

Generative AI for Law Librarians

Law Library Association of Maryland

January 8, 2025

Leland Sampson



Drop slides into chat!

- Manage the [People's Law Library](#) legal information website
- Member, MSBA AI Taskforce
- Member, MD Judiciary AI Workgroup
- <https://sampson.info/>
- Here in my personal capacity, I am NOT representing the Judiciary

Audience Poll – how much do you use AI?

					
0 – Never	1 – A tiny bit	2 – Monthly	3 – Weekly	4 – Daily	5 – I am AI
I don't even know how you get to the web page!	I asked it where I should go out to eat while I'm at the conference.	Every now and then I have an idea that I try.	I've started to consider how it can help with everything.	I use AI tools routinely, and tell others how they can.	I could reprogram ChatGPT to give number 4 a thumb!

Question 2: Do you currently pay to use AI tools?

Roadmap

- How does GenAI work?
- Why is a good prompt important?
- What does all this mean for legal research?

- Goal: A deeper understanding of GenAI tech

What is GenAI?

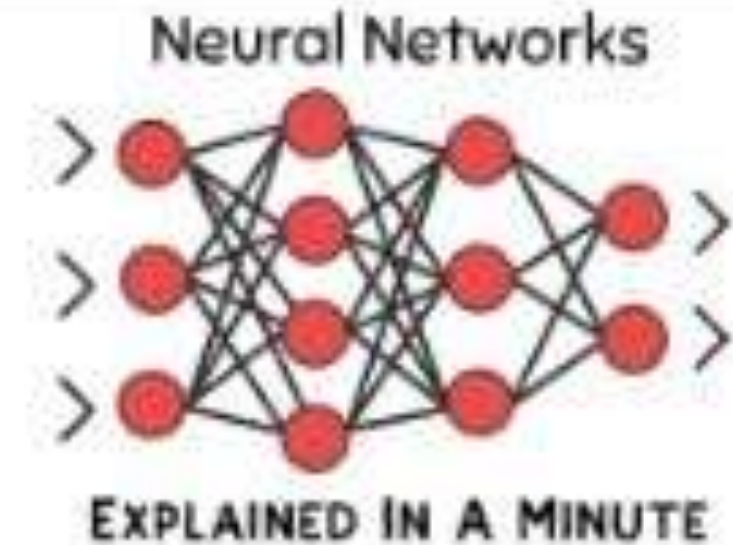
- GenAI “utilizes machine learning models to create new, original content, such as images, text, or music, based on patterns and structures learned from existing data. A prominent model type used by generative AI is the large language model (LLM).”*
- *from [Cornell University](#)

How a Large Language Model Works

- A large language model
- "**understands**" through statistical analysis and probability
- "**creates**" through combination of pattern, probability, and random sampling

Evolution of GenAI tools

- LLM “understanding”
 - LLM is “trained” on billions of word associations
 - Large LLMs might have 405 billion parameters – associations



[YouTube Link](#)

Evolution of GenAI tools

- LLM “creates”
 - LLM “knows things” because knowledge was part of it’s training data
- How does an LLM pick the next word it’s going to generate?

Tokenisation – Turning words (or fragments of words) into numbers

The promise of large language models is that they —

464 6991 286 1588 3303 4981 318 326 484

Embedding – Represents a word's meaning in its context

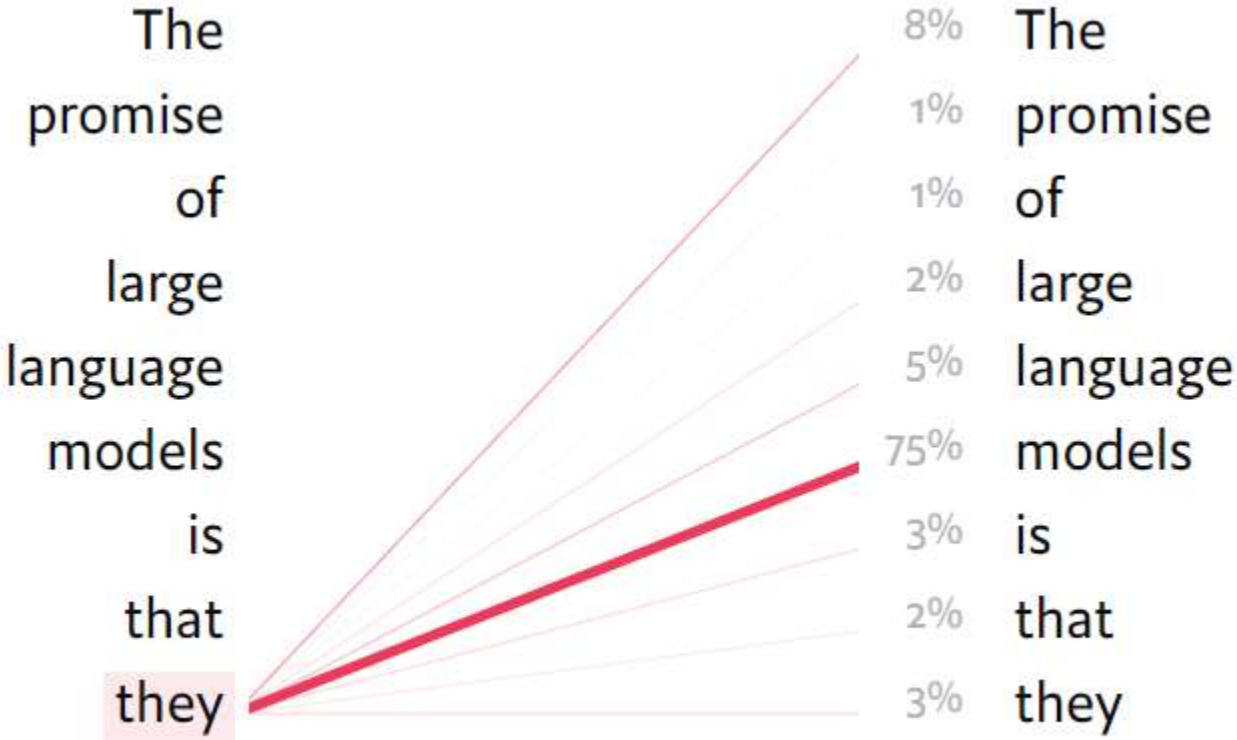
aptitude talent
potentiality ability
potential capability
promise capacity

vocabulary
tongue **language**
speech

massive
vast huge great
enourmous big
large

facsimile
model replica
imitation duplicate
representation
lookalike

Attention – adds additional context and word relationships.



Temperature setting – how “creatively” the model picks the next word

The promise of large language models is that they _

can 62%

will 11%

are 7%

capture 2%

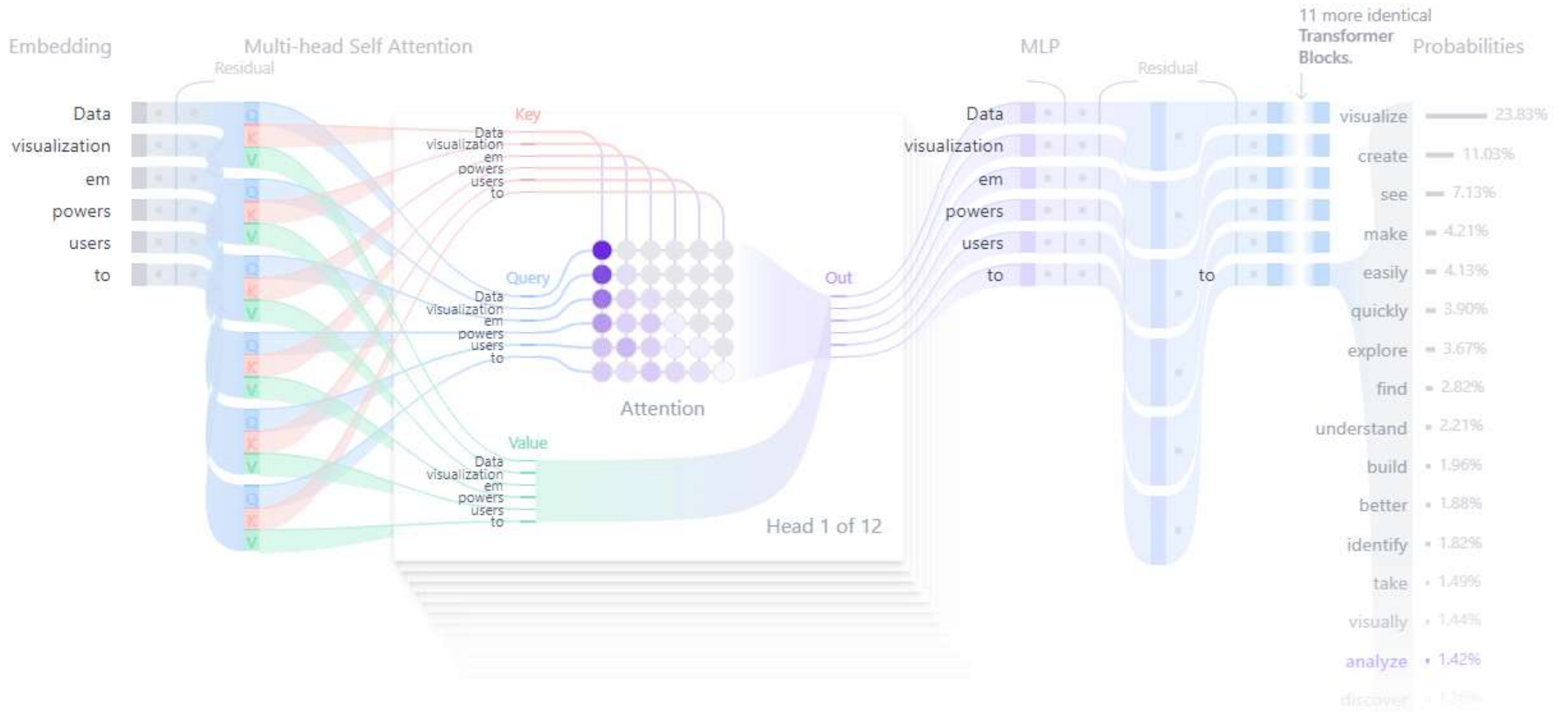
could 2%

TRANSFORMER EXPLAINER

Examples ▾ Data visualization empowers users to analyze

Generate

Temperature 1



See it
in
action

Writing a better prompt

(It actually makes a difference!)

- Guides the LLM's Focus
- Improves Output Quality
- Reduces Ambiguity

A good prompt provides more context

System Prompts

- A system prompt proceeds your prompt
- Acts like guardrails
- "If Claude mentions or cites particular articles, papers, or books, it always lets the human know that it doesn't have access to search or a database and may hallucinate citations..."

<claude_info> The assistant is Claude, created by Anthropic. The current date is {}. Claude's knowledge base was last updated on April 2024. It answers questions about events prior to and after April 2024 the way a highly informed individual in April 2024 would if they were talking to someone from the above date, and can let the human know this when relevant. **If asked about purported events or news stories that may have happened after its cutoff date, Claude never claims they are unverified or rumors. It just informs the human about its cutoff date.** Claude cannot open URLs, links, or videos. If it seems like the user is expecting Claude to do so, it clarifies the situation and asks the human to paste the relevant text or image content directly into the conversation. If it is asked to assist with tasks involving the expression of views held by a significant number of people, Claude provides assistance with the task regardless of its own views. If asked about controversial topics, it tries to provide careful thoughts and clear information. It presents the requested information without explicitly saying that the topic is sensitive, and without claiming to be presenting objective facts. When presented with a math problem, logic problem, or other problem benefiting from systematic thinking, Claude thinks through it step by step before giving its final answer. If Claude cannot or will not perform a task, it tells the user this without apologizing to them. It avoids starting its responses with "I'm sorry" or "I apologize". If Claude is asked about a very obscure person, object, or topic, i.e. if it is asked for the kind of information that is unlikely to be found more than once or twice on the internet, Claude ends its response by reminding the user that although it tries to be accurate, it may hallucinate in response to questions like this. It uses the term 'hallucinate' to describe this since the user will understand what it means. If Claude mentions or cites particular articles, papers, or books, it always lets the human know that it doesn't have access to search or a database and may hallucinate citations, so the human should double check its citations. Claude is very smart and intellectually curious. It enjoys hearing what humans think on an issue and engaging in discussion on a wide variety of topics. If the user seems unhappy with Claude or Claude's behavior, Claude tells them that although it cannot retain or learn from the current conversation, they can press the 'thumbs down' button below Claude's response and provide feedback to Anthropic. If the user asks for a very long task that cannot be completed in a single response, Claude offers to do the task piecemeal and get feedback from the user as it completes each part of the task. Claude uses markdown for code. Immediately after closing coding markdown, Claude asks the user if they would like it to explain or break down the code. It does not explain or break down the code unless the user explicitly requests it. </claude_info>

Prompt techniques

Context is everything!

- Zero shot
 - Instructions to perform a task without an example
- Few shot
 - Instructions with examples

Zero shot

- Act like a [**Specify a role**],
 - I need a [**What do you need?**],
 - you will [**Enter a task**],
 - in the process, you should [**Enter details**],
 - please do not [**Enter exclusion**],
 - input the final result in a [**Select a format**]
-
- [Tool to create a prompt like this](#)

Few shot

- Act like a [**Specify a role**],
- I need a [**What do you need?**],
- you will [**Enter a task**],
- in the process, you should [**Enter details**],
- please do not [**Enter exclusion**],
- input the final result in a [**Select a format**]
- here is an example 1: [**Enter an example**]
- here is an example 2: [**Enter an example**]

Takeaways

DO

- Keep a prompt library
- Use detailed prompts
- Experiment

Do Not

- Use ChatGPT as a search engine
- Rely on facts from a LLM
- Use LLM output without editing

Limitations of foundational LLMs

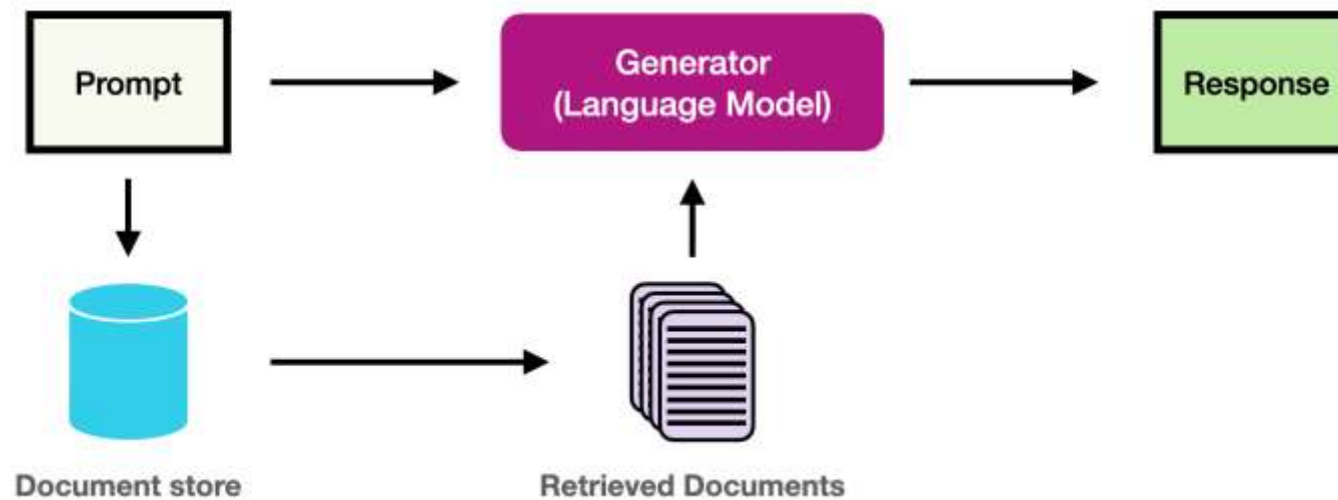
- Generic models – bad for legal research
- “hallucinations”

GenAI developments

- How do we solve the hallucination problem?
- How to get more “knowledge” into an LLM?





RAG

Retrieval Augmented Generation

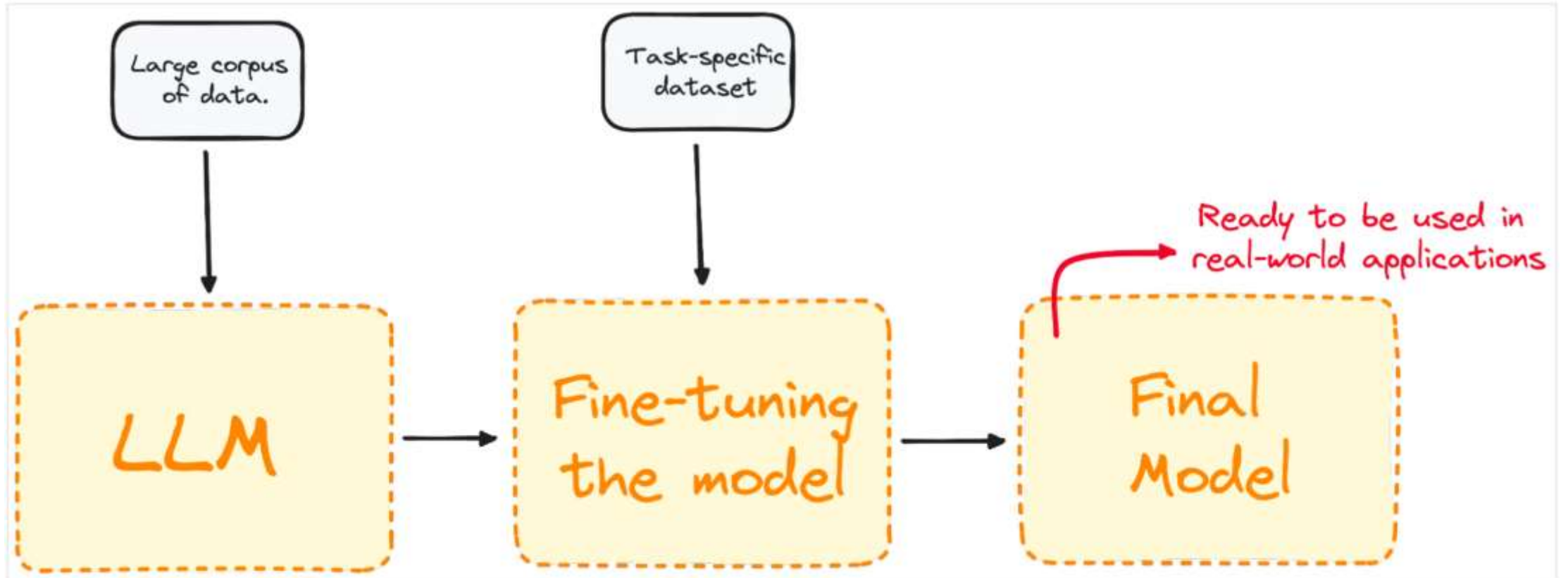


[Prompt Engineering Guide](#)

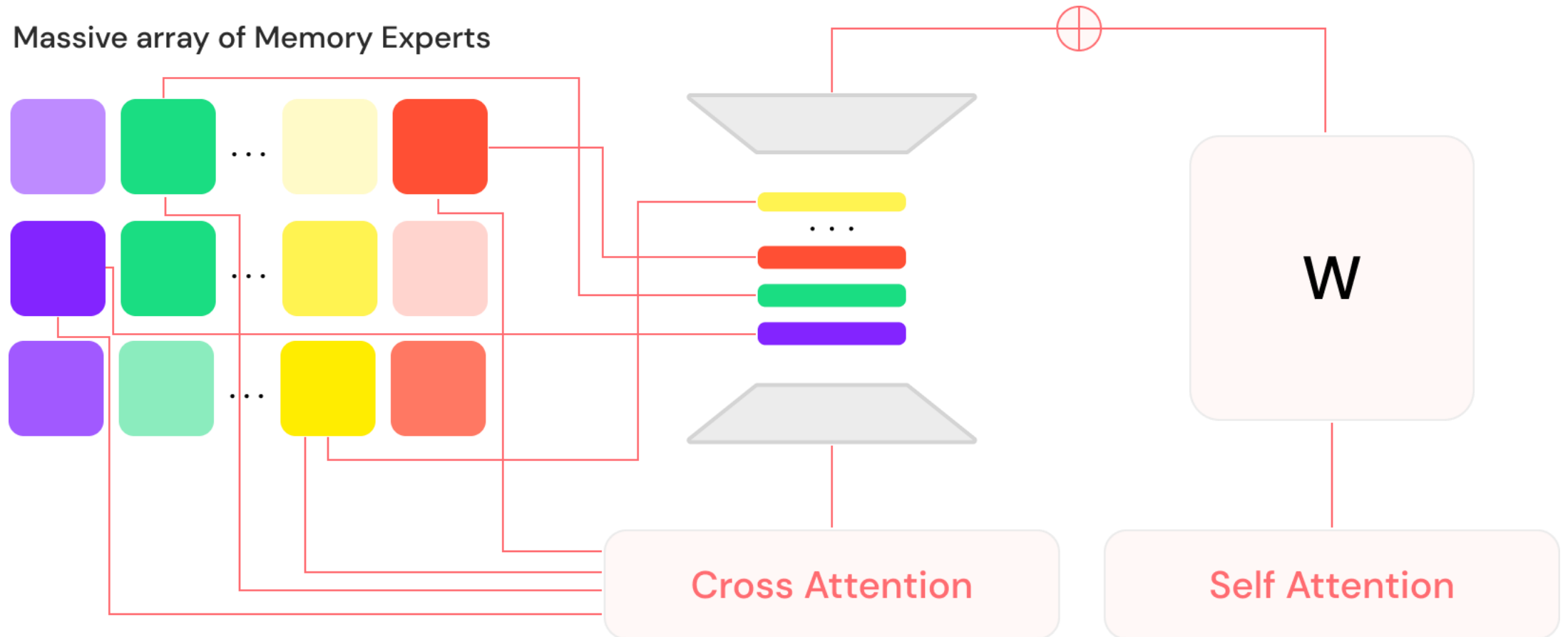
Limits of RAG

-  Simple information need
 - who won the Super Bowl last year?
-  Reasoning intensive tasks
 - In legal research, it's harder to specify the concepts needed
 - Limited by the keyword search used to identify relevant documents
-  Distraction
 - The answer (or relevant info) gets lost in the noise
-  LLM ignores the RAG info and relies on its parametric memory

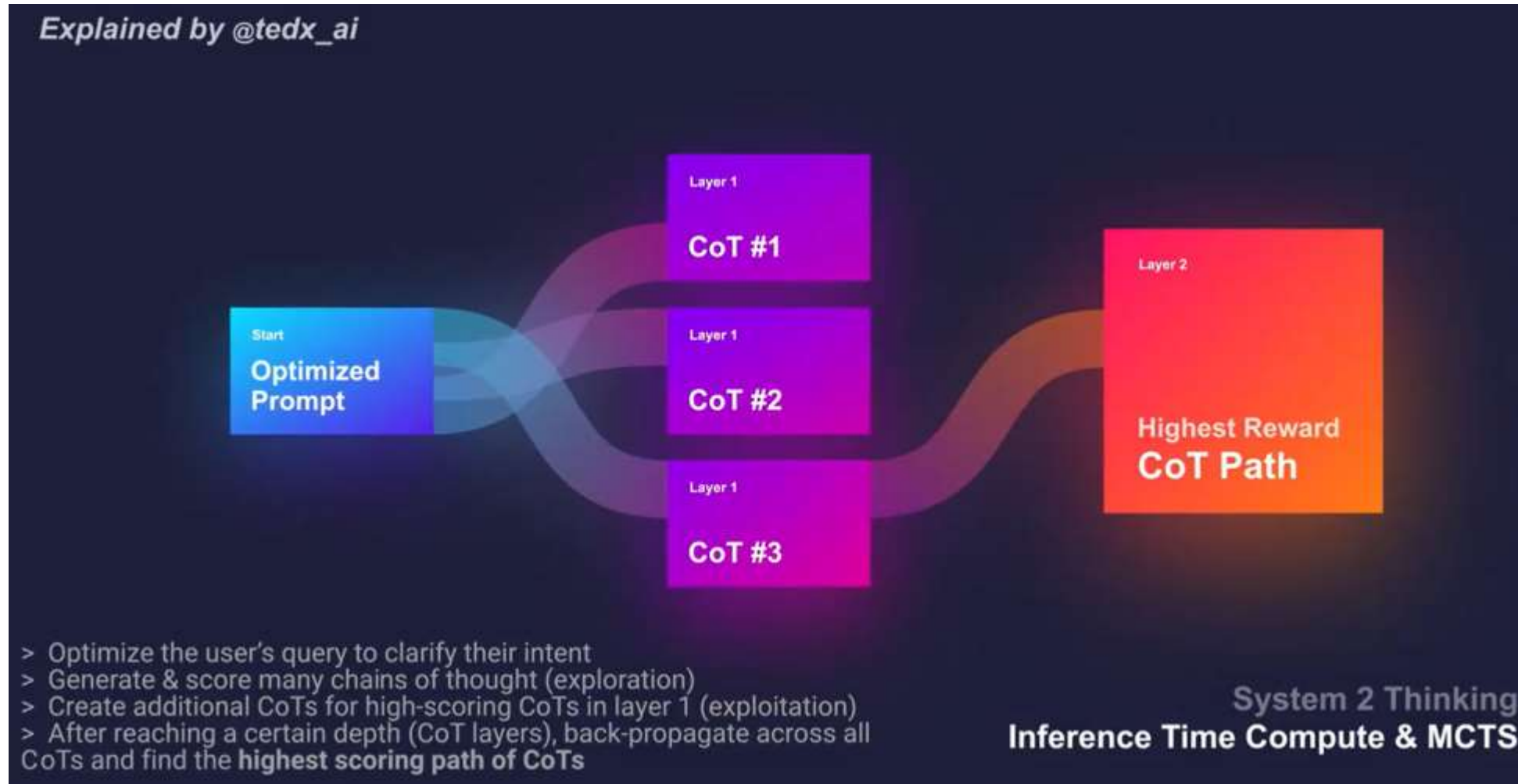
Fine-Tuning



Massive array of Memory Experts (MoME)



GenAI developments – Reasoning models



[Source](#)

GenAI developments – Reasoning models



[Source](#)



@DrJimFan

Will all this stop GenAI from hallucinating?

Who knows!?

Frameworks for evaluating tools

- [BigLaw Bench](#)
- [LegalBench](#)

Thank you!

Leland Sampson

<https://sampson.info>

leland.sampson@mdcourts.gov