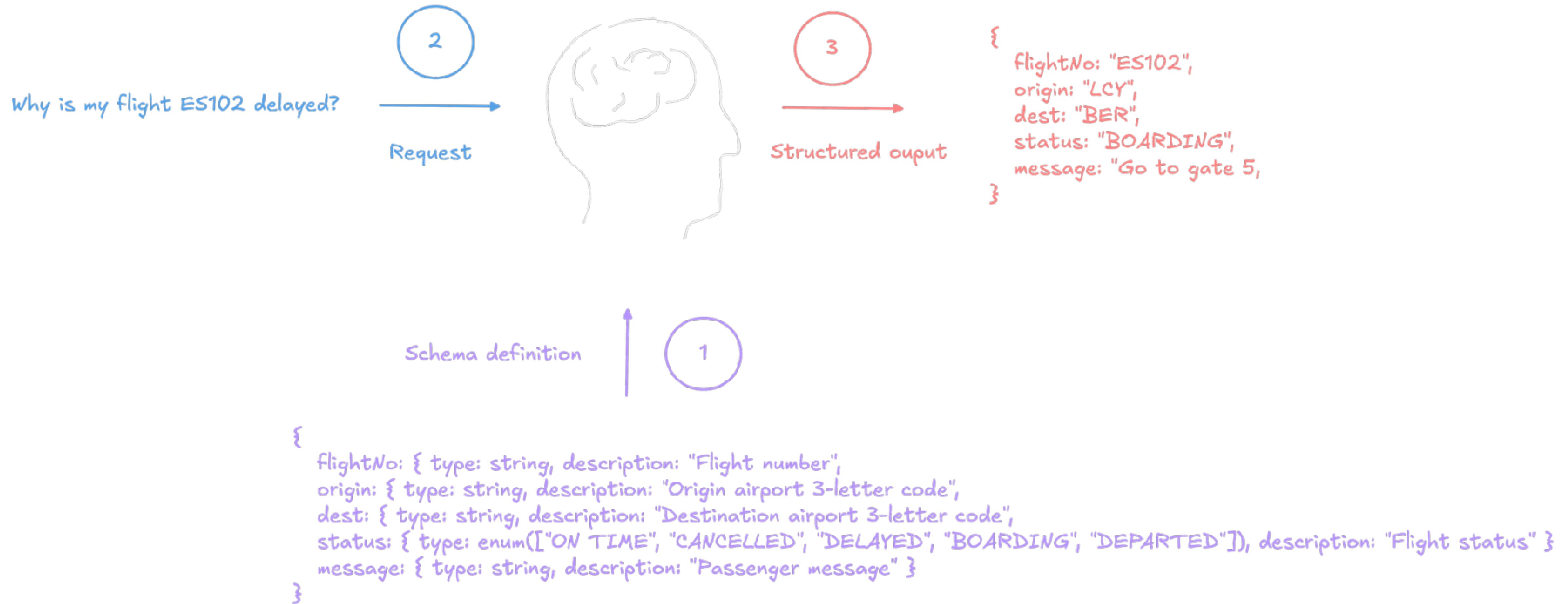
¹ Beijing University of Post and Communications, Beijing, China

² Queen Mary University of London, London, UK

jp2022213034@qmul.ac.uk, jp2022213402@qmul.ac.uk

Abstract. Large language models (LLMs) struggle to effectively utilize a growing number of external tools, such as those defined by the Model Context Protocol (MCP)[1], due to prompt bloat and selection complexity. We introduce RAG-MCP, a Retrieval-Augmented Generation framework that overcomes this challenge by offloading tool discovery. RAG-MCP uses semantic retrieval to identify the most relevant MCP(s) for a given query from an external index before engaging the LLM. Only the selected tool descriptions are passed to the model, drastically reducing prompt size and simplifying decision-making. Experiments, including an MCP stress test, demonstrate RAG-MCP significantly cuts prompt tokens (e.g., by over 50%) and more than triples tool selection accuracy (43.13% vs 13.62% baseline) on benchmark tasks. RAG-MCP enables scalable and accurate tool integration for LLMs.



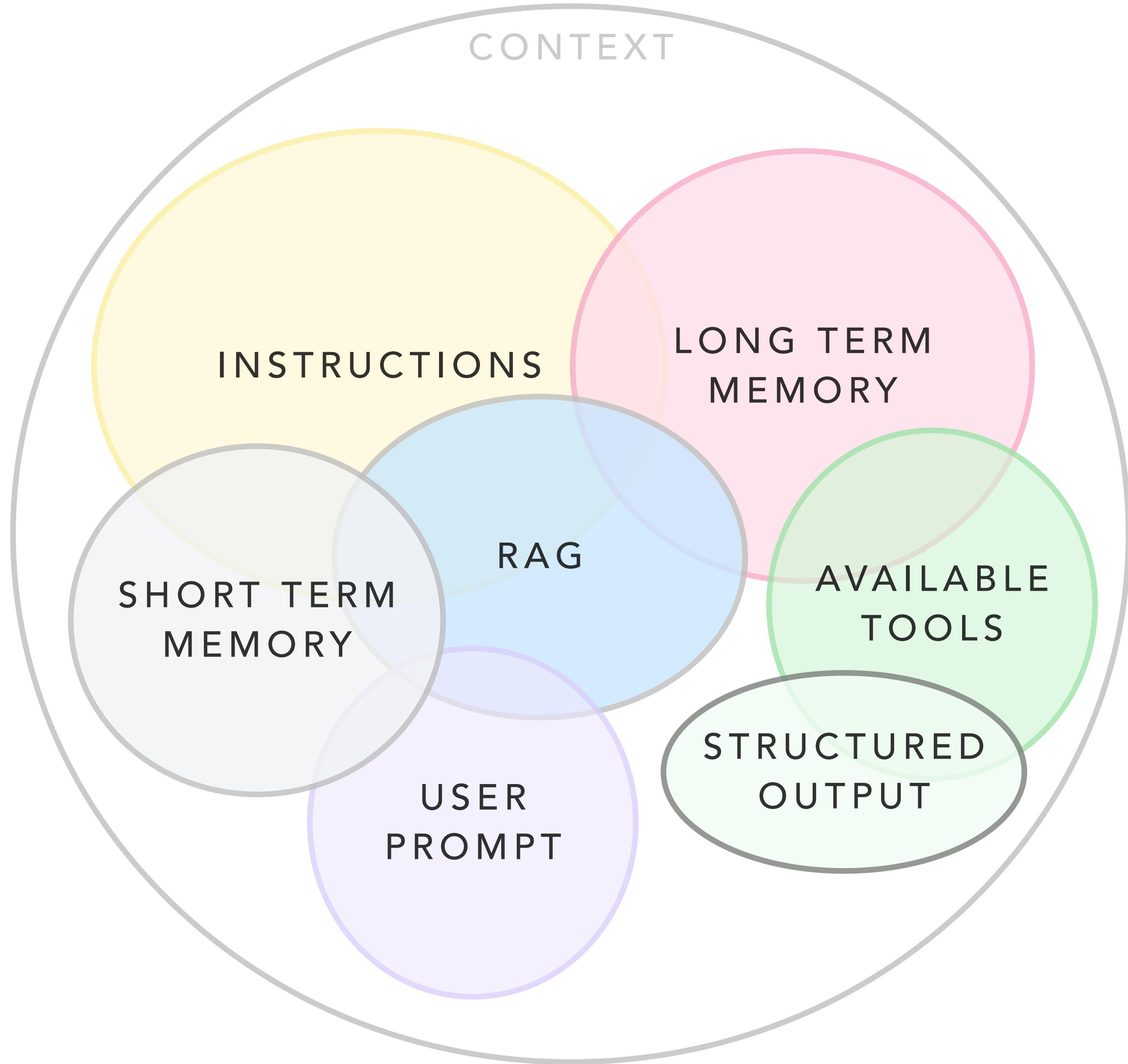


DEMO



SCAN ME





RESOURCES

- [What is Context Engineering | Carly Richmond @ Elasticsearch Labs](#)
- [12-Factor Agents - Principles for building reliable LLM applications | Dexter Horthy](#)
- [How to Fix Your Context | Drew Breunig](#)
- [The Prompt Report: A Systematic Survey of Prompt Engineering Techniques | Schulhoff et al.](#)
- [Prompt Engineering Guide](#)
- [What is Memory | LangChain](#)
- [Managing Agentic Memory with Elasticsearch | Som Mohankumar @ Elasticsearch Labs](#)
- [Don't Build Multi-Agents | Cognition](#)
- [Structured Outputs | LangChain](#)
- [RAG-MCP: Mitigating Prompt Bloat in LLM Tool Selection via Retrieval-Augmented Generation | Gan and Sun](#)
- [Why Language Models Hallucinate | OpenAI](#)



SCAN ME





SCAN ME

