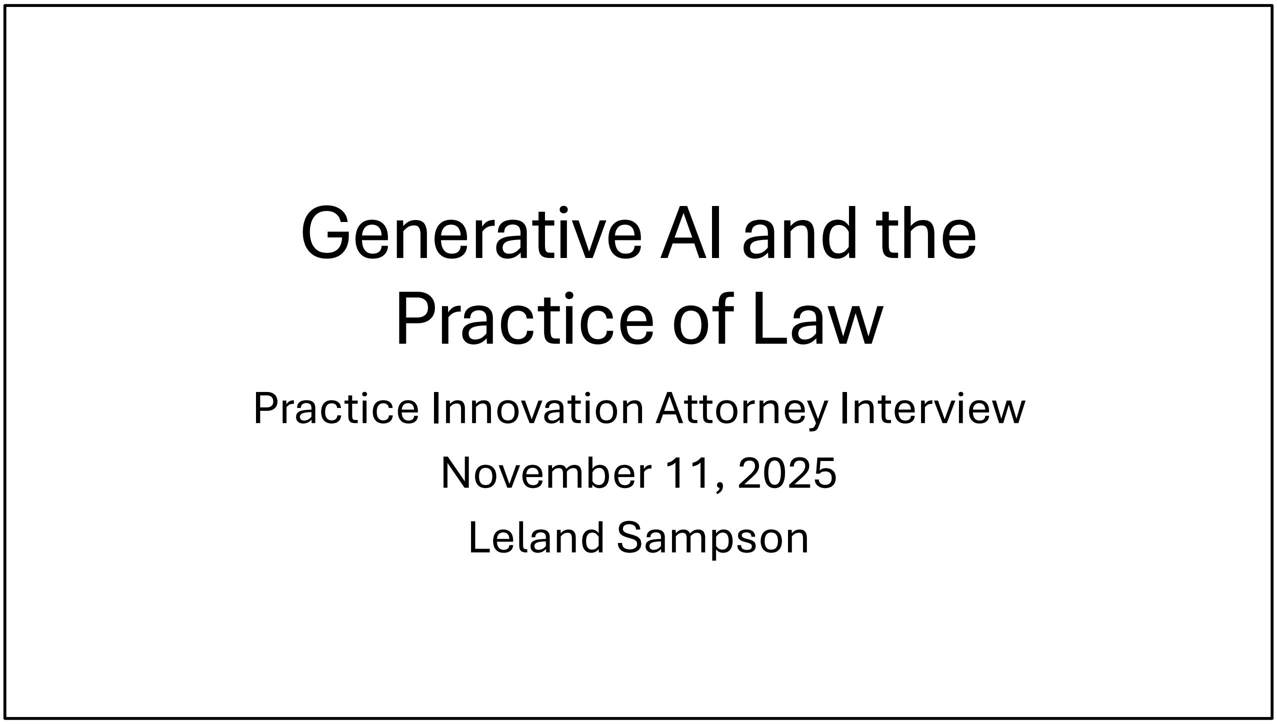


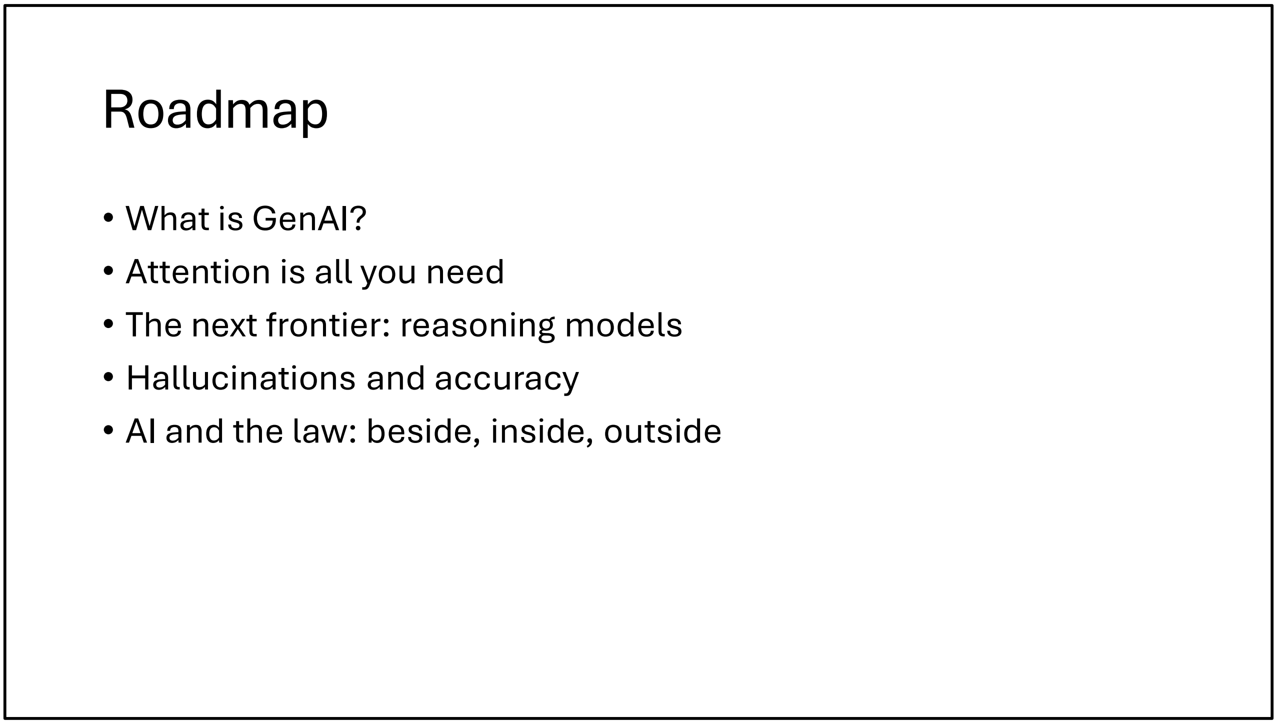
Generative AI and the Practice of Law

Practice Innovation Attorney Interview

November 11, 2025

Leland Sampson

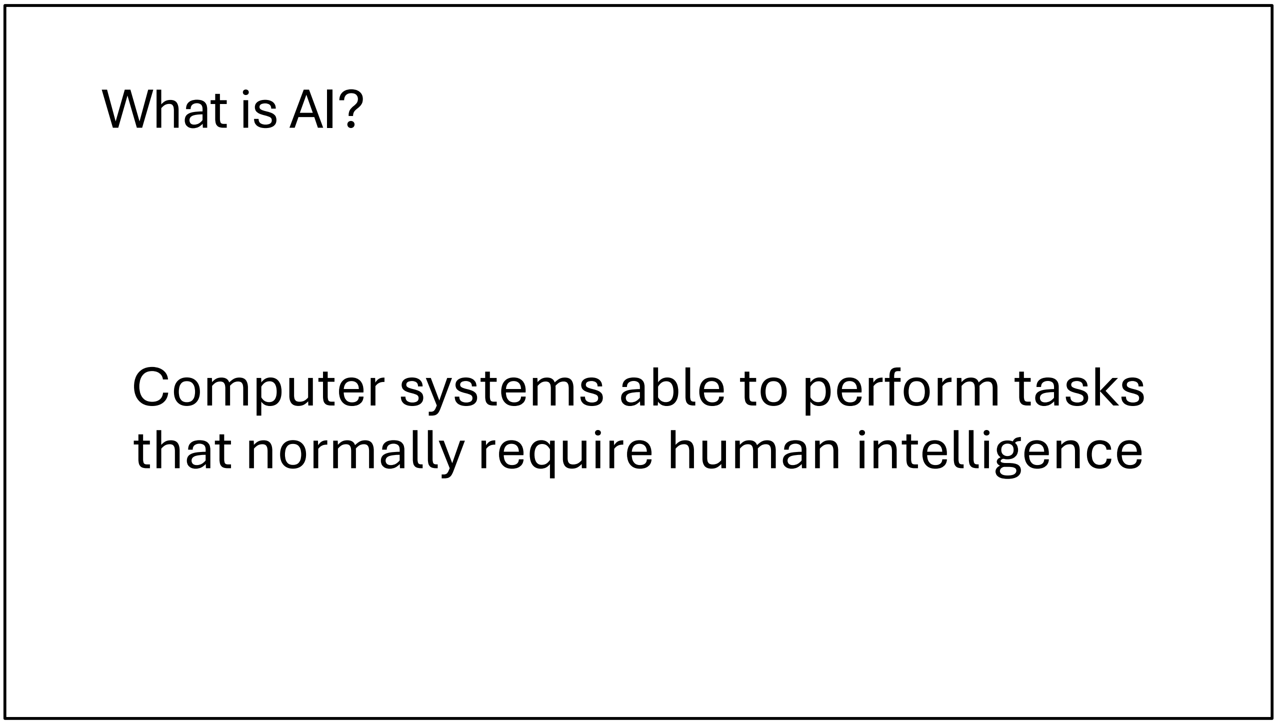
Hi, my name is Leland Sampson. Thank you for joining me today for a presentation about generative AI and the practice of law.



Roadmap

- What is GenAI?
- Attention is all you need
- The next frontier: reasoning models
- Hallucinations and accuracy
- AI and the law: beside, inside, outside

This presentation touches on several topics in in abbreviated form. I will distinguish generative AI from other types of AI, describe the research paper that sparked classical GPTs, briefly describe the reasoning models that are the current state of the art, discuss hallucinations and accuracy throughout, and wrap up with why I think every lawyer needs to pay attention to AI

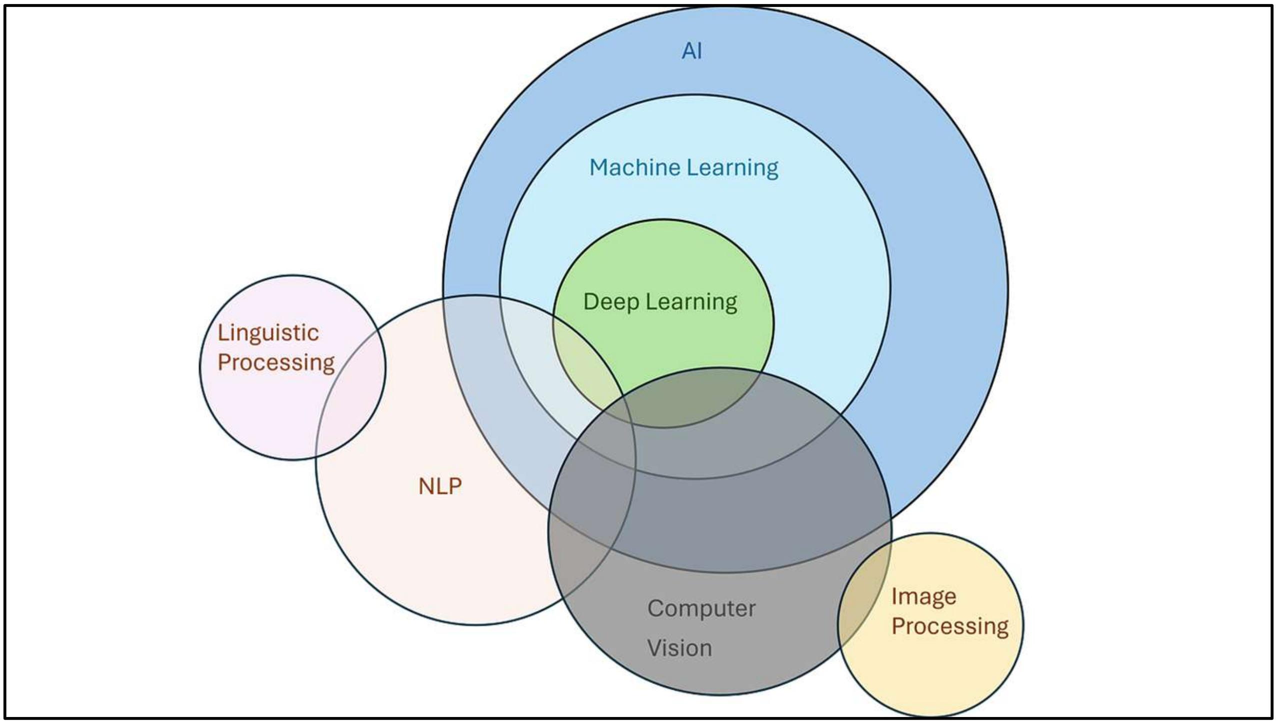


What is AI?

Computer systems able to perform tasks that normally require human intelligence

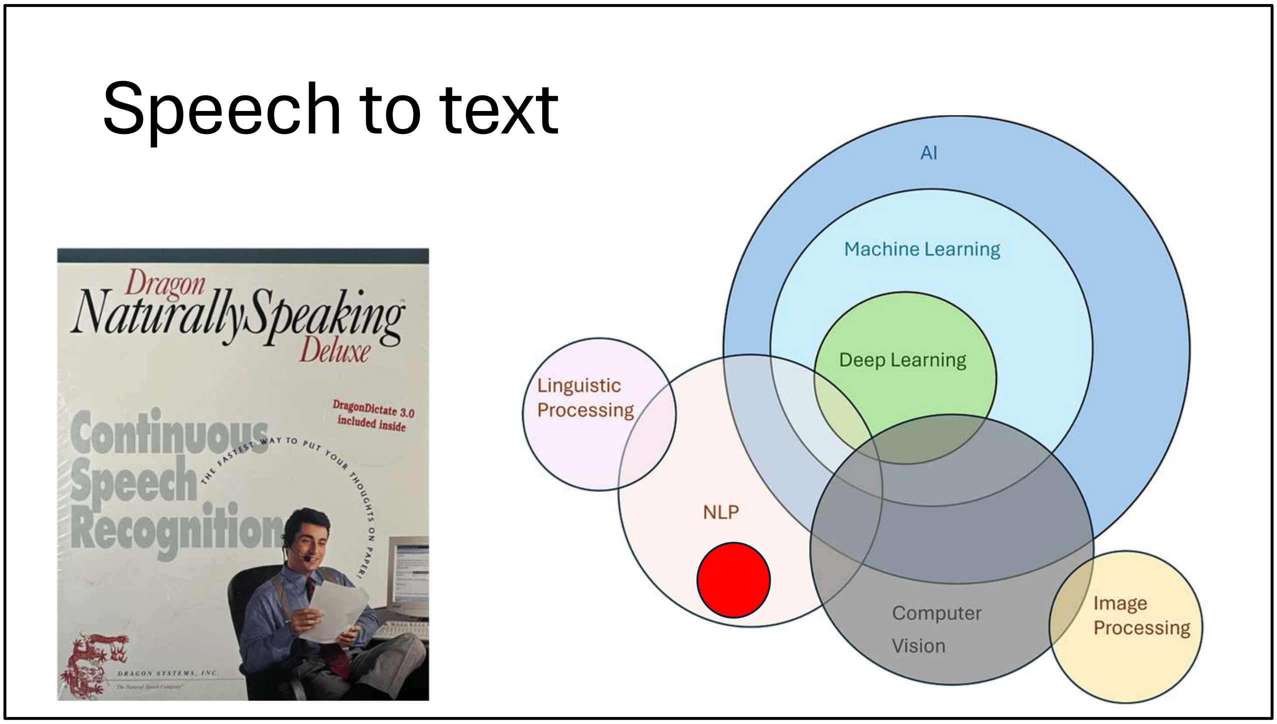
So what is AI? Broadly defined, AI is the ability of computer systems to perform tasks that normally require human intelligence.

https://www.oxfordreference.com/display/10.1093/acref/9780198609810.001.0001/acref-9780198609810-e-423



This diagram provides a good visual that categorizes computer tasks that we would normally associate with human intelligence. Let’s look at some examples.

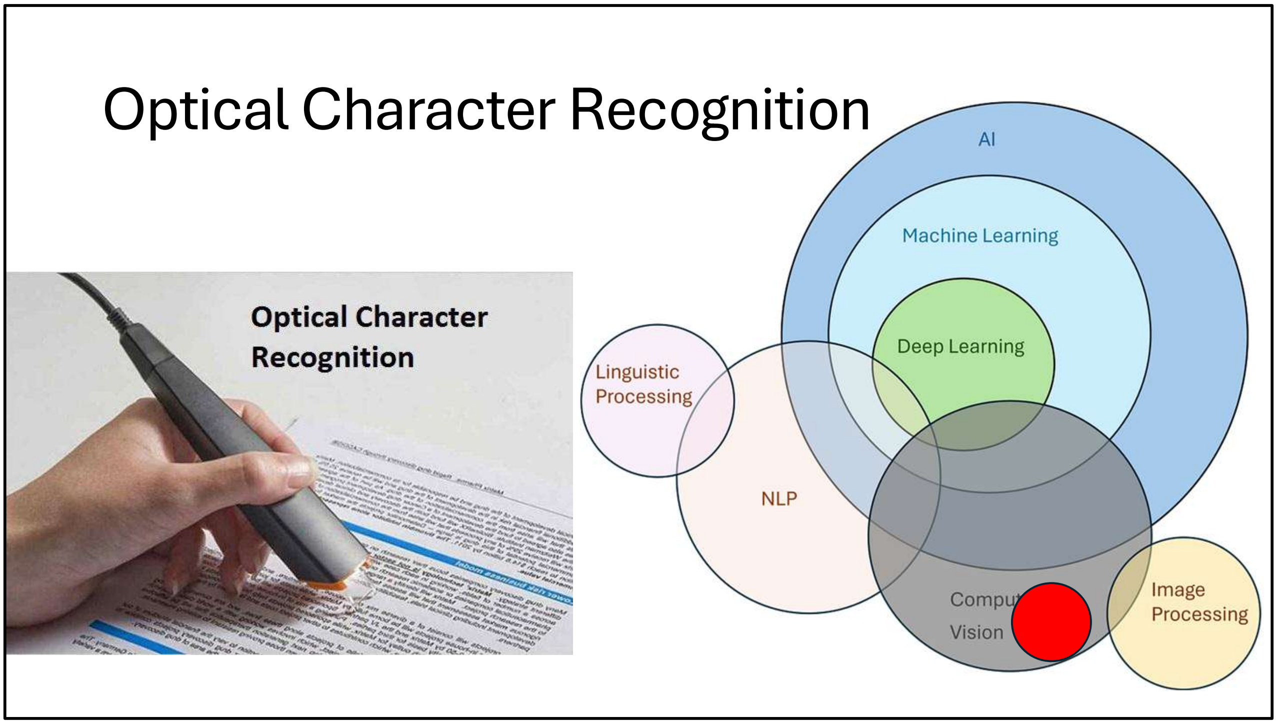
https://medium.com/@jainpalak9509/breakdown-simplify-ai-ml-nlp-deep-learning-computer-vision-c76cd982f1e4



Speech to text

Consider a natural language processing app that does text to speech, such as Dragon Naturally Speaking. This software is deterministic. For the same input, the user would get the same output.

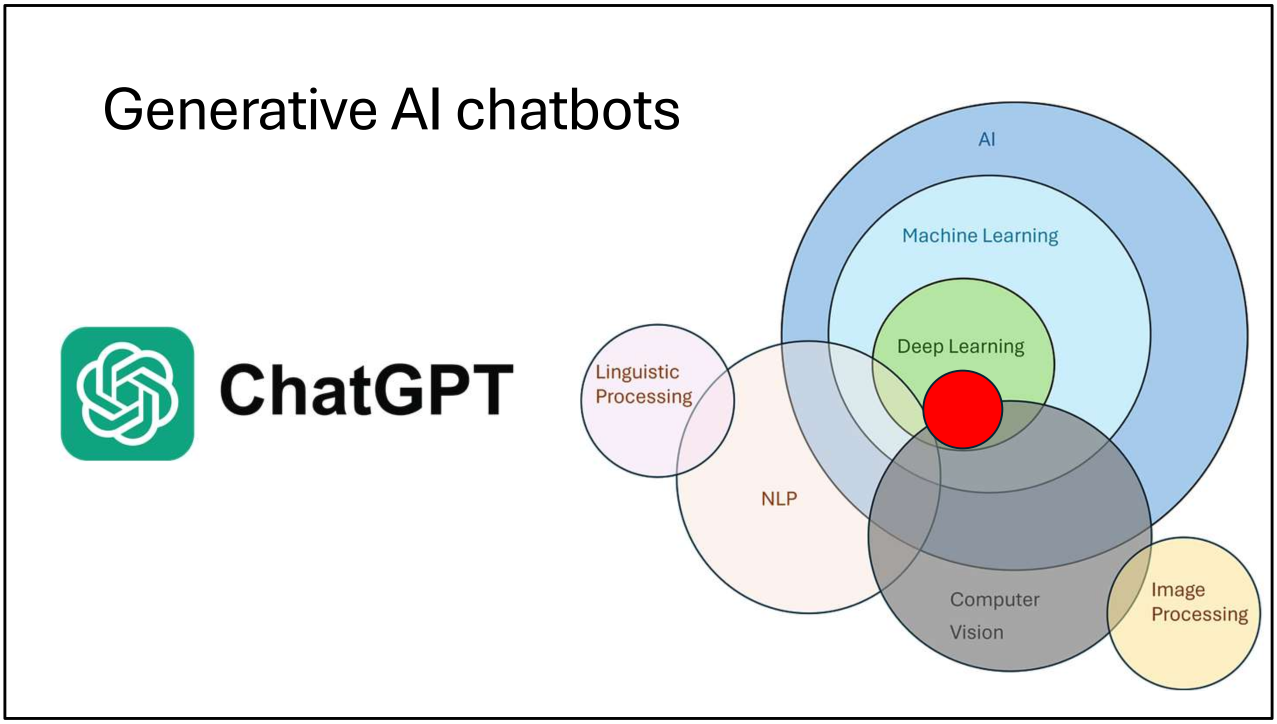
https://www.ebay.com/itm/326201828484?chn=ps&google_free_listing_action=view_item



Optical Character Recognition

The same for optical character recognition. This form of computer vision can be done using algorithms and heuristics with predictable outputs.

https://www.openpr.com/news/2057056/optical-character-recognition-ocr-market-survey-report-2020

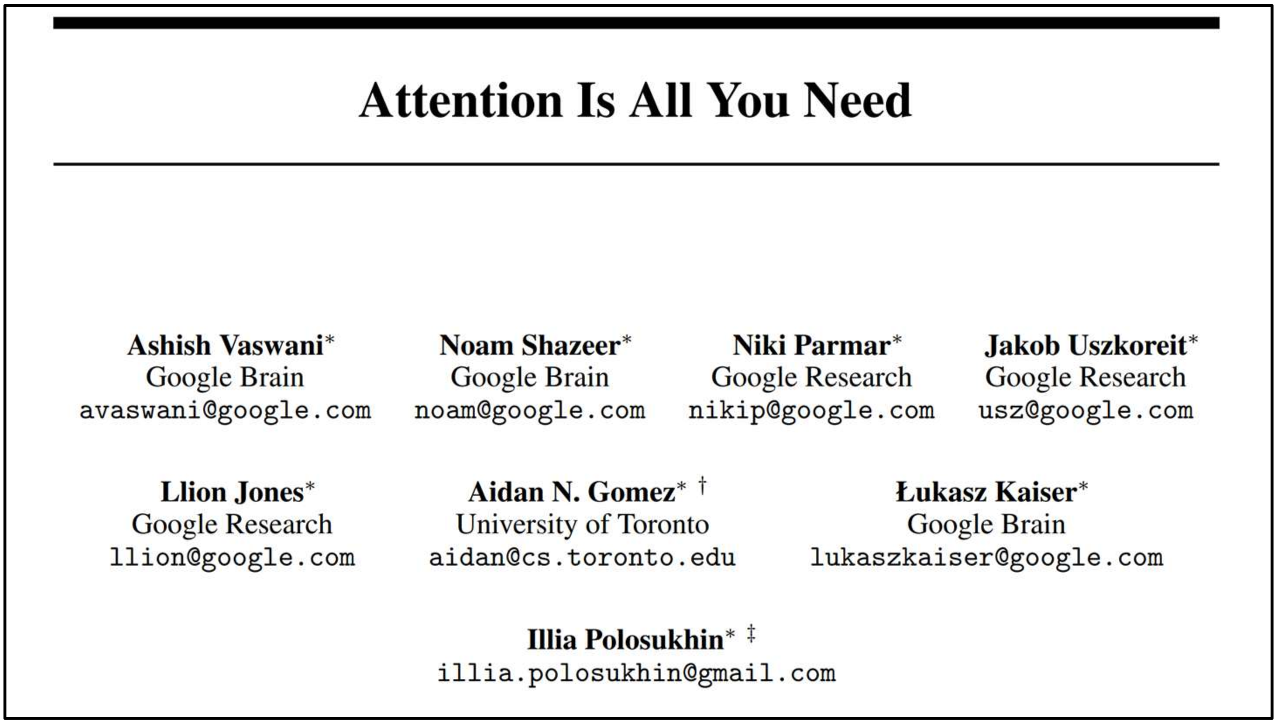


Generative AI chatbots

ChatGPT

https://www.openpr.com/news/2057056/optical-character-recognition-ocr-market-survey-report-2020

Chatbots like ChatGPT are examples of deep learning AI where, the output is probabilistic and the same input may yield different outputs.



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

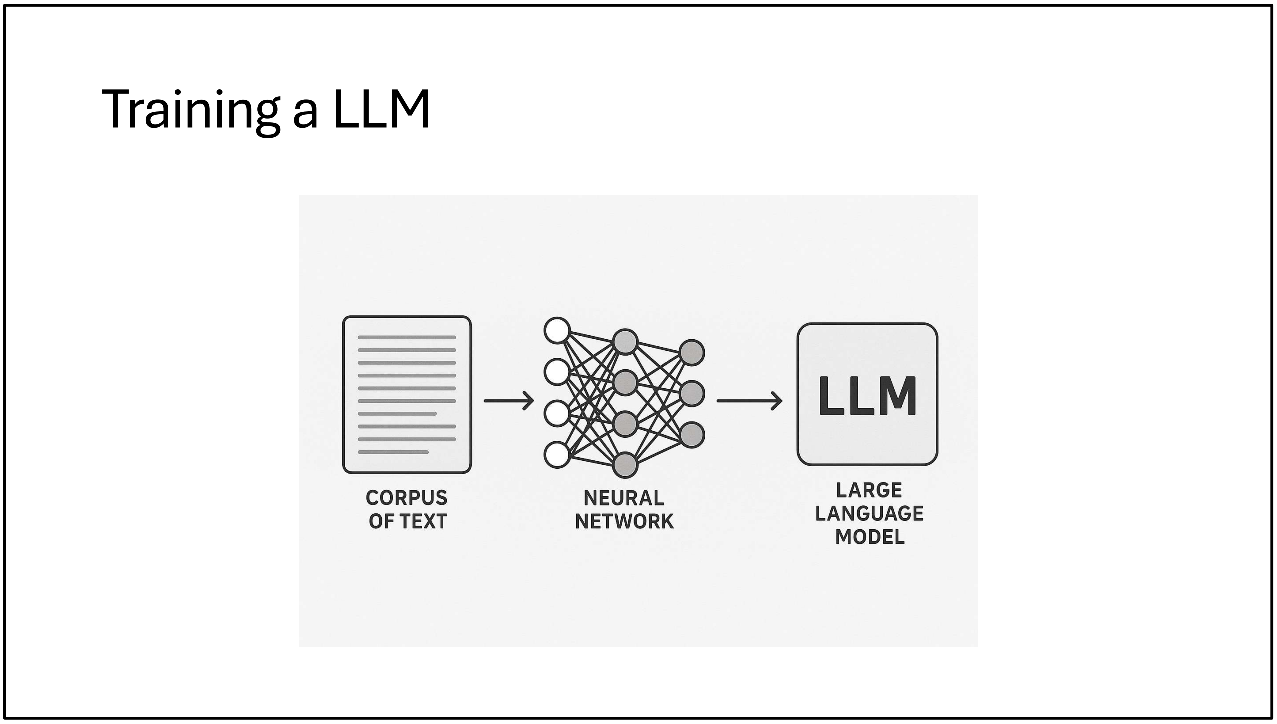
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

AI research is older than the transistor. This is the paper that proposed the Transformer in 2017 and sparked the current AI summer..

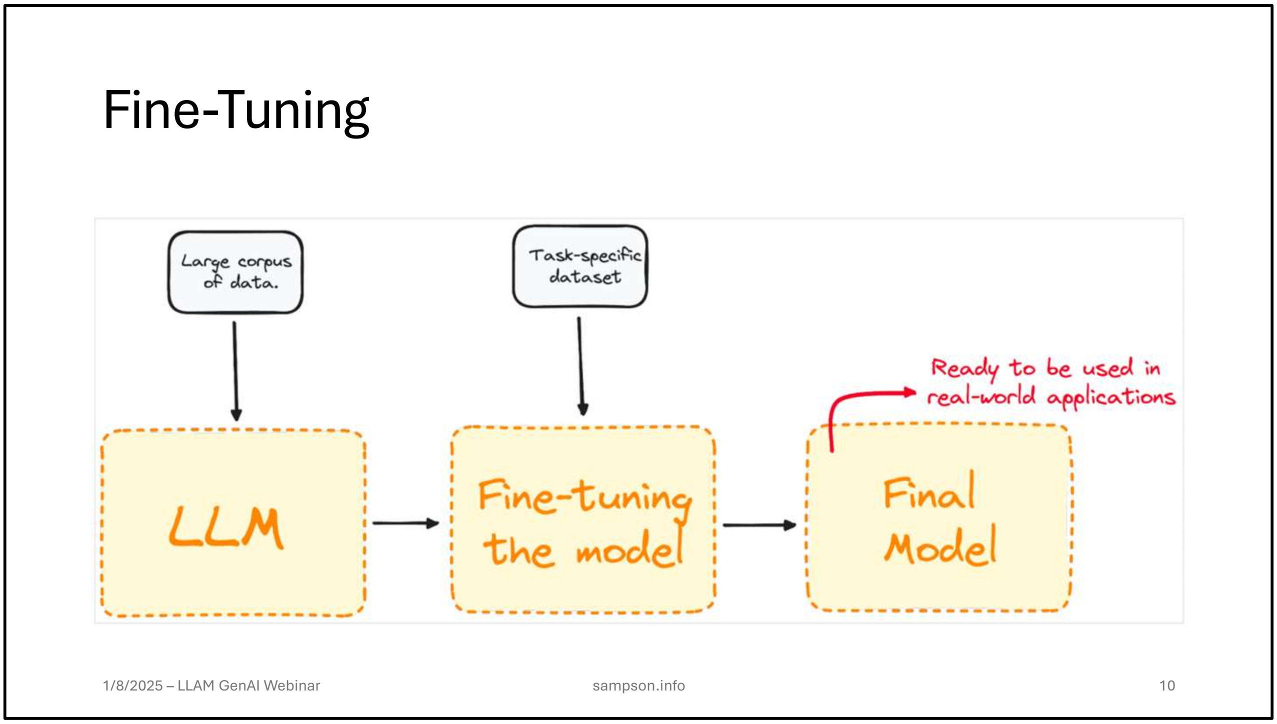
https://arxiv.org/pdf/1706.03762



Training a LLM

Using the Transformer described in the paper, computer neural networks were built by analyzing trillions of associations between words. The weights associated with these relationships became known as large language models or LLMs for short.

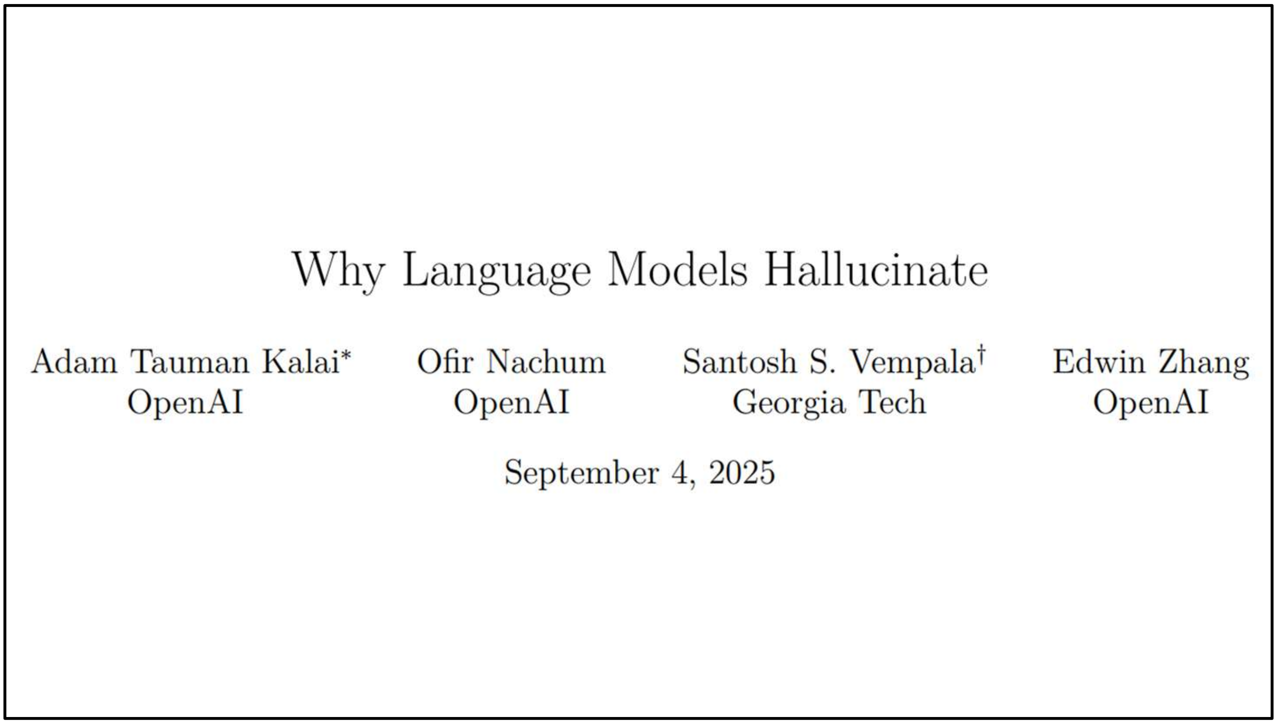
Image generated by ChatGPT



Fine-Tuning

All LLMs are fine-tuned before they are used. The response to a prompt is scored by a human and the LLM tries again. The scoring affects the model weights and influences all future outputs.

https://www.datacamp.com/tutorial/fine-tuning-large-language-models



Why Language Models Hallucinate

Adam Tauman Kalai[*] OpenAI
Ofir Nachum OpenAI
Santosh S. Vempala[†] Georgia Tech
Edwin Zhang OpenAI

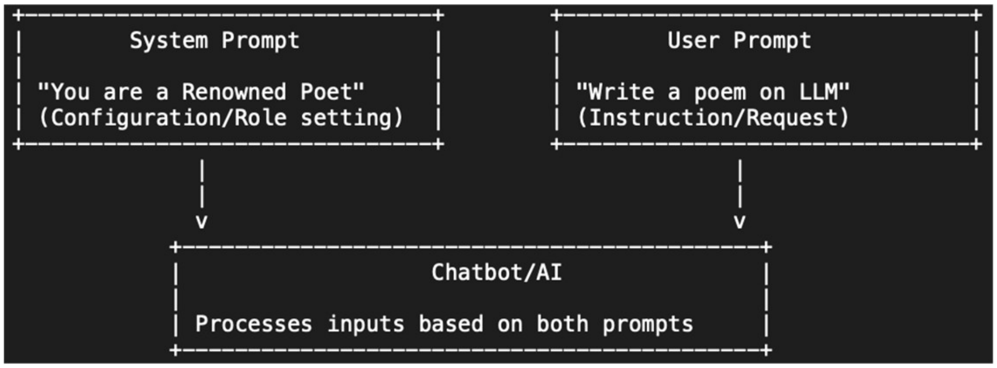
September 4, 2025

Another great research paper on why LLMs hallucinate suggests that the problem is the incentives of benchmarking and fine tuning. The way benchmarks work, the LLM is rewarded for guessing and so it scores higher, but at the cost of inaccuracies like hallucinations.

https://arxiv.org/pdf/2509.04664

System prompts

- Set of instructions, guidelines, and additional contextual information that defines how LLM responds to user prompt.



```
+--------------------------------+       +--------------------------------+
|         System Prompt          |       |          User Prompt           |
|                                |       |                                |
| "You are a Renowned Poet"      |       | "Write a poem on LLM"          |
| (Configuration/Role setting)   |       | (Instruction/Request)          |
+--------------------------------+       +--------------------------------+
              |                                        |
              |                                        |
              v                                        v
            +--------------------------------------------+
            |                 Chatbot/AI                 |
            |                                            |
            | Processes inputs based on both prompts     |
            +--------------------------------------------+
```

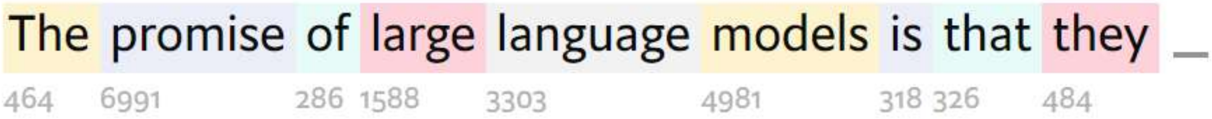
System prompts are sets of instructions and guardrails that tell the LLM how to respond to a user prompt. As an example, Anthropic publishes the system prompt for their chatbot Claude and the system prompt has a guardrail that says Claude “does not provide information that could be used to make chemical or biological or nuclear weapons.

https://promptengineering.org/system-prompts-in-large-language-models/

https://generativeai.pub/understanding-system-and-normal-prompts-in-llm-how-they-work-and-differ-df9ea490546c

https://docs.claude.com/en/release-notes/system-prompts

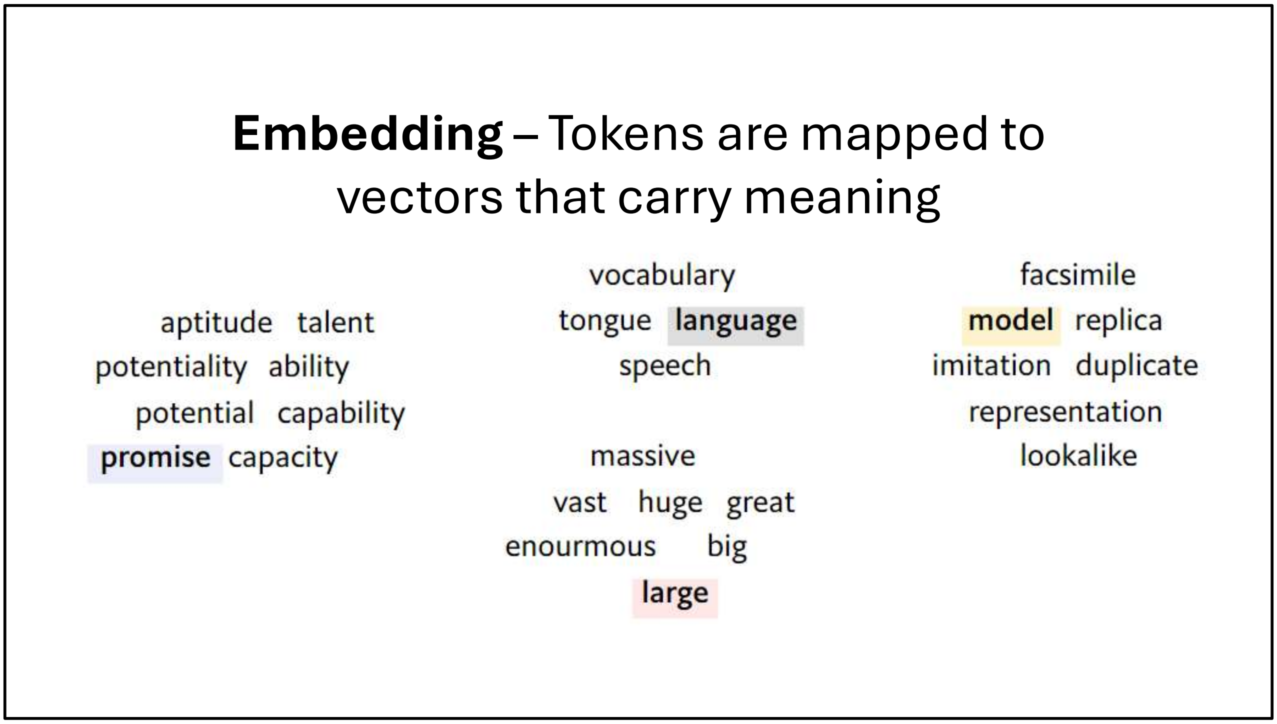
Tokenisation – Turning words (or fragments of words) into numbers



The	promise	of	large	language	models	is	that	they	_
464	6991	286	1588	3303	4981	318	326	484	

Now let’s look at how a LLM picks the next word in an output. First, words are transformed into tokens.

Visual taken from https://www.economist.com/interactive/science-and-technology/2023/04/22/large-creative-ai-models-will-transform-how-we-live-and-work



Embedding – Tokens are mapped to vectors that carry meaning

aptitude talent
potentiality ability
potential capability
promise capacity

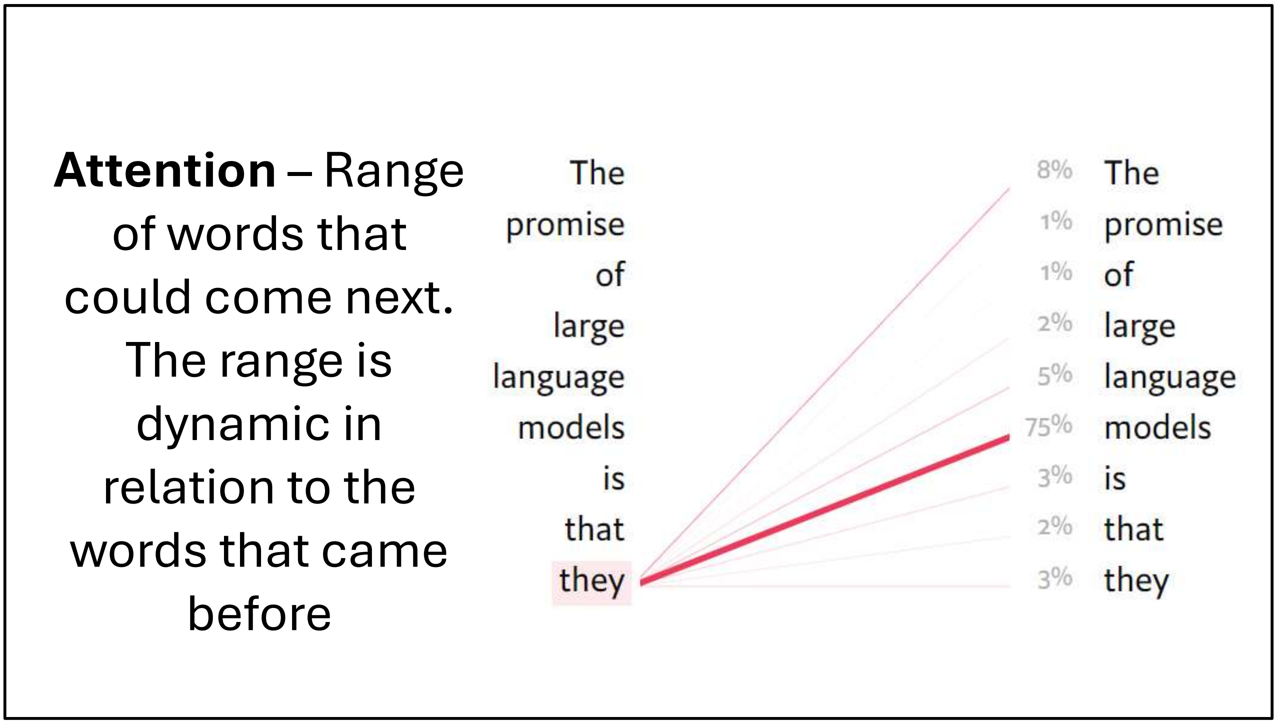
vocabulary
tongue **language**
speech

massive
vast huge great
enourmous big
large

facsimile
model replica
imitation duplicate
representation
lookalike

Next the tokens are assigned a vector that comes from the training set. This begins the process of teasing out meaning, Similar words have similar vectors. it’s called embedding.

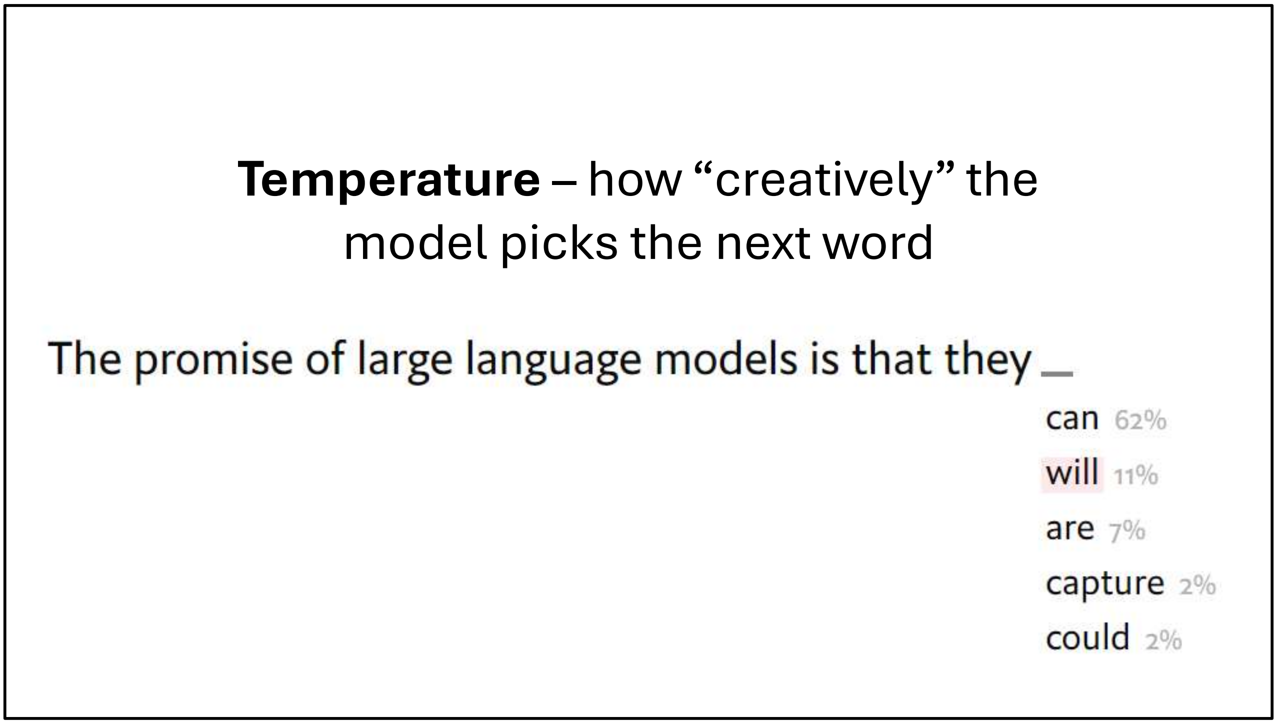
Visual taken from https://www.economist.com/interactive/science-and-technology/2023/04/22/large-creative-ai-models-will-transform-how-we-live-and-work



Attention – Range of words that could come next. The range is dynamic in relation to the words that came before

Attention was the innovation of the 2017 paper and it allows the LLM to capture meaning and relationships even between words that are far apart in a text. LLMs do the same attention task in parallel looking at different parts of the text. It’s the parallel nature of attention has made nVidia a 5 trillion dollar company.

Visual taken from https://www.economist.com/interactive/science-and-technology/2023/04/22/large-creative-ai-models-will-transform-how-we-live-and-work



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

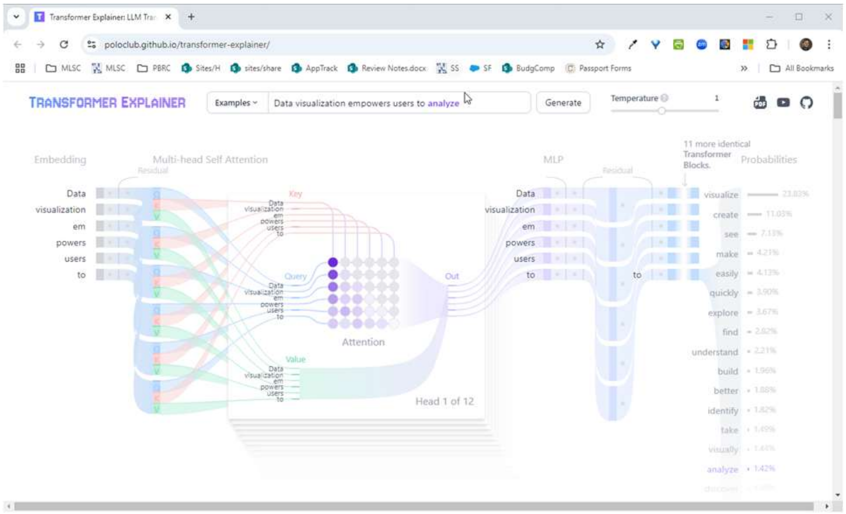
The promise of large language models is that they _

can	62%
will	11%
are	7%
capture	2%
could	2%

There are many nobs to adjust in an LLM, but temperature is one of the easiest for users to understand and tweak. Increasing the temperature will increase the randomness and result in more creativity at the cost of more hallucinations.

Visual taken from https://www.economist.com/interactive/science-and-technology/2023/04/22/large-creative-ai-models-will-transform-how-we-live-and-work

See it in action



poloclub.github.io/transformer-explainer/

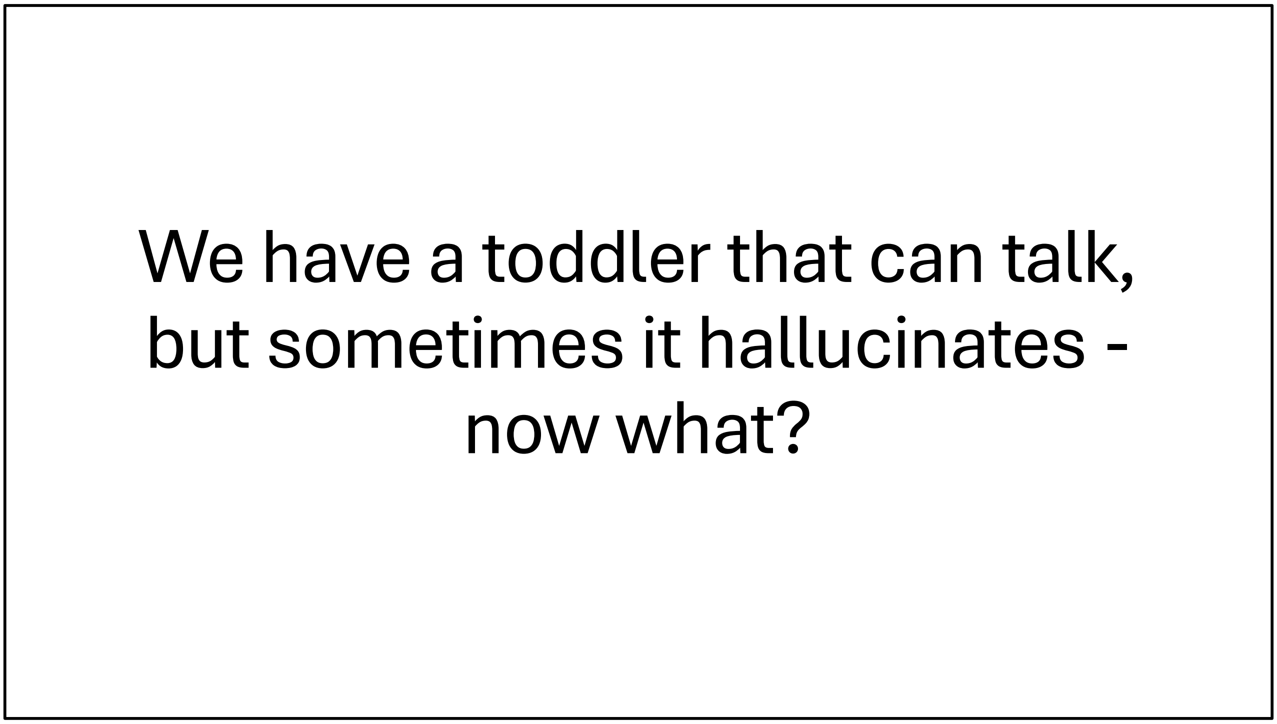
Examples ˅ Data visualization empowers users to analyze

Temperature 1

visualize	23.83%
create	11.03%
see	7.13%
make	4.21%
easily	4.13%
quickly	3.90%
explore	3.67%
find	2.82%
understand	2.21%
build	1.96%
better	1.88%
identify	1.82%
take	1.49%
visually	1.44%
analyze	1.42%

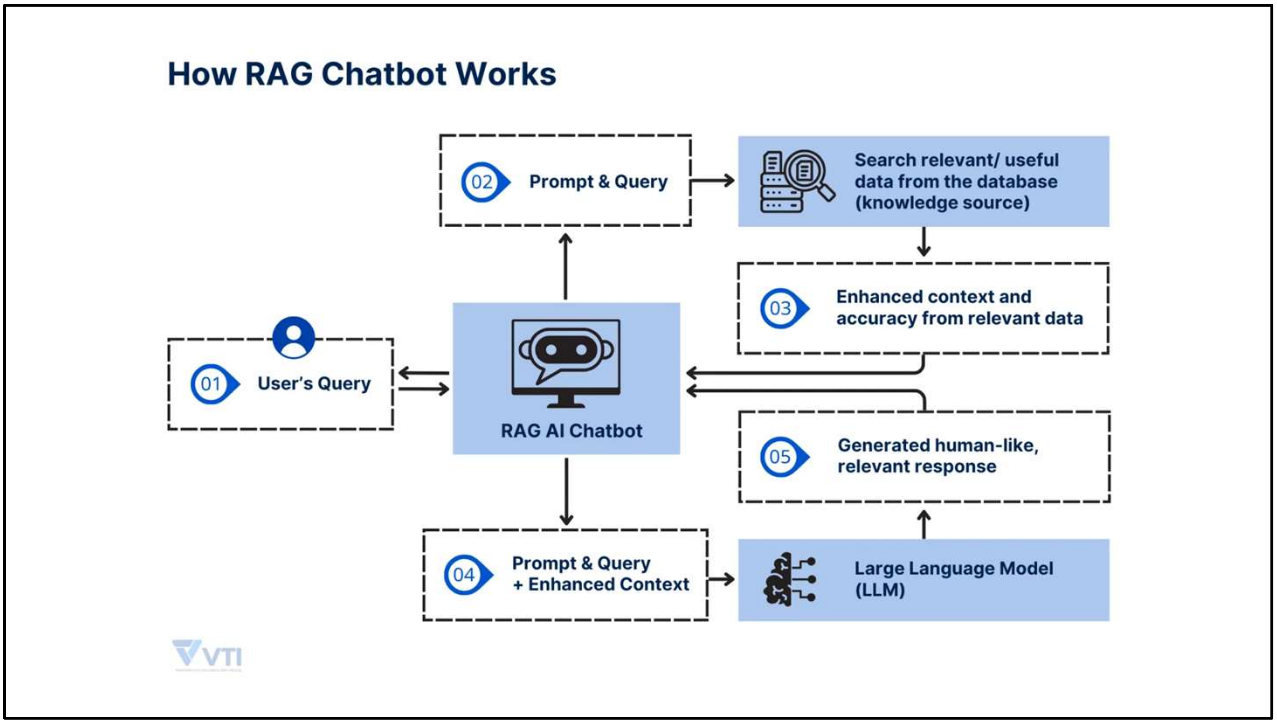
This website uses the GPT 2 model to visually show you how the LLM is generating the next word in a sentence. I suggest you play with it, it’s very fun.

https://poloclub.github.io/transformer-explainer/



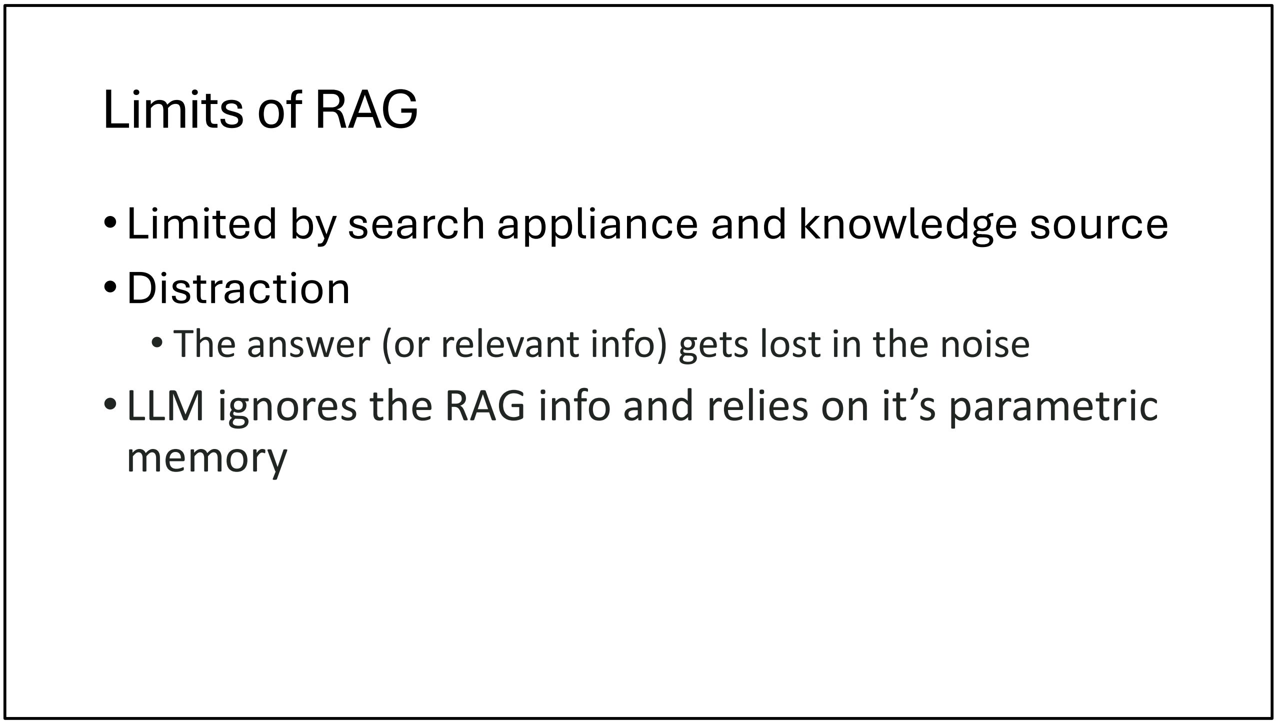
We have a toddler that can talk, but sometimes it hallucinates - now what?

So this bot can speak coherently, but what does that get you? It only “knows” things from its training data and the temperature that makes it creative can make things up. How to we get knowledge into the bot?


How RAG Chatbot Works
02 Prompt & Query
Search relevant/ useful data from the database (knowledge source)
03 Enhanced context and accuracy from relevant data
01 User's Query
RAG AI Chatbot
05 Generated human-like, relevant response
04 Prompt & Query + Enhanced Context
Large Language Model (LLM)
VTI

One common approach is Retrieval Augmented Generation. Here, the user's prompt first goes to a search tool that finds relevant information in a knowledgebase. The original prompt and context taken from the knowledgebase then go to the LLM for processing and the user gets an output.

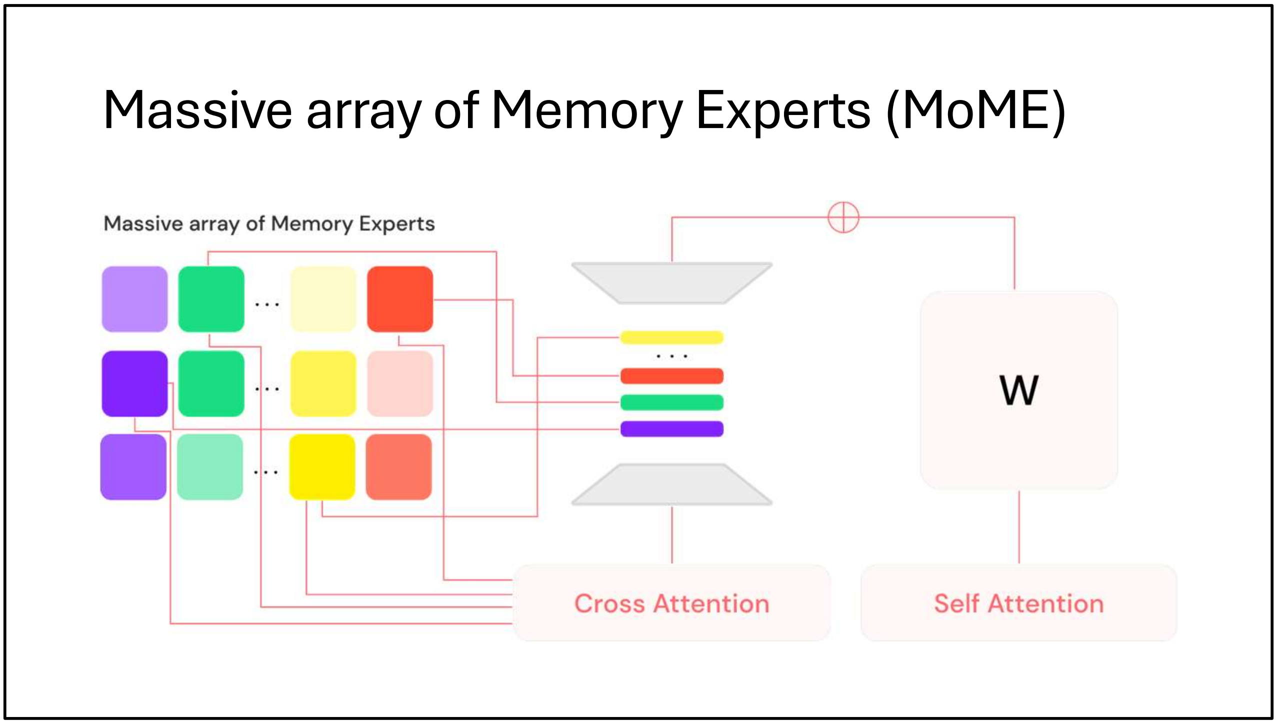
https://vti.com.vn/a-comprehensive-guide-to-rag-genai



Limits of RAG

- Limited by search appliance and knowledge source
- Distraction
  - The answer (or relevant info) gets lost in the noise
- LLM ignores the RAG info and relies on it’s parametric memory

RAG can be useful, but the user quickly discovers the limits. If the front end search or knowledge base are poor, outputs will likewise be useless. If the answer is one small needle in the haystack the LLM might not find it for the output. Also, the randomness of the LLM can sometimes cause it to ignore the knowledge source and use the output from it’s parametric memory that comes from it’s training data.

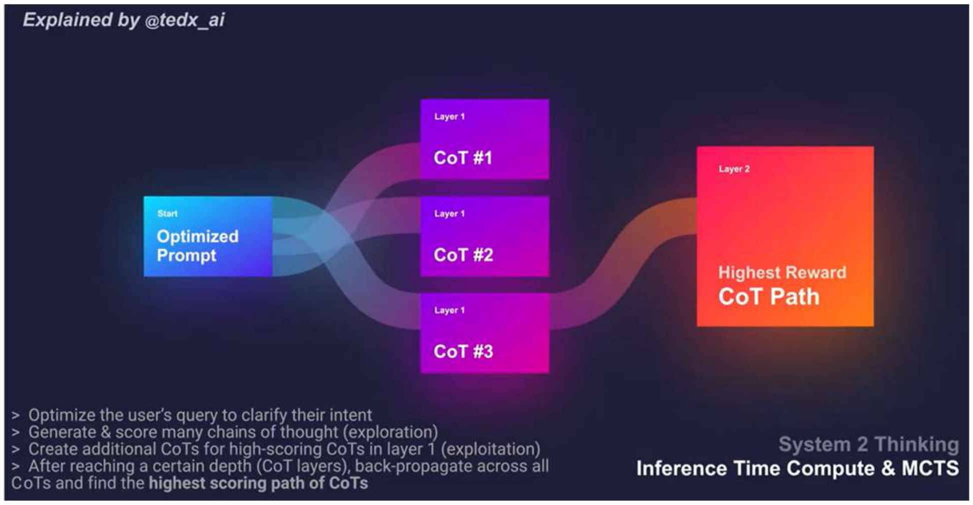


Massive array of Memory Experts (MoME)

The massive array of memory experts is another attempt to improve outputs. Here an LLM has thousands of small SLMs fine tuned on specific areas of knowledge. For example there may be model trained on the UCC statutory text, another on cases involving the UCC, etc. Here, the LLM is an orchestrator that picks which SML is appropriate for a subset of the prompt. The SLM generates an output and those tokens are used to increase the context of the original prompt. Eventually the orchestrating LLM provides a response.

https://arxiv.org/html/2406.17642v1

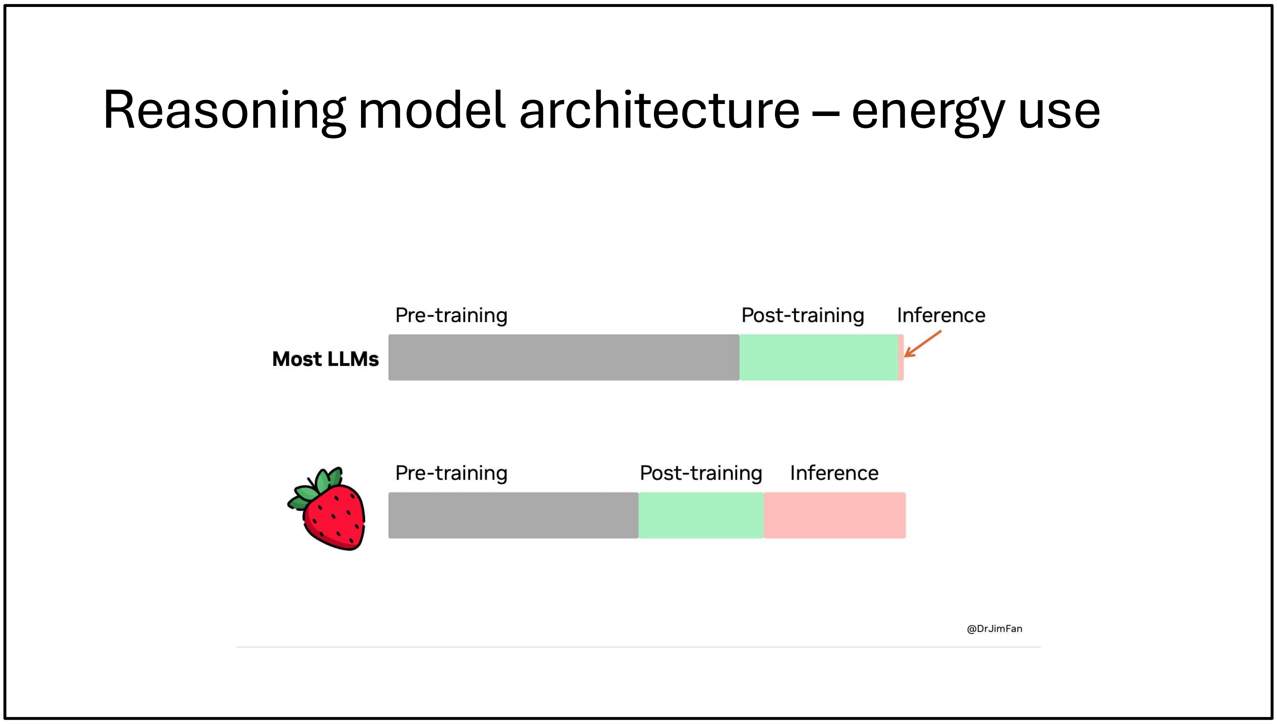
Reasoning model architecture



> Optimize the user's query to clarify their intent
> Generate & score many chains of thought (exploration)
> Create additional CoTs for high-scoring CoTs in layer 1 (exploitation)
> After reaching a certain depth (CoT layers), back-propagate across all CoTs and find the **highest scoring path of CoTs**

Reasoning models are the next frontier. This approach breaks prompts down into chains of thought where the LLM steps through the response many times. Another part of the LLM evaluates the chains of thought and picks the best one. When the models say the are “thinking”this is what’s going on in the background.

https://x.com/tedx_ai/status/1834346066482987340

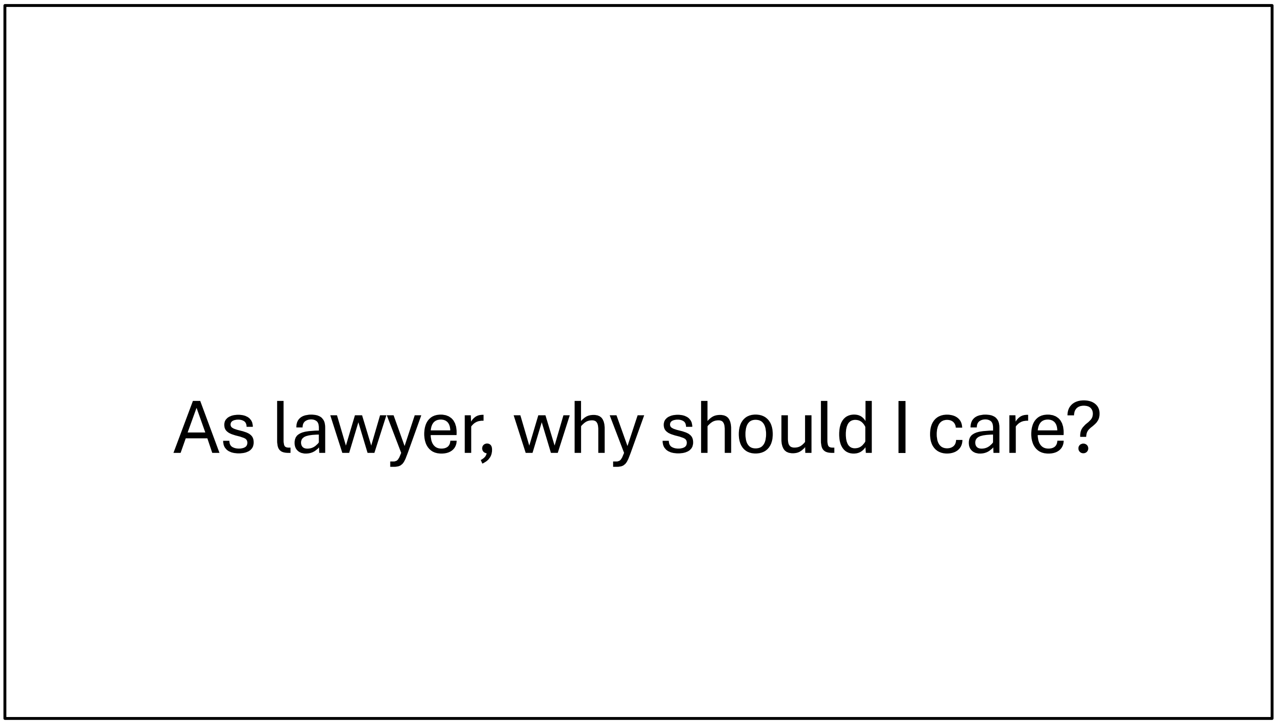


Reasoning model architecture – energy use

The reasoning models are using compute resources when the prompt is processed. Classic LLM models take a lot of resources to train but then the marginal cost to generate tokens is minimal. The reasoning models use significant amounts of processing power to produce and evaluate the chains of thought. And this means the marginal cost of generating outputs is a lot higher.

This energy cost is why big tech companies are leasing nuclear reactors to fuel their AI datacenters.

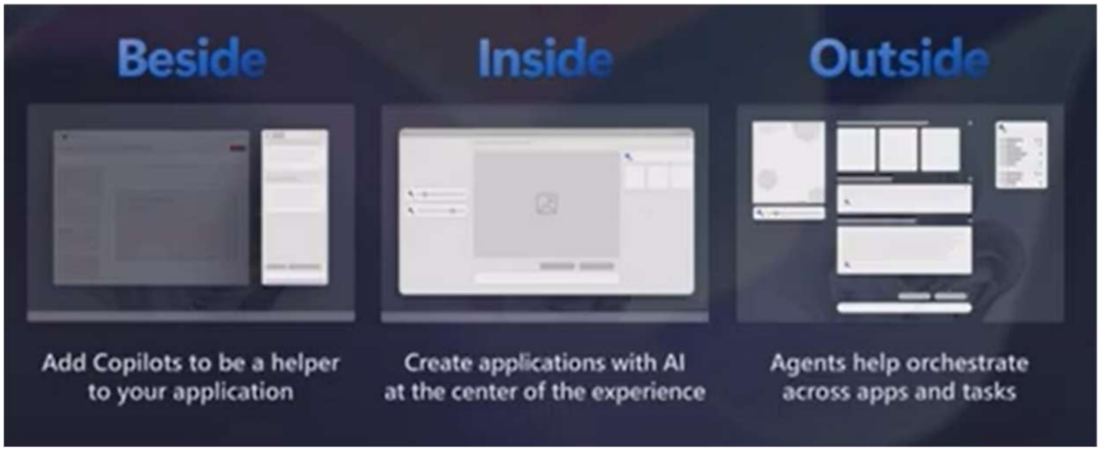
https://x.com/DrJimFan/status/1834279865933332752/photo/1



As lawyer, why should I care?

For the last section of our time together, I want to explain what I think all this has to do with the practice of law.

Beside, Inside, Outside



Beside

Add Copilots to be a helper to your application

Inside

Create applications with AI at the center of the experience

Outside

Agents help orchestrate across apps and tasks

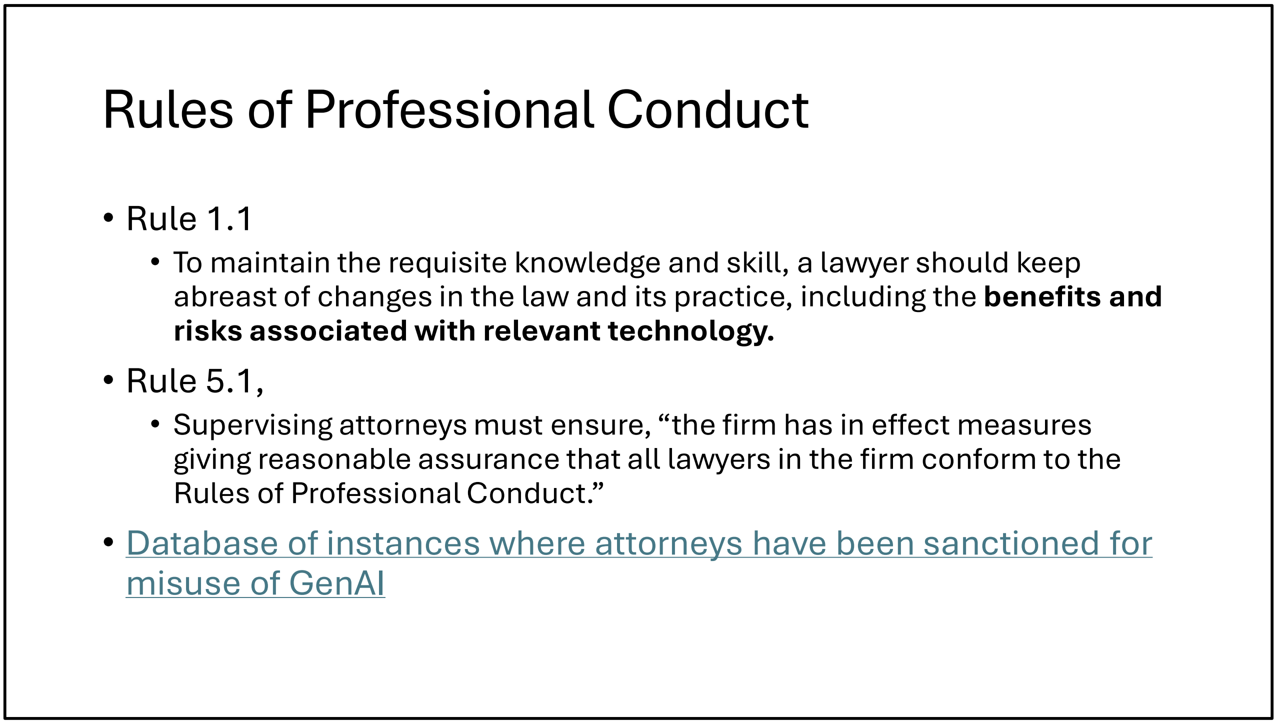
First I need to introduce a concept called beside, inside, and outside. This defines the way lawyers intact with generative AI. A chatbot like Claude sits beside the user. The user has to enter a prompt, give it some data, get an output, then copy that output into their Word document. This is a return to the command line interface. You don't have to know specific commands and syntax, but the volume of context you have to dump into the prompt increases the friction of the interaction.

Increasingly we are seeing AI integrated inside of apps, such as Copilot in Microsoft Word. Menus and user interfaces will be simplified and apps will perform complicated tasks in with simple commands. Imagine telling Word to build a table of authorities for your brief automatically.

I don't think you can buy a piece of legal tech software that doesn't claim to have some AI magic.

Eventually, the AI interface will sit outside of your apps and data. The interface will be an agent that is aware of the context in which you are working. Based on data feeds like email and shared repositories like SharePoint, systems will proactively ask you if you want it to accomplish a task or take an action. The agent will then orchestrate your apps and data to accomplish the task.

https://www.youtube.com/watch?time_continue=205&v=h41Uc73xph4



Rules of Professional Conduct

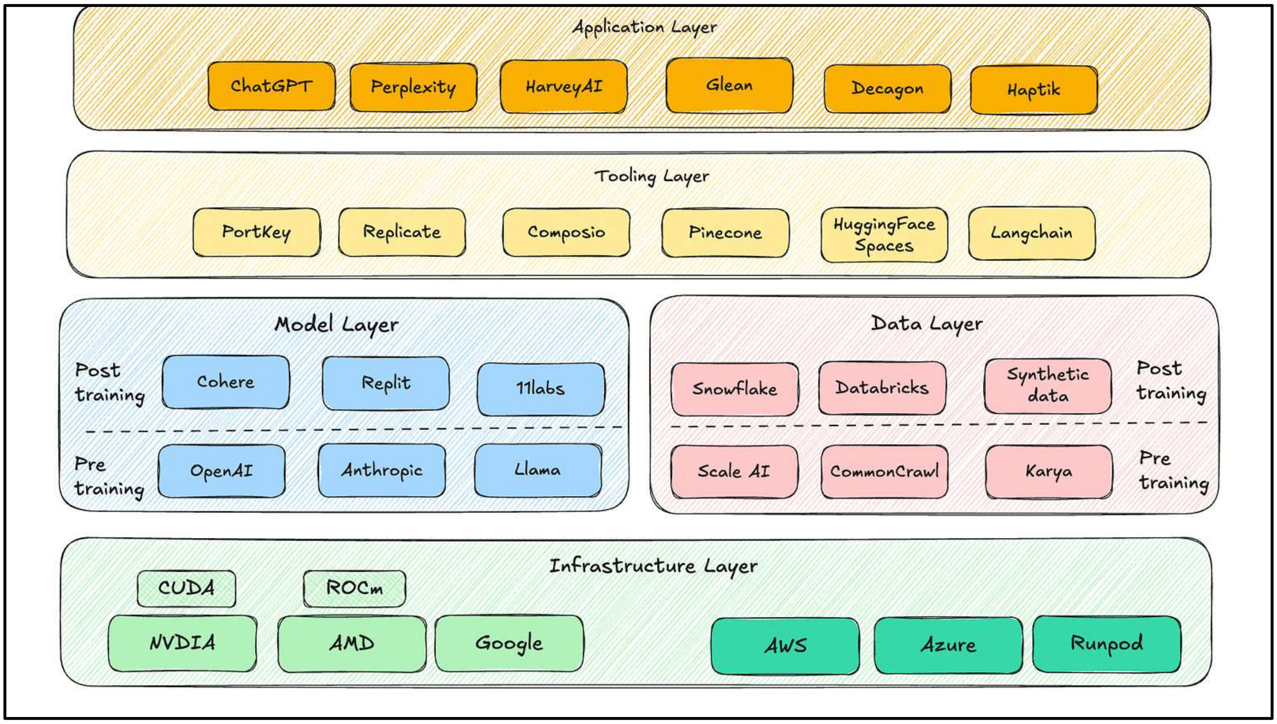
- Rule 1.1
  - To maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, including the **benefits and risks associated with relevant technology.**
- Rule 5.1,
  - Supervising attorneys must ensure, “the firm has in effect measures giving reasonable assurance that all lawyers in the firm conform to the Rules of Professional Conduct.”
- Database of instances where attorneys have been sanctioned for misuse of GenAI

Modern practice of law requires using computers. Rule 1.1 of the Rules of Professional Conduct require that lawyers remain abreast of the risks and benefits of technology. Supervising attorneys also have a duty to ensure the firm has things like AI policies to govern ethical use. The flurry of attorneys being sanctioned for the inappropriate use of GenAI systems shows the downside risks. If the systems that we use to practice law are infused with generative AI, then I think some knowledge of it is not only advisable, but mandatory for all attorneys.

https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_1_competence/comment_on_rule_1_1/

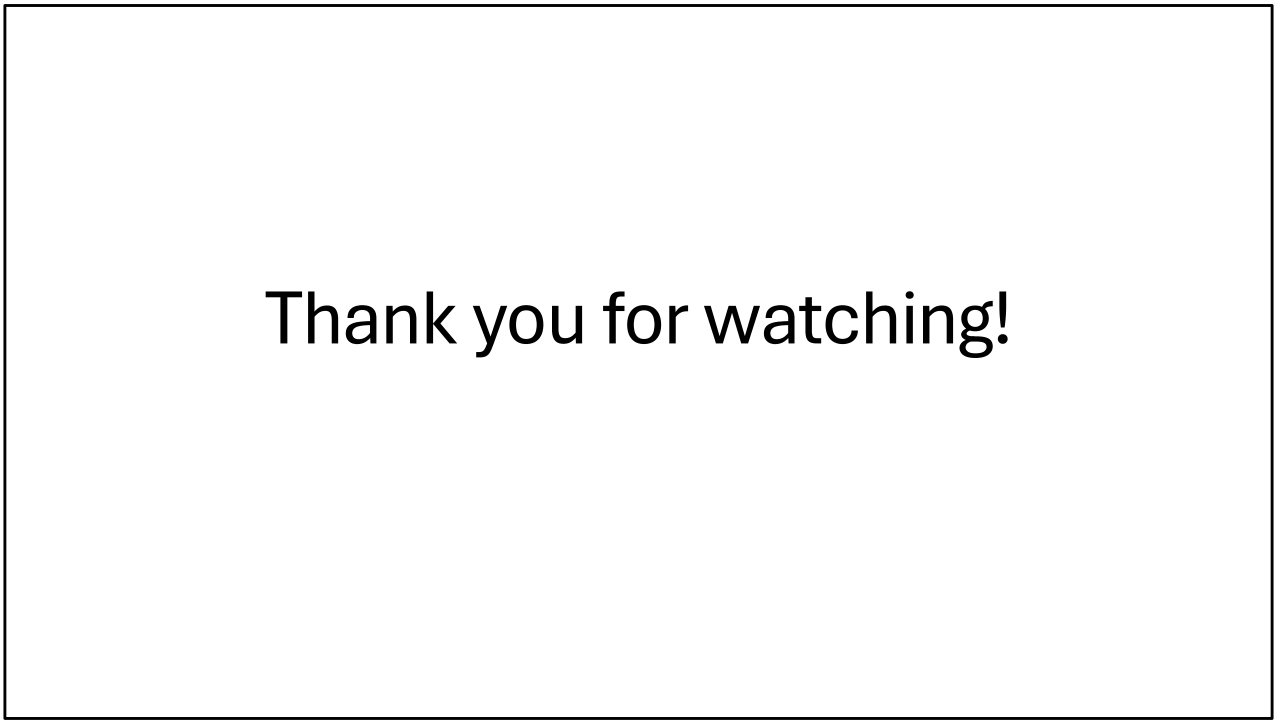
https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_5_1_responsibilities_of_a_partner_or_supervisory_lawyer/

https://www.damiencharlotin.com/hallucinations/



The last little bit of detail that I want to end on is that the application layer that you're working at with Harvey and Perplexity is the just the tip of the AI stack. There are opportunities for innovation at all levels of the ecosystem.

https://www.srajdev.com/p/the-ai-stack



Thank you for watching!

That is the end of my presentation, thank you for your attention.