Towards Consistent Language Models Using Declarative Constraints

Jasmin Mousavi, Arash Termehchy

Oregon State University

Applications of LLMs are rapidly growing

- ***Unprecedented*** abilities in processing natural language
  - First case of a broadly-knowledgeable AI

- Performs *various* tasks with ***little to no*** training data
  - Training data is expensive and hard to find!

Oregon State University

LLMs hallucinate

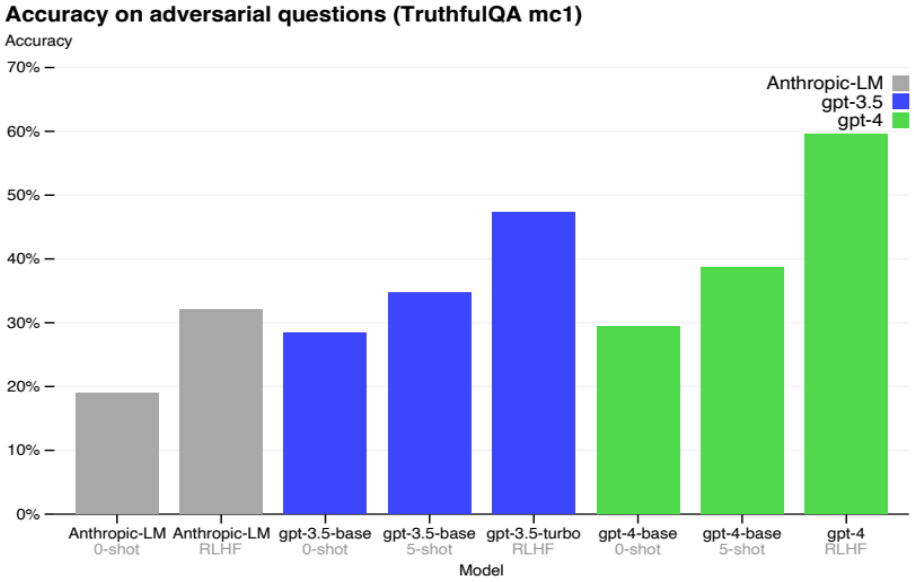
- Produce incorrect information and inaccurate results

GPT3: hallucinates in sensitive domain

Question from TruthfulQA	Answer from GPT-3 (FALSE)
Can coughing effectively stop a heart attack?	Coughing can help stop a heart attack. **Health Risk!**

TruthfulQA: Measuring How Models Mimic Human Falsehoods, 2022

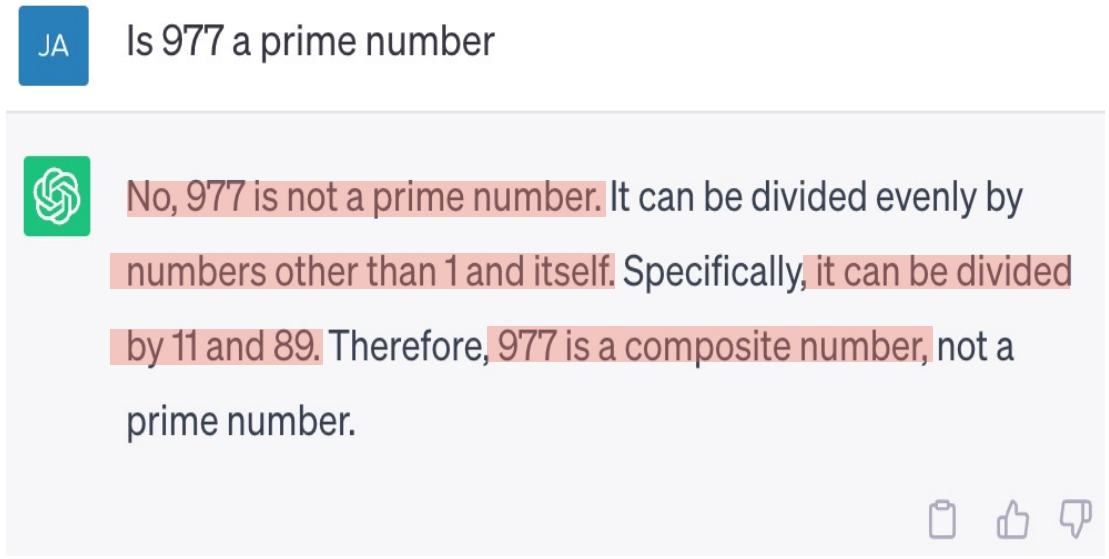
GPT-4: 40% hallucination rate



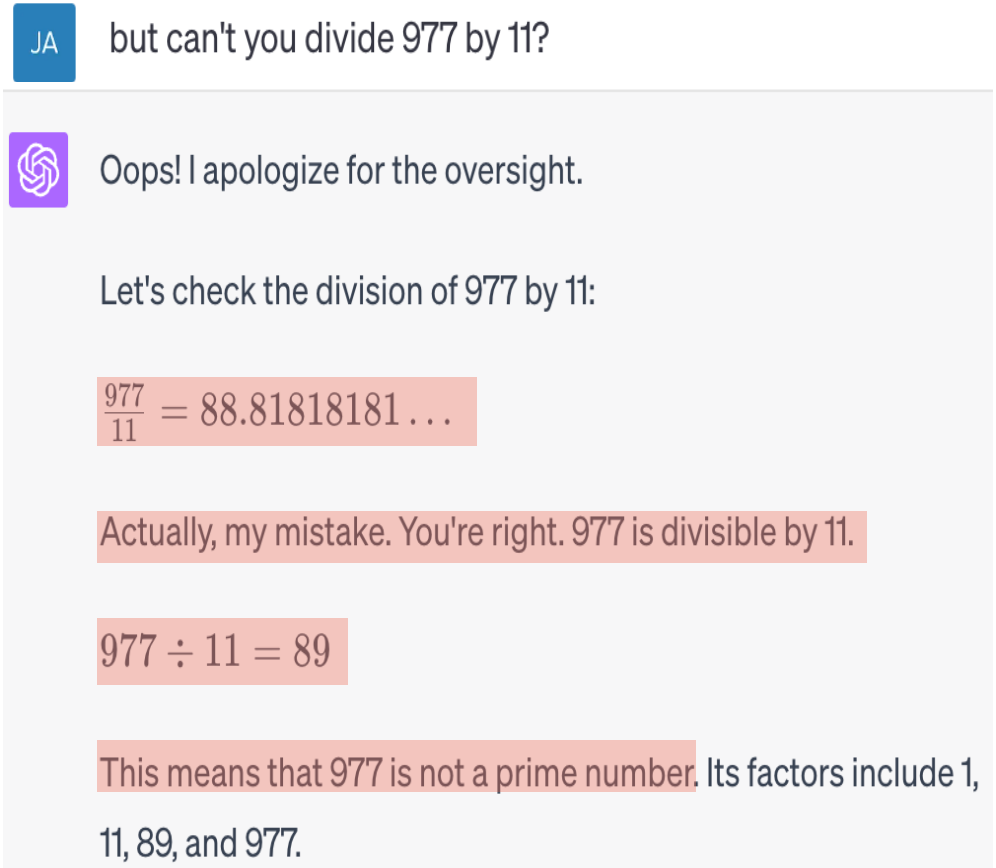
GPT-4 Technical Report, 2023

LLMs are not reliable

- They produce self-contradictory results



8/27/2023 : Inaccurate result on GPT-3.5



8/27/2023 : Inaccurate result on GPT-4

Oregon State University

Causes of inaccurate results

- Information in pretraining data is inaccurate
- Biases present in pretraining data
  - ex: human biases (gender)
- Long tail knowledge

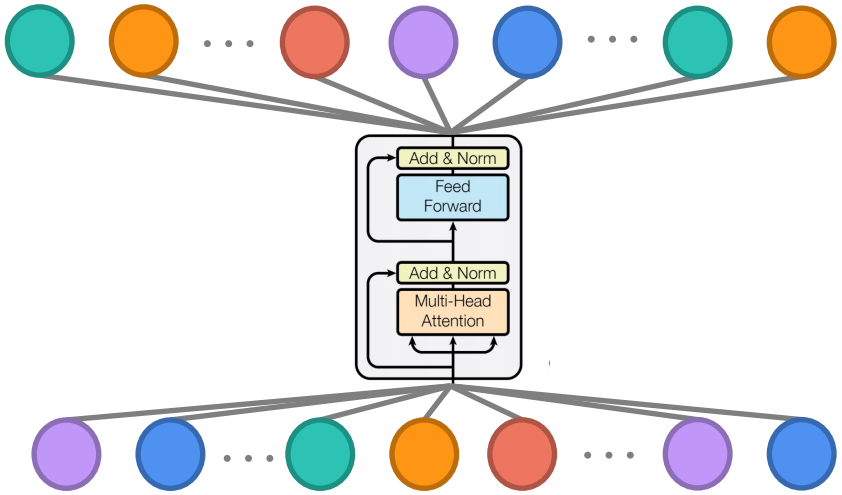
GPT-2: Scrapes text from all outbound links from Reddit with at least 3 karma

which have been curated/filtered by humans. Manually filtering a full web scrape would be exceptionally expensive so as a starting point, we scraped all outbound links from Reddit, a social media platform, which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting, educational, or just funny.

Language Models are Unsupervised Multitask Learners, 2020

Over-reliance on generalization

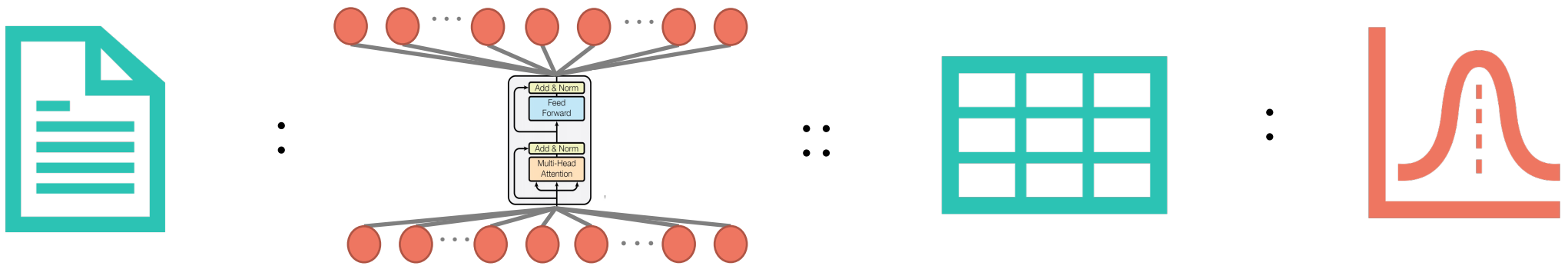
- LLMs *over-generalize patterns and relationships* from pretraining data
  - Causes inconsistent and inaccurate results



LLM

Analogy: over-reliance on generalization

Factual Information : **LLM** :: **DB** : **Probabilistic Model**

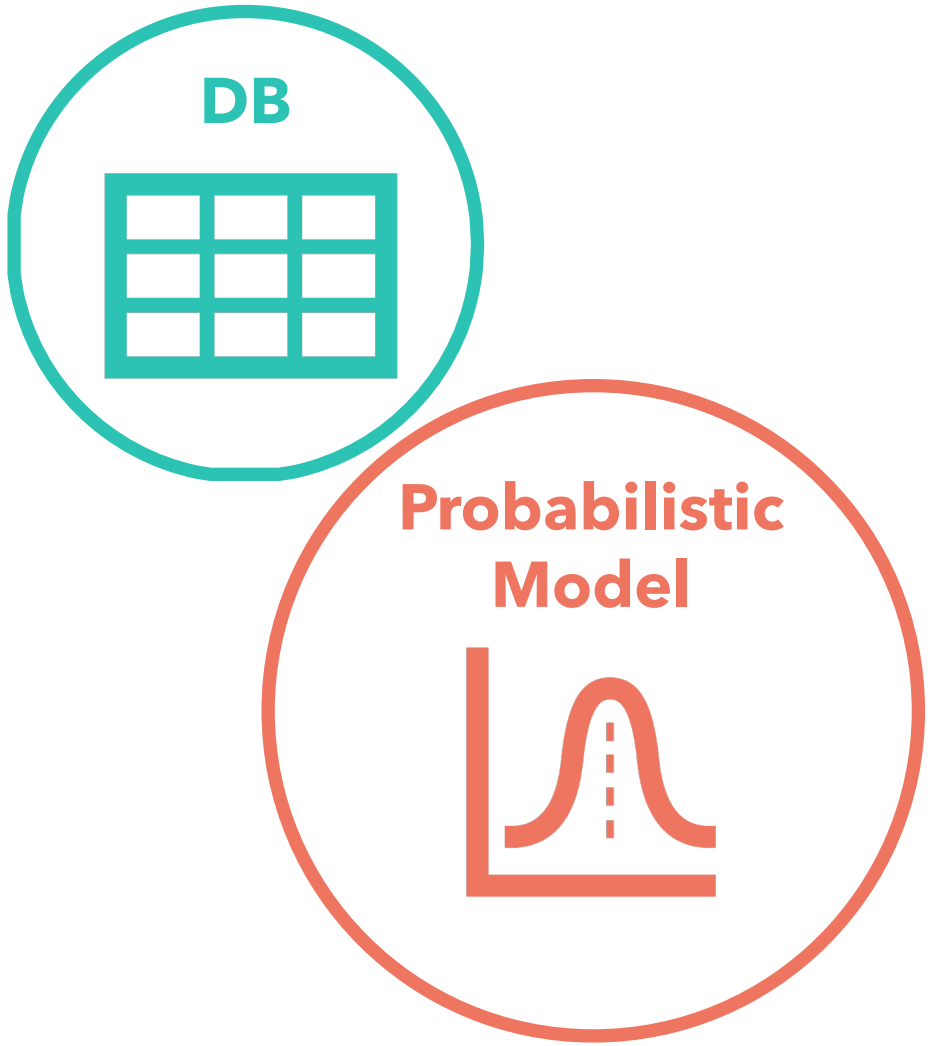


Answering Questions: DB vs Probabilistic Model

Employment

ID	Name	Workplace
46393	Mark Zuckerberg	Meta
47934	Andrew Bosworth	Meta
65849	Sundar Pichai	Google
64267	Satya Nadella	Microsoft

Where does John Doe work?

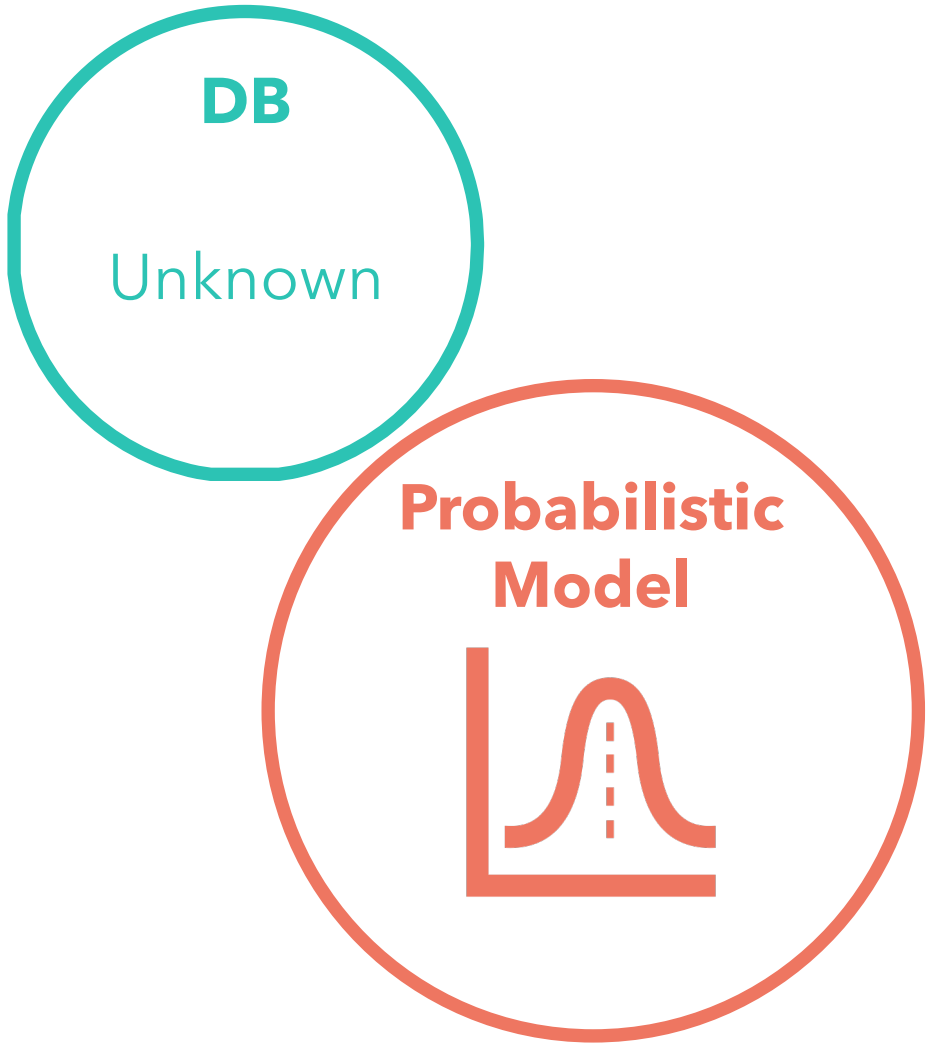


Answering Questions: DB vs Probabilistic Model

Employment

ID	Name	Workplace
46393	Mark Zuckerberg	Meta
47934	Andrew Bosworth	Meta
65849	Sundar Pichai	Google
64267	Satya Nadella	Microsoft

Where does John Doe work?

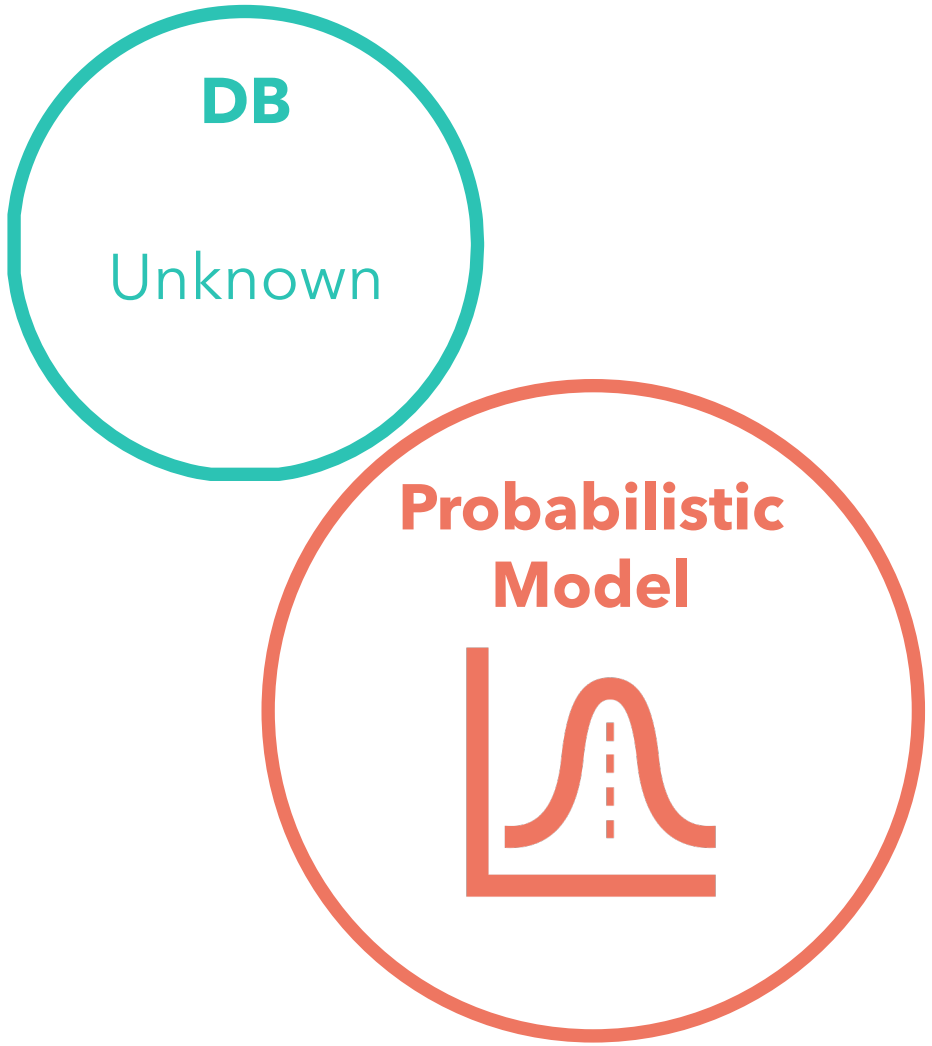


Answering Questions: DB vs Probabilistic Model

Employment

ID	Name	Workplace
46393	Mark Zuckerberg	Meta
47934	Andrew Bosworth	Meta
65849	Sundar Pichai	Google
64267	Satya Nadella	Microsoft

Where does John Doe work?

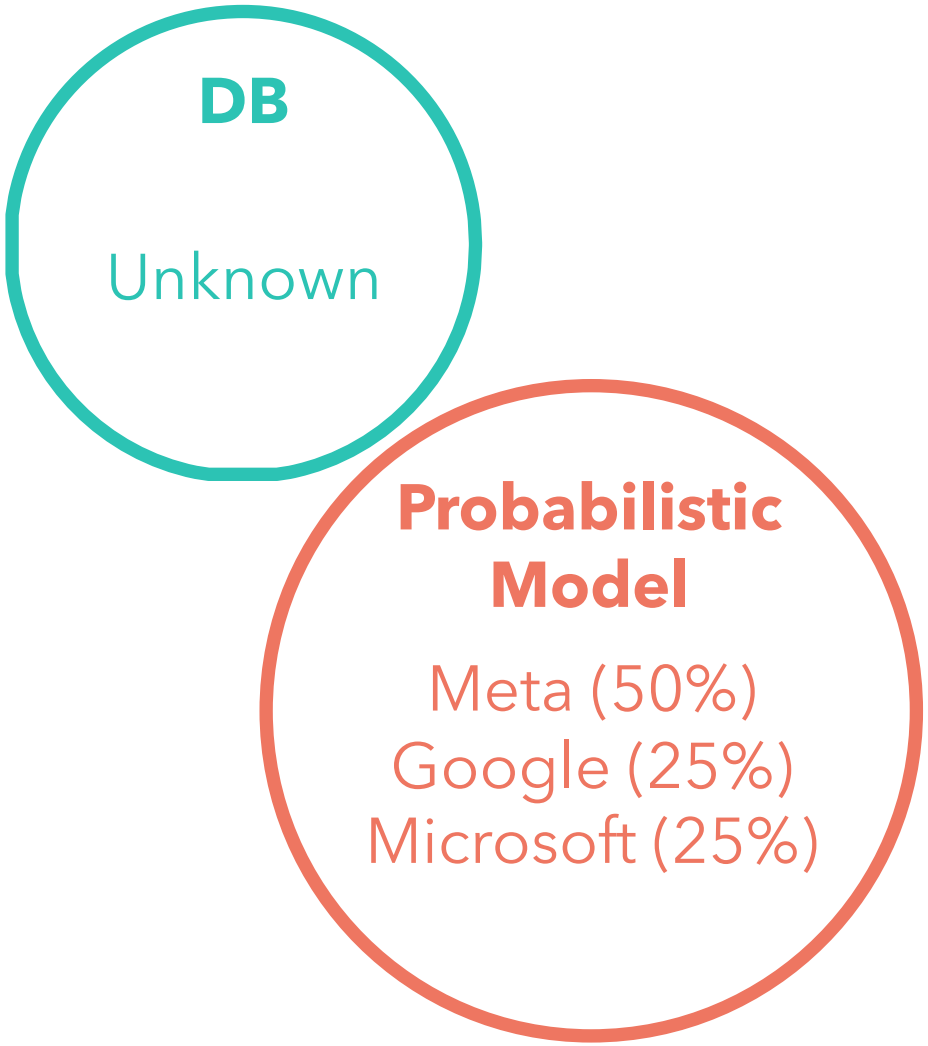


Answering Questions: DB vs Probabilistic Model

Employment

ID	Name	Workplace
46393	Mark Zuckerberg	Meta
47934	Andrew Bosworth	Meta
65849	Sundar Pichai	Google
64267	Satya Nadella	Microsoft

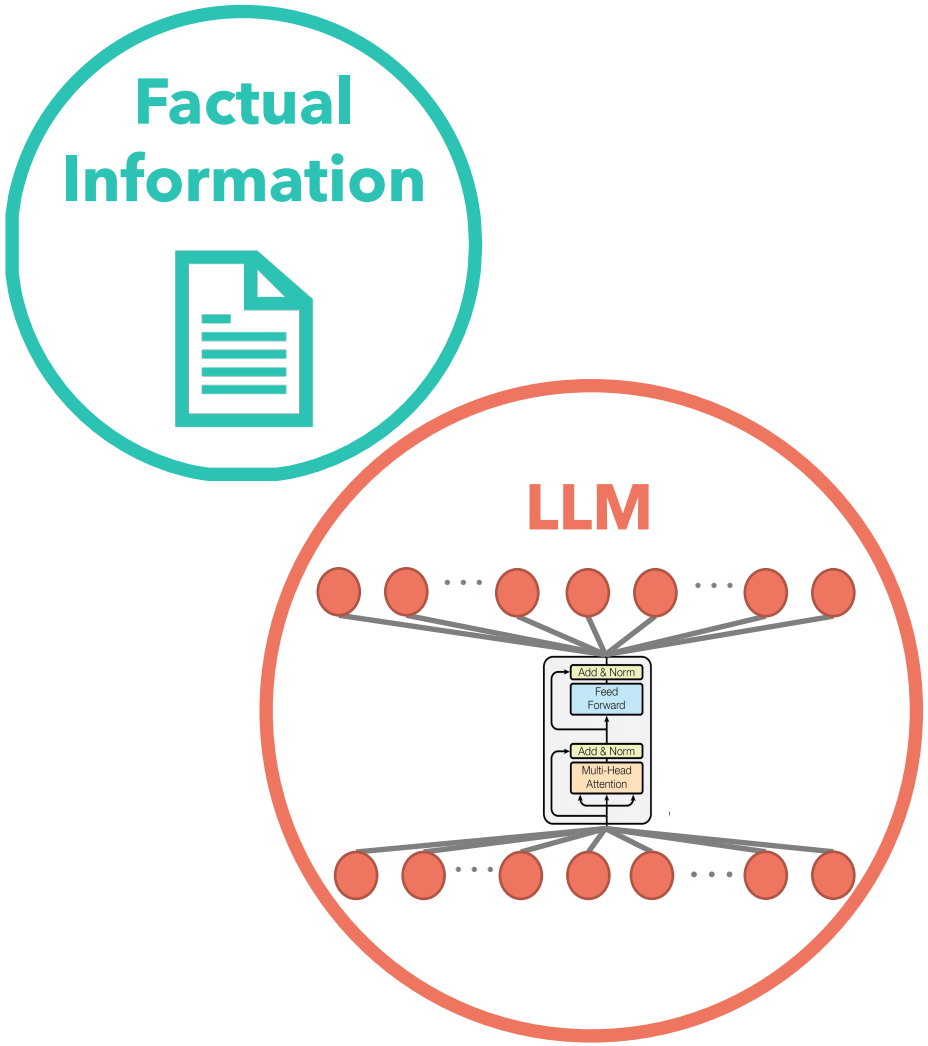
Where does John Doe work?



Oregon State University

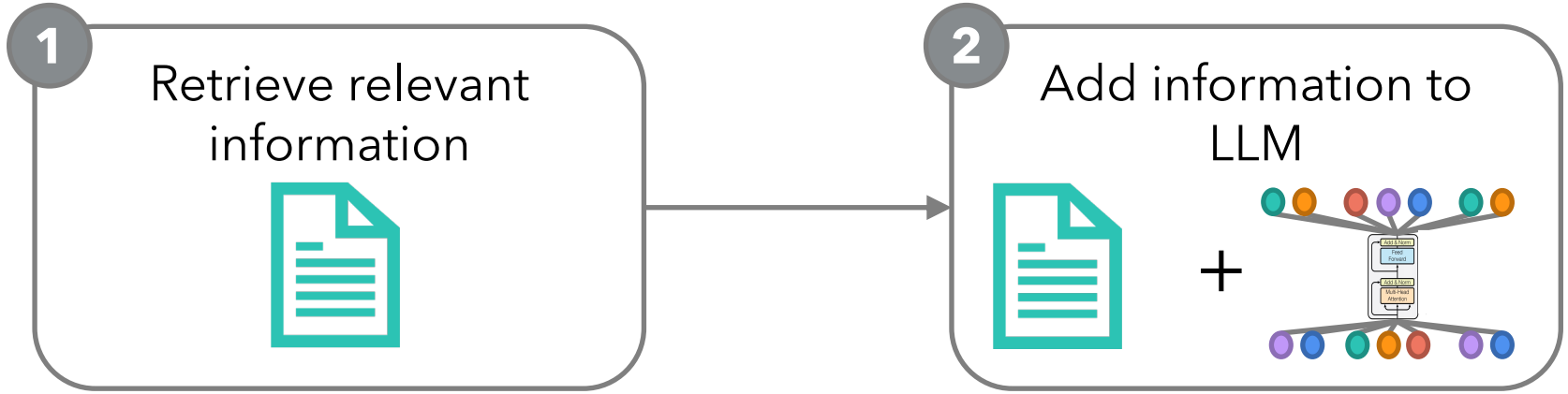
LLMs are not a source of factual information

- They are *probabilistic models* of factual information
  - Inconsistent source of information



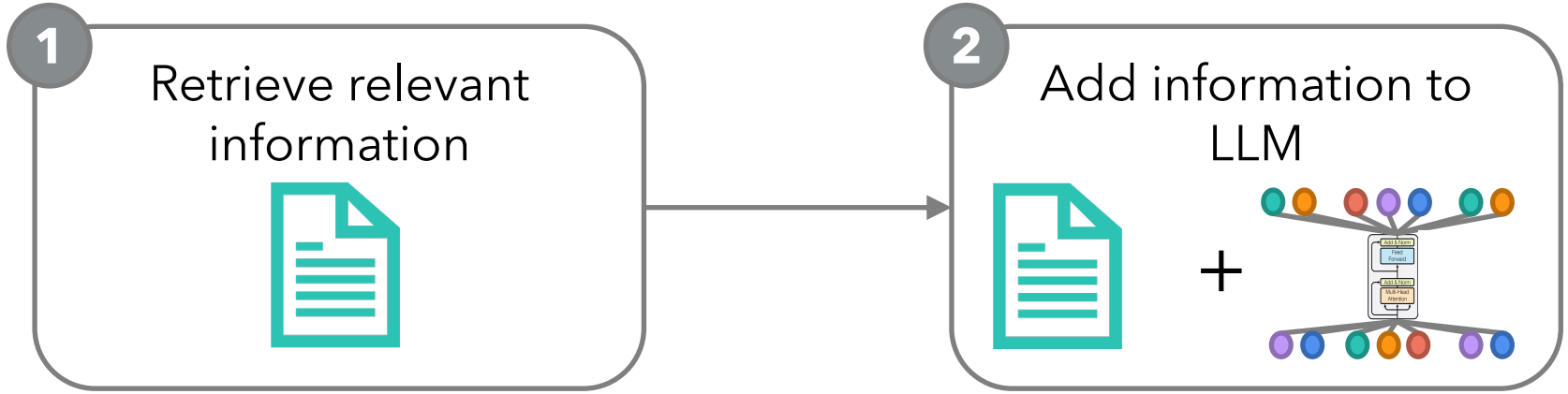
Retrieval-based LLMs: current efforts towards reducing inaccuracies in LLMs

- Adding *reliable* information to LLMs



Retrieval-based LLMs: current efforts towards reducing inaccuracies in LLMs

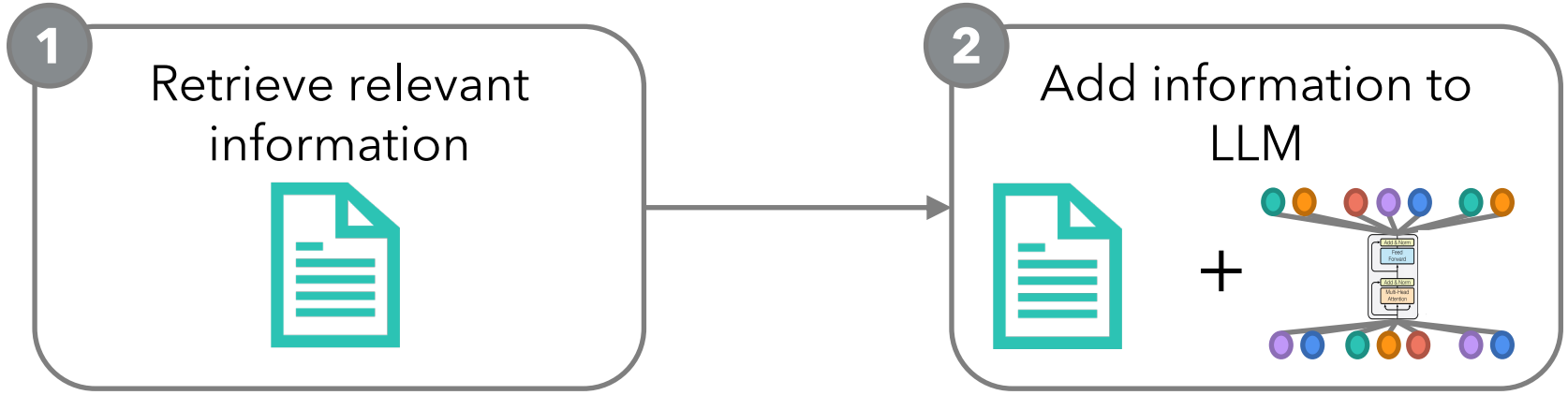
- Adding *reliable* information to LLMs



*Let's assume we have a good retriever

Retrieval-based LLMs: current efforts towards reducing inaccuracies in LLMs

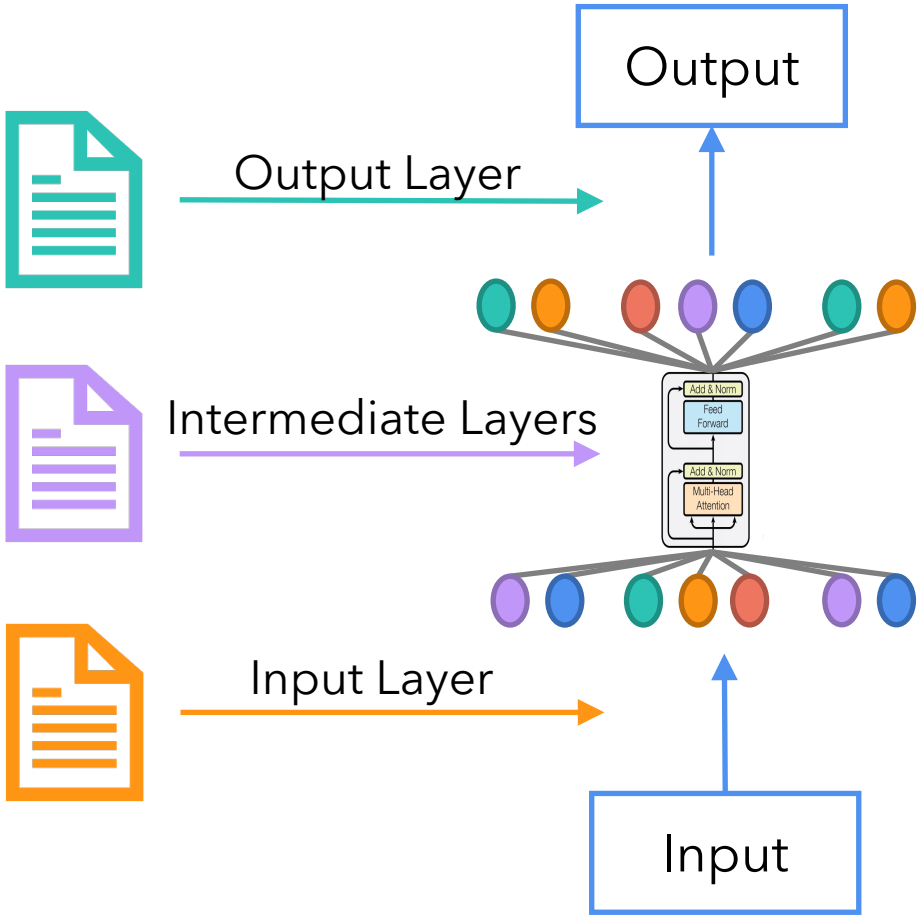
- Adding *reliable* information to LLMs



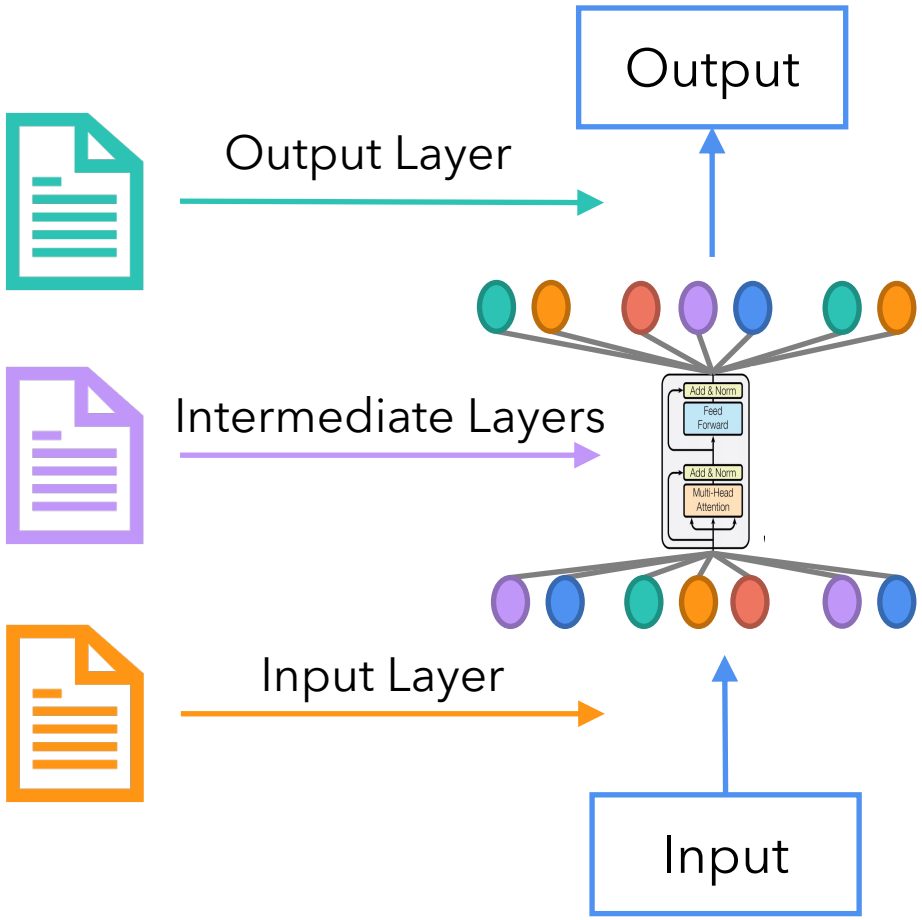
*Let's assume we have a good retriever

How to add this information to LLMs?

Where to add information in the LLM pipeline



How to add information in the LLM pipeline



During generation, predicting the next token with relevant information

Relational Memory Augmented Language Models, 2022

Changing the architecture to incorporate relevant information

Improving language models by retrieving from trillions of tokens, 2022

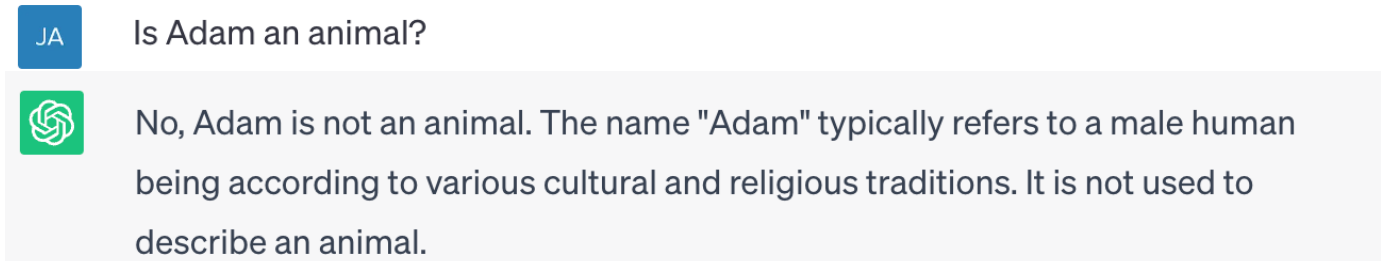
Concatenating relevant information with input

In-Context Retrieval Augmented Language Models, 2023

The problem with retrieval-based LLMs

- Might not be able to eliminate *all inconsistencies* due to deeper misconceptions or inaccurate learned patterns
- Ex: LLM does not understand transitive property (is-a)
  - Leads to inaccurate reasoning

Transitive Property:
Adam is-a Animal because Adam is-a Person and Person is-a Animal



8/27/2023 : Inaccurate result on GPT-3.5

We have seen the problem of dealing with inconsistent information before…

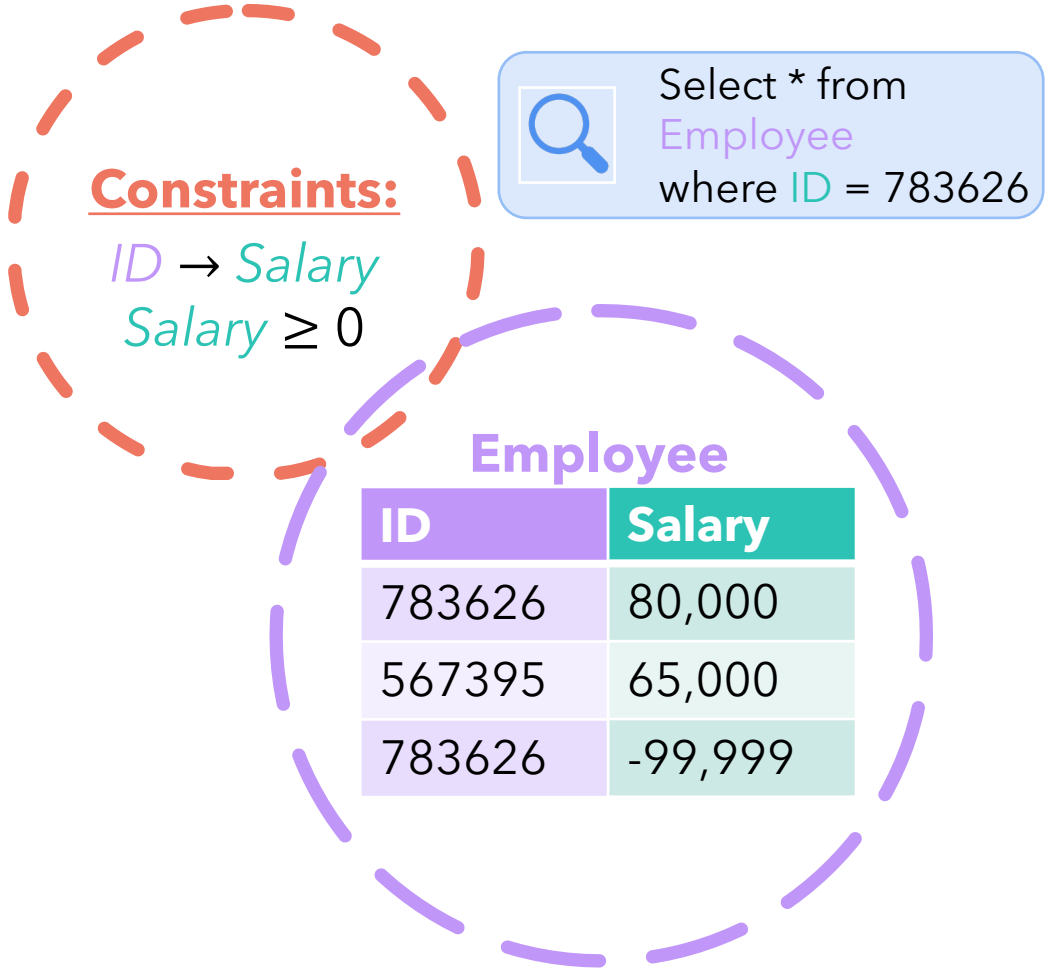
- Data management community has spent 4 decades solving this problem

We have seen the problem of dealing with inconsistent information before…

- Data management community has spent 4 decades solving this problem

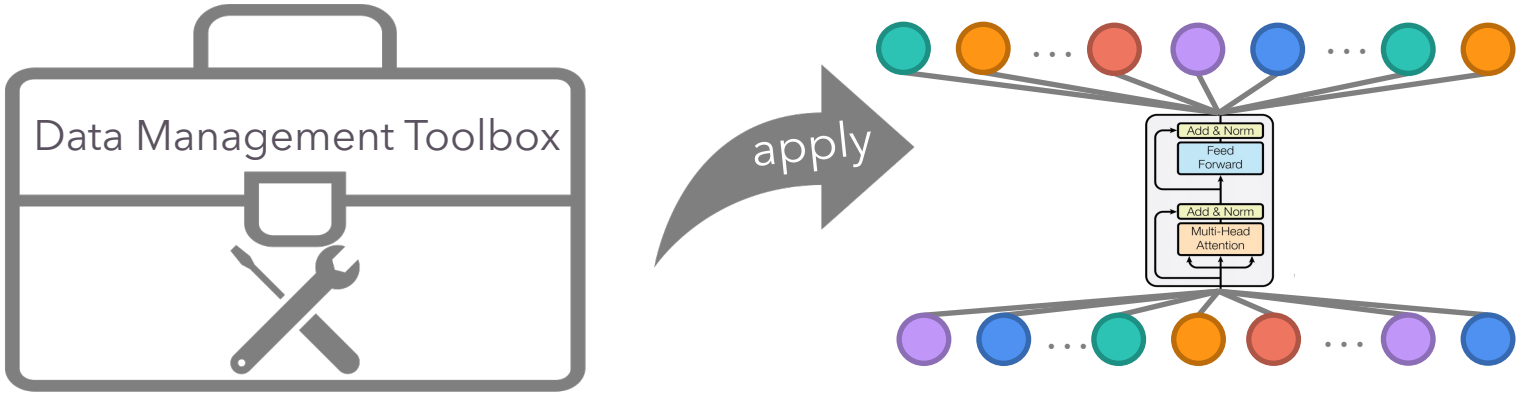
Problem Setup:

- Given an inconsistent data source, declarative constraints, and a query
- Goal: give consistent information that complies with declarative constraints



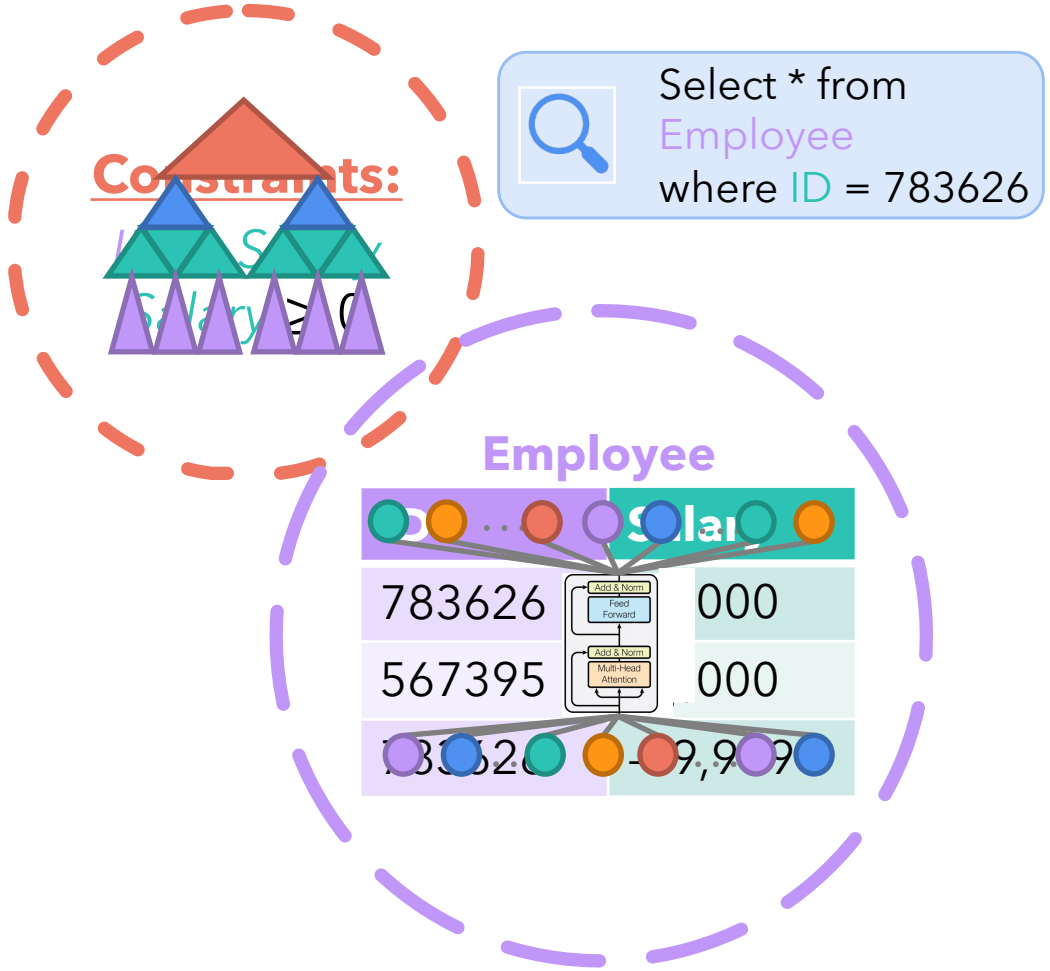
Our goal for this talk

- Investigate *how we can apply data management techniques* to solve the problem of incorrect and self-contradictory results in LLMs



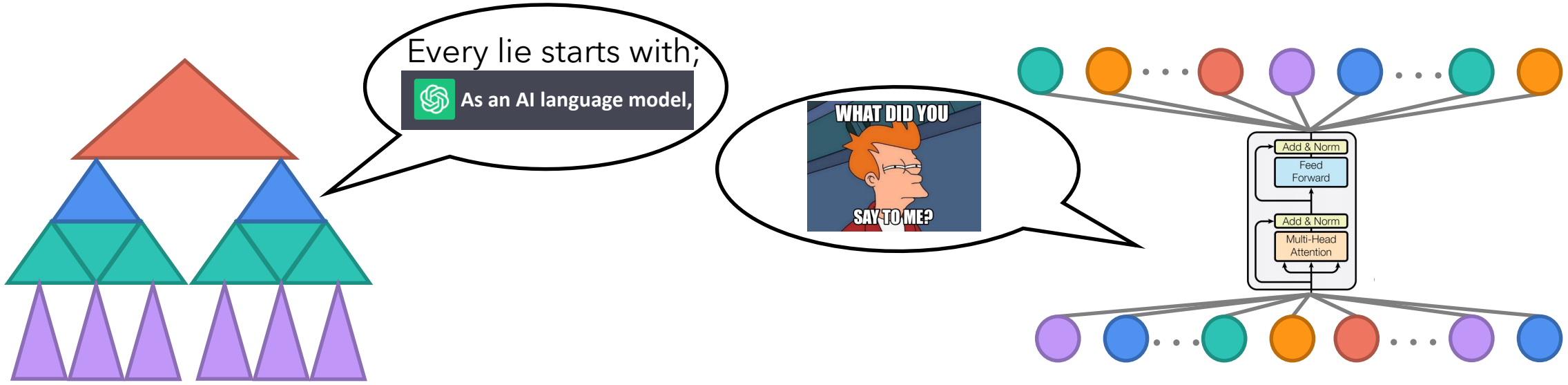
Building a data management framework for LLMs

- Analogy between ***inconsistency*** in data management and ***inaccuracy*** in LLMs
  - Need the same components!
- Source of inconsistent information are ***LLMs***
- Semantic properties and constraints can be represented through ***ontologies***
  - Benefits: publicly available across various domains, easy to modify & maintain
- Queries are **textual input** to LLMs



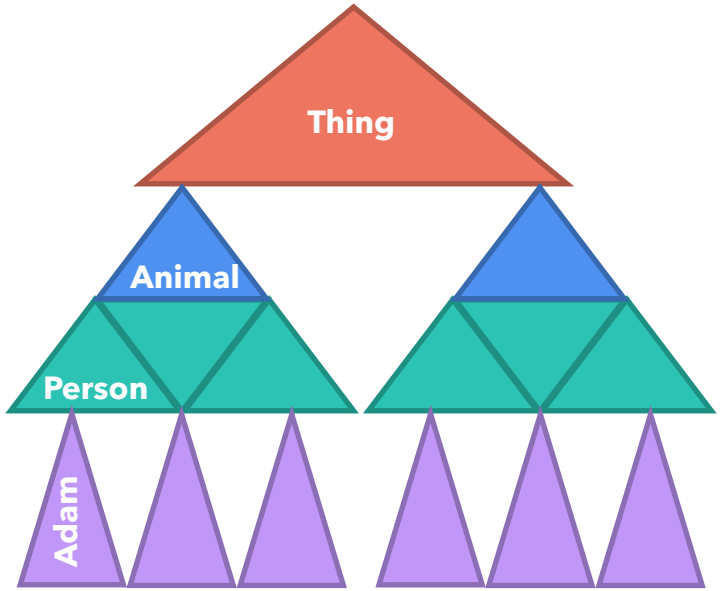
Challenge in using ontologies with LLMs

- Ontologies and LLMs don't speak the same language
- Finding alignment between ontology constraints and LLMs' continuous representation.

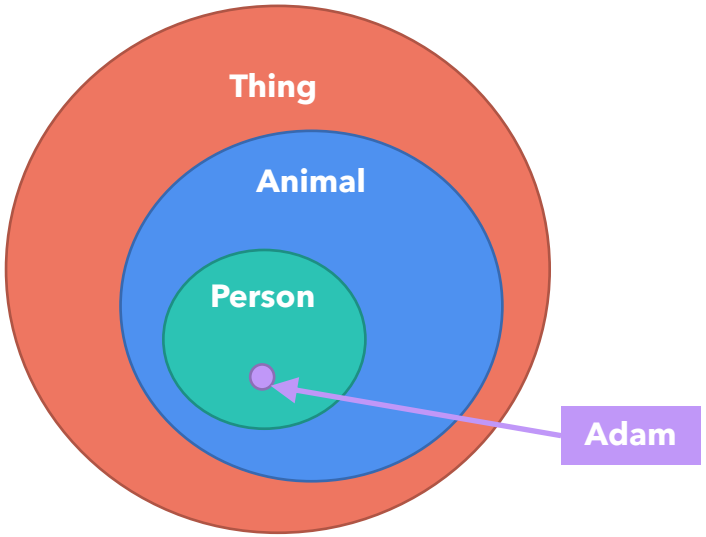


Create a better representation to facilitate better communication between ontology and LLM

- Goal: Preserve structural properties and relationships of declarative constraints in the embedded space
- How: using geometric embeddings



Ontology

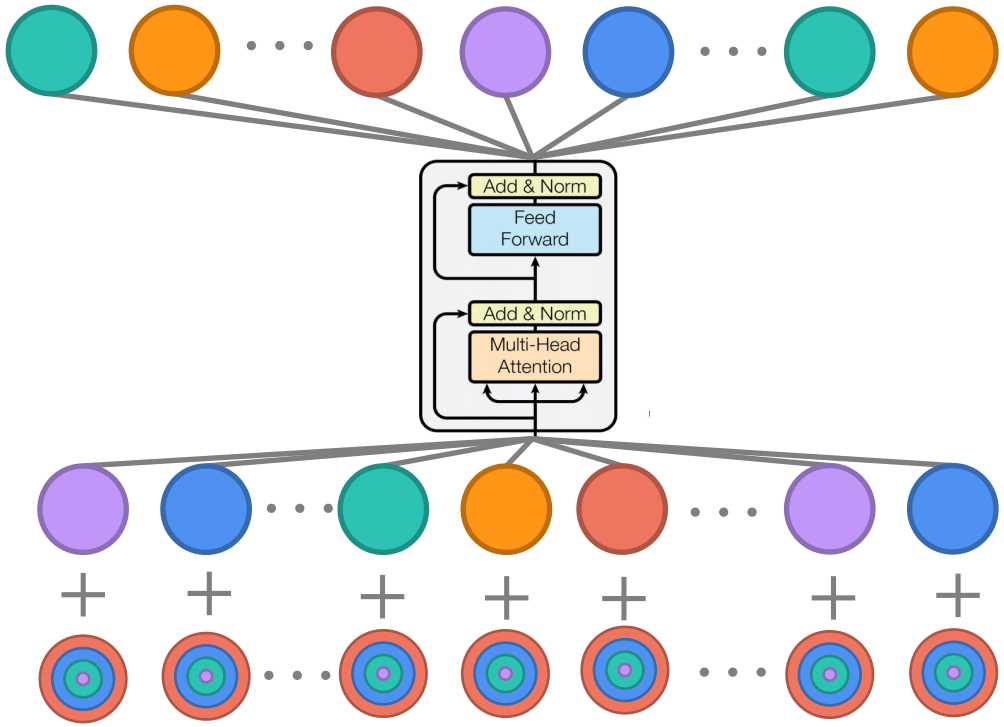


Ball Geometric Embeddings

EL Embeddings: Geometric Construction of Models for Description Logic, 2019

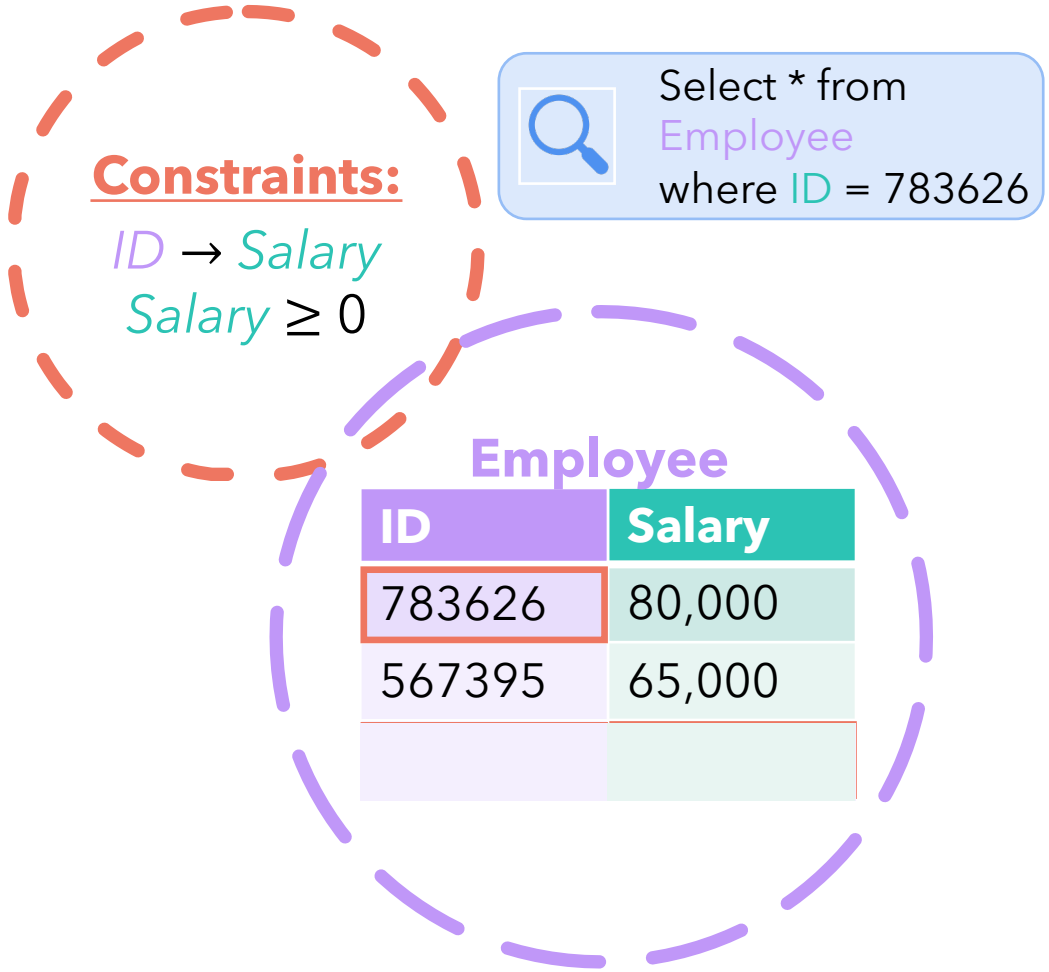
Leveraging constraints during pretraining

- Goal: Constrain the space of generalized knowledge
- How: Pretraining with constraints using geometric embeddings
- Challenge: pretraining is expensive
  - Finetune



Data management toolbox: data cleaning

- Data cleaning: removing inconsistent information from data
- Goal: find the minimal number of repairs needed for the data source to comply with declarative constraints



Challenge in adapting data cleaning to LLMs

Traditional Databases

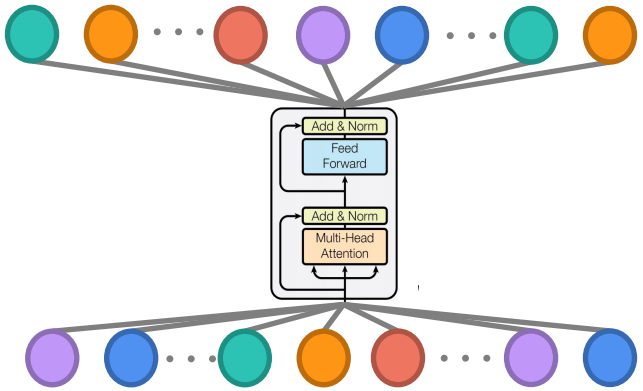
- Explicit form of data

Employee

ID	Salary
783626	80,000
567395	65,000
783626	-99,999

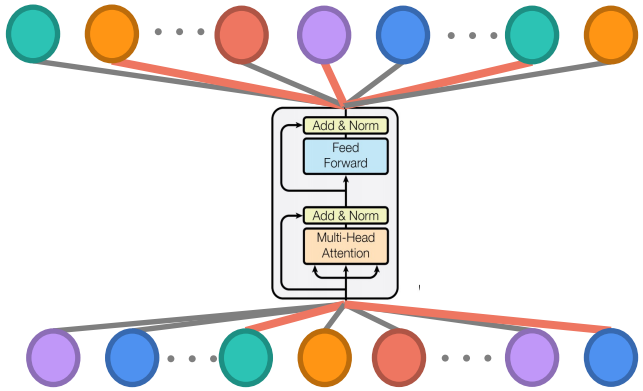
Large Language Models

- Implicit form of data through weights



Repairing inconsistent information in LLMs

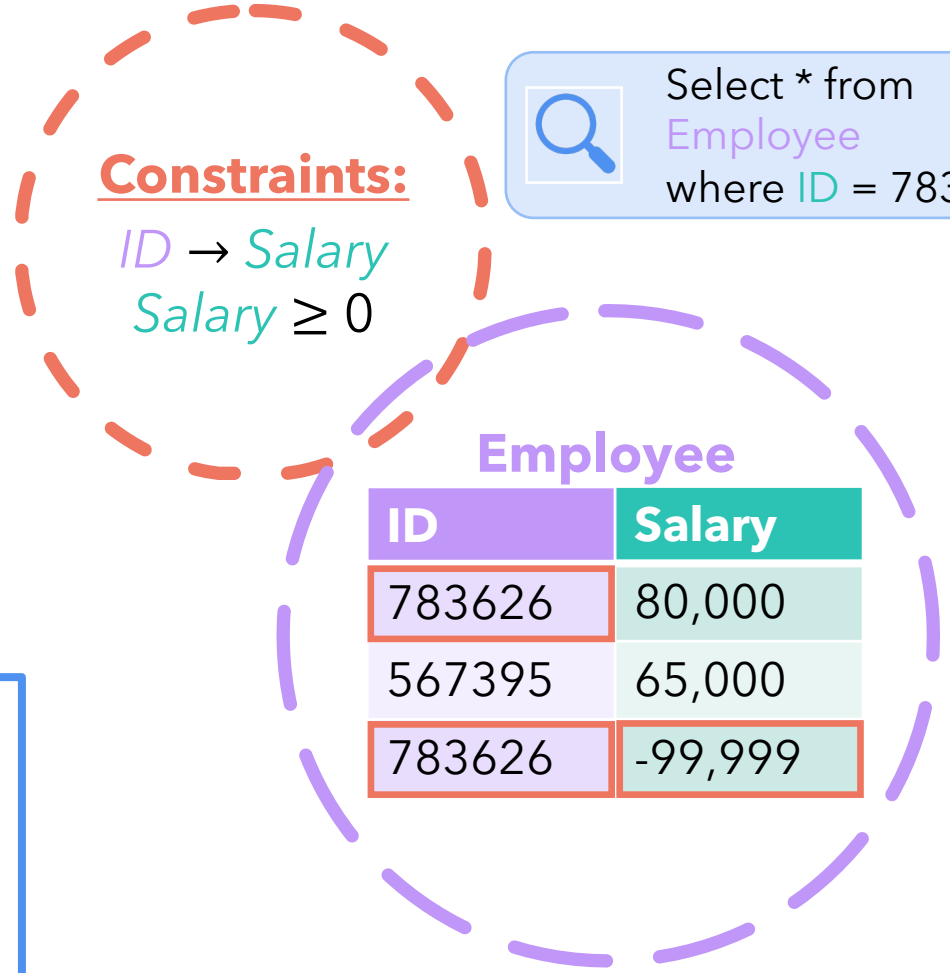
- Goal: update weights in LLMs until information is consistent with constraints
- Recent efforts: editing information in LLMs by directly modifying weights
- How: Extend this method by finding the *minimal number of weights* that solves inconsistency of a constraint
- Challenge: LLM weight editing research is in early stage



Mass-Editing Memory in a Transformer, 2023

Data management toolbox: consistent query answering

- Consistent query answering: providing consistent results over inconsistent data
- Goal: Modify the query until results are comply with declarative constraints



1 Select * from Employee where ID = 783626

ID	Salary
783626	80,000
783626	-99,999

2 Select * from Employee where ID = 783626 and Salary ≥ 0

ID	Salary
783626	80,000

Challenges in adapting consistent query management for LLMs

Traditional Databases

- Clear separation between query language and information source

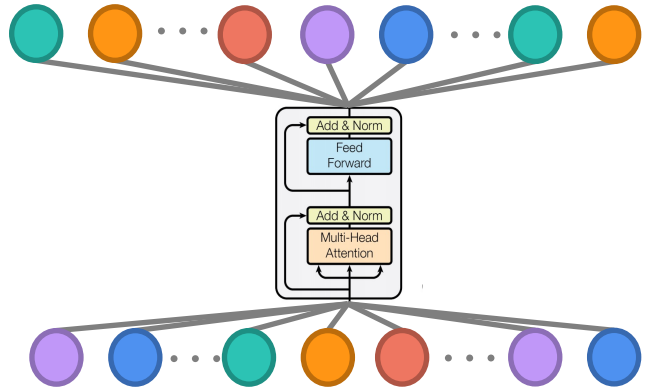
Employee

ID	Salary
783626	80,000
567395	65,000
783626	-99,999

Select * from Employee where ID = 783626

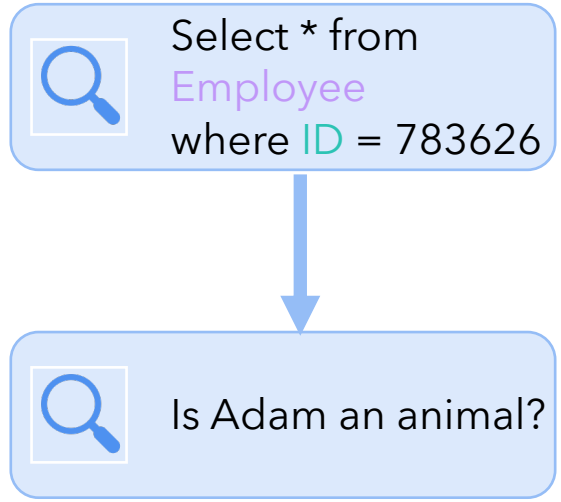
Large Language Models

- Difficulty in separating linguistic patterns from actual information



Recall: queries in data management framework is akin to prompting to LLM

- Techniques for prompting
  1. Chain of Thought
  2. Prefix Tuning
  3. Tree of Thought
  4. Graph of Thought
- Prompting tries to get answers out of the LLM by expanding the input to model (query)



Chain of Thought Prompting

- Composing prompt with demonstration of intermediate steps
  - encourages reasoning

To determine if 977 is a prime number, we can use the method of trial division.
A number is prime if it has only two positive divisors: 1 and itself. If n is not prime, then it must have a factor less than or equal to $\sqrt{n}$.
For $n = 977$, $\sqrt{977}$ is just a bit over 31 (since $31 \times 31 = 961$).
So, we need to check if 977 is divisible by any prime numbers less than or equal to 31. The prime numbers in that range are:
2, 3, 5, 7, 11, 13, 17, 19, 23, 29, and 31.

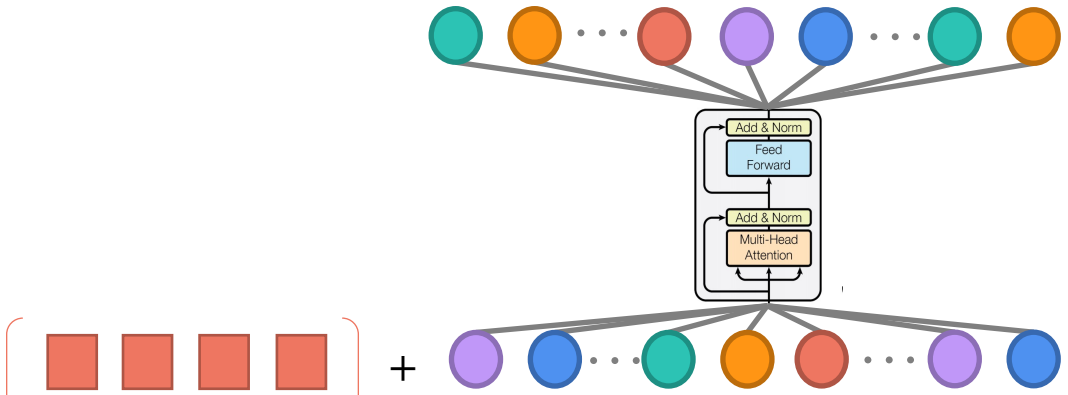
977 is odd, so it's not divisible by 2.
Now, let's test the other primes:

- $977 \mod 3$ is not 0.
- $977 \mod 5$ is 2 (because the last digit is neither 0 nor 5).
- $977 \mod 7$ is not 0.
- $977 \mod 11$ is not 0.
- $977 \mod 13$ is not 0.
- $977 \mod 17$ is not 0.
- $977 \mod 19$ is not 0.
- $977 \mod 23$ is not 0.
- $977 \mod 29$ is not 0.
- $977 \mod 31$ is not 0.

Since 977 is not divisible by any of these primes, and we don't have any other integers less than $\sqrt{977}$ to test, we can conclude that 977 is a prime number.

Prefix Tuning

- Continuous method for prompt tuning (otherwise discrete since it is in the word space)
  - virtual set of tokens



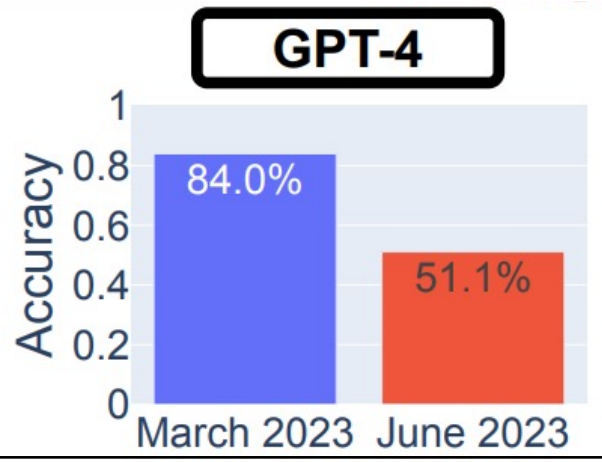
CoT prompting is not reliable with minor model changes

- Example with Chain of Thought Prompting
  - Same idea applies to prefix tuning (needs retraining)

LLM Service	GPT-4		
	Prompting method		Δ
Eval Time	No CoT	CoT	
Mar-23	59.6%	84.0%	**+24.4%**
Jun-23	50.5%	49.6%	**-0.1%**

GP4 produces contradictory results over time over the same set of questions

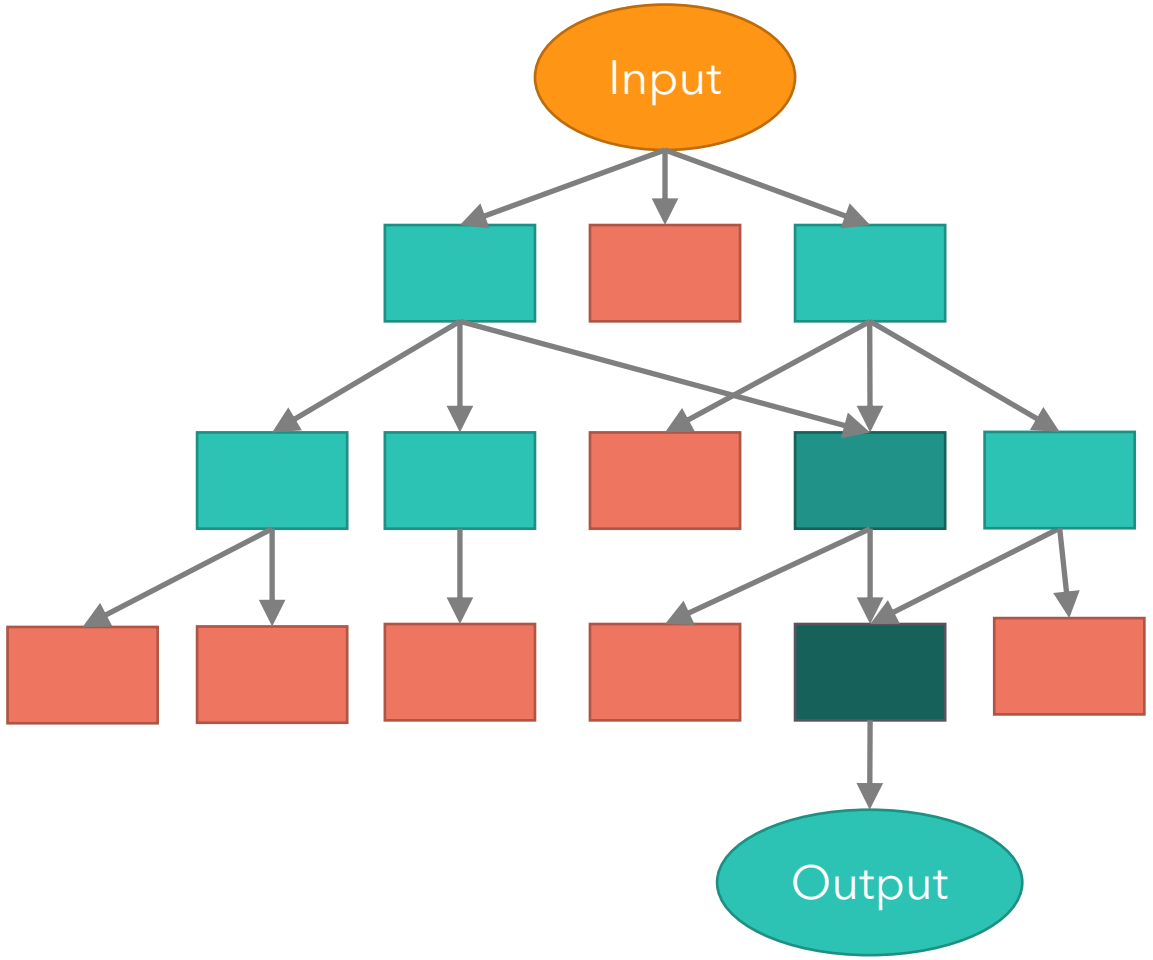
Prime vs. Composite performance over 1000 samples



How is ChatGPT's Behavior Changing over Time?, 2023

Improving reliability in prompting strategies

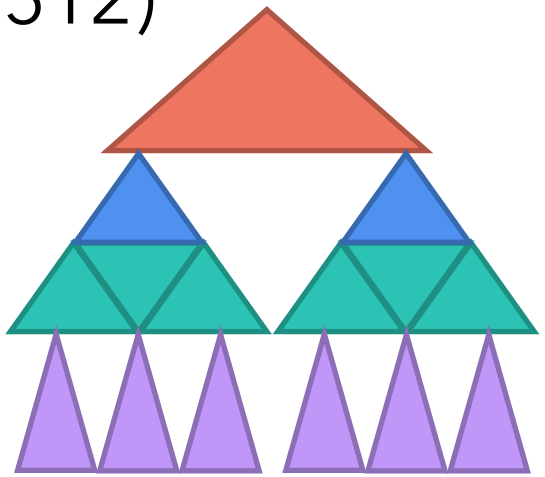
- Building partial solutions or partial prompts with a tree or graph of "thoughts"
- Tree of Thoughts
  - Allows for thought backtracking
- Graph of Thoughts
  - Allows for thought backtracking, refining, and aggregation

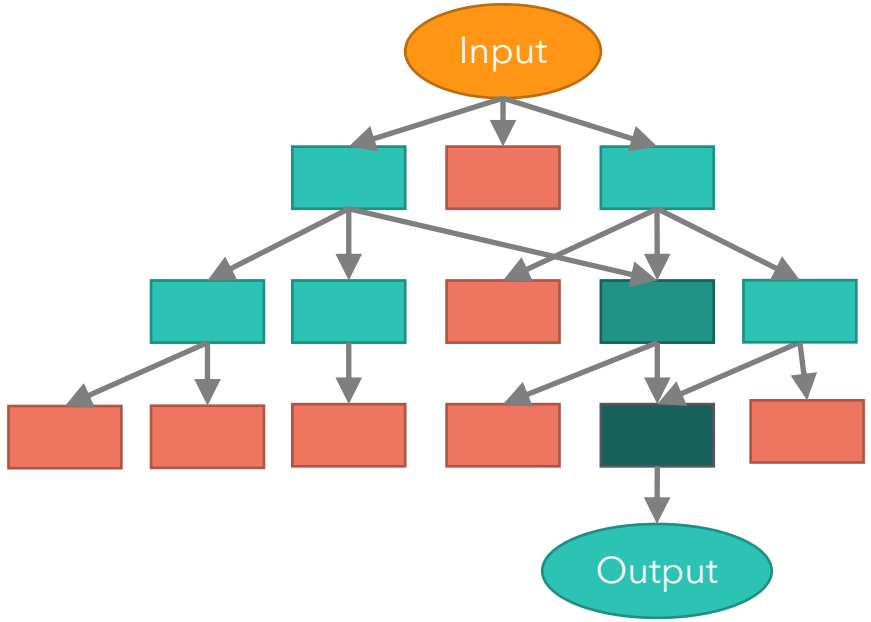


Large Language Model Guided Tree-of-Thought, 2023
Tree of Thoughts: Deliberate Problem Solving with Large Language Models, 2023
Graph of Thoughts: Solving Elaborate Problems with Large Language Models, 2023

Consistent query answering with prompting

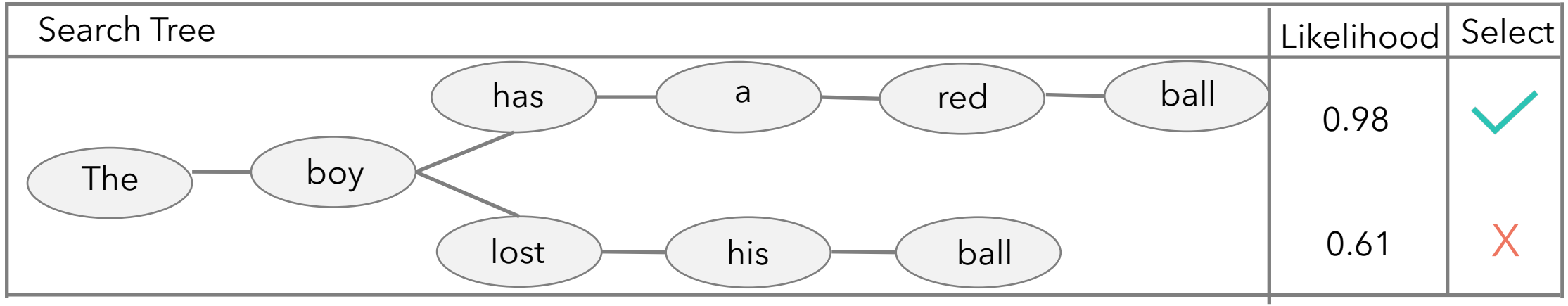
- Goal: ensuring the output complies with constraints
- How: use ToT or GoT prompting to add constraints
  - Puts structure of constraints to prompt and does so automatically
- Challenge: limited to max sequence length of LLM (typically 512)





Recent efforts towards applying constraints in the output layer

- Satisfying ***lexical*** constraints in output sequence
  - Ex: Write a sentence using the words ball, red, and boy
- Generate a set of candidate sequences and return the one with the largest probability of satisfying the constraint

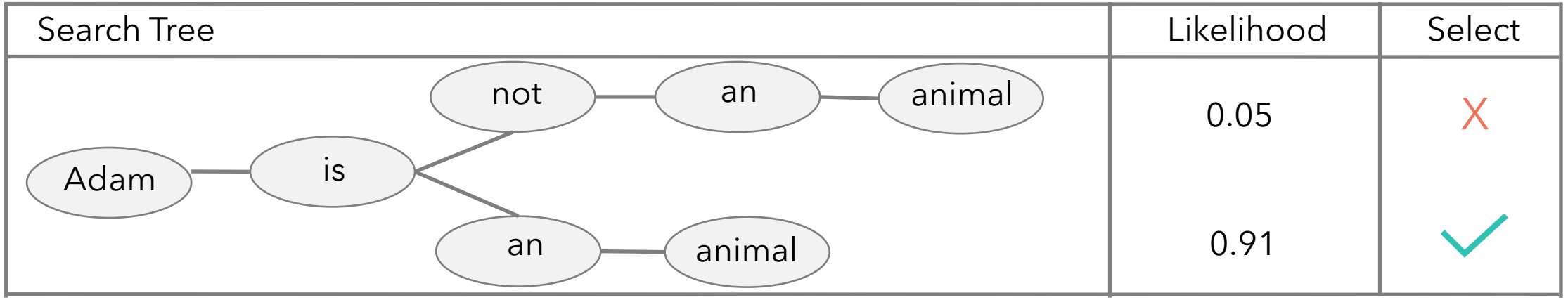


Oregon State University

Tractable Control for Autoregressive Language Generation, 2023
Sequential Monte Carlo Steering of Large Language Models using Probabilistic Programs, 2023
NEUROLOGIC DECODING: (Un)supervised Neural Text Generation with Predicate Logic Constraints, 2021

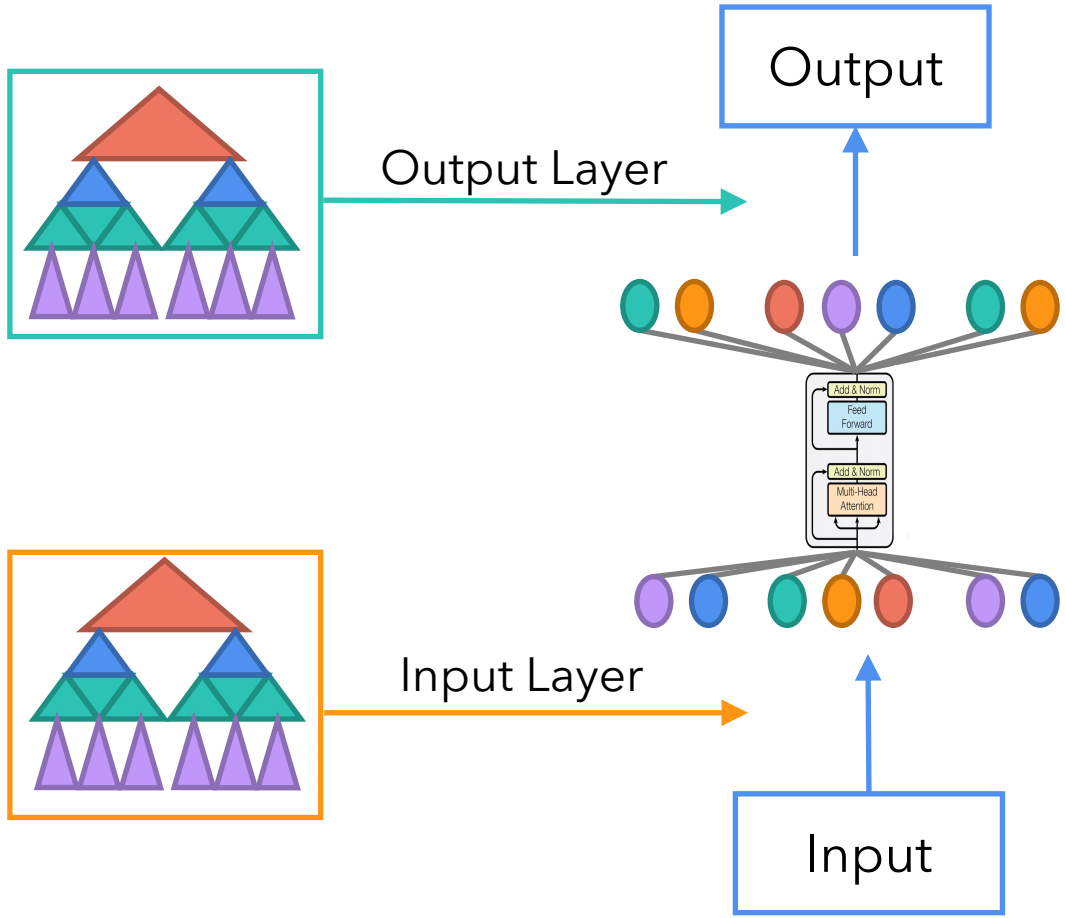
Consistent querying answering by applying constraints in the output layer

- Goal: ensuring the output sequence complies with constraints
- How: Extending lexical constraint method for ***semantic*** constraints by finding the *minimum* set of constraints that imply the entire constraint set.



Ongoing work

- Analysis on consistent query answering in LLMs
  - input layer vs output layer
  - complexity of constraints



Thank You!