



Onepoint
TechTalk

Decoding AI series

Session 2

Unleashing the power of Large Language Models

Part 2 – Workflows and complex interactions

Our awesome speaker



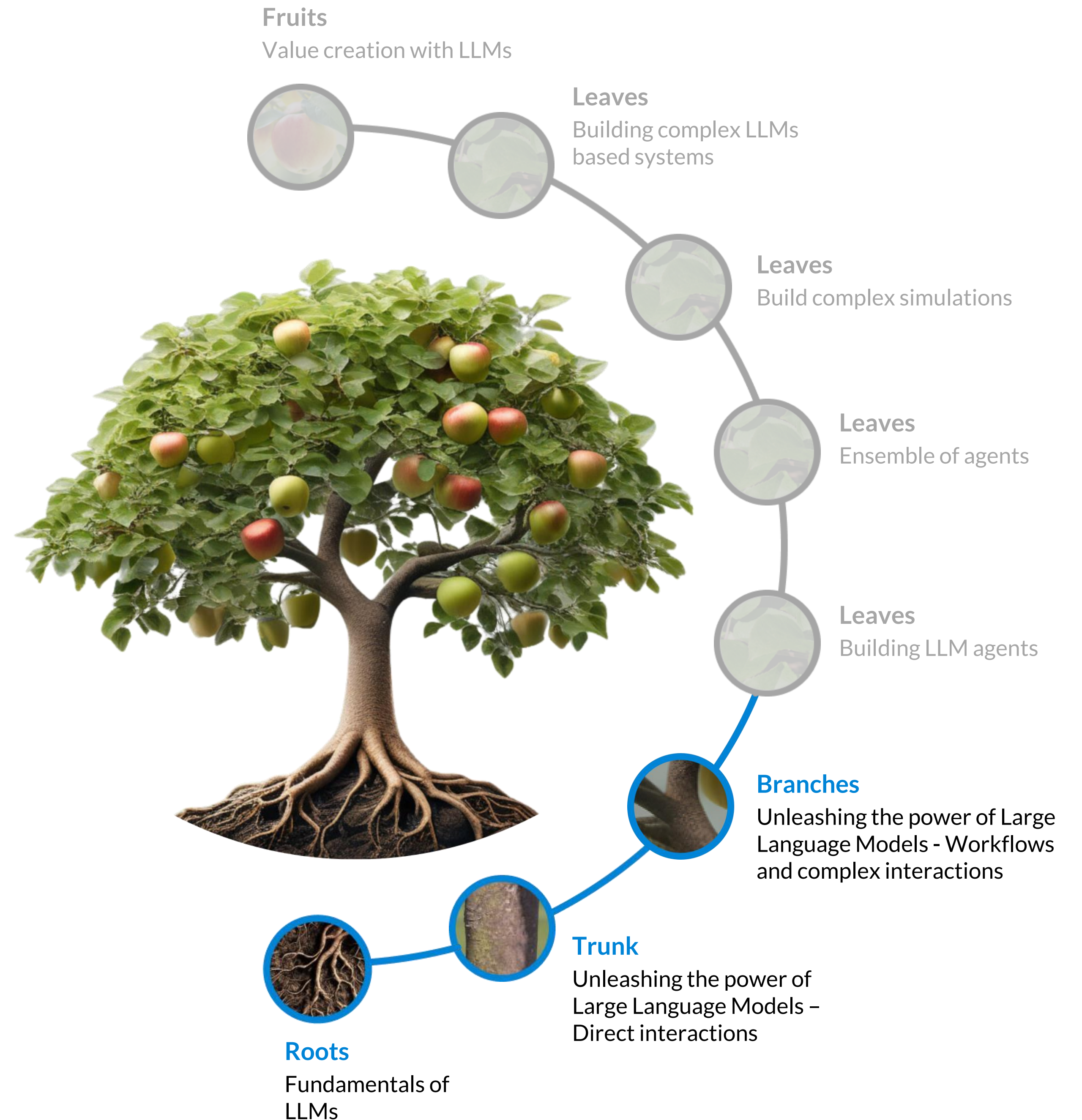
Gil Fernandes

AI Solutions Engineer

Welcome

From Roots to Fruits

- Journey to create value from LLMs



Agenda

Insights from previous webinar	<u>04</u>
In-context Learning	<u>11</u>
RAG (Retrieval Augmented Generation)	<u>14</u>
Reasoning strategies	<u>18</u>
D-Wise Demo	<u>46</u>
Credits	<u>48</u>
Thank you for joining	<u>49</u>

Insights from previous webinar

Audience poll

For which type of tasks are you using LLMs and MLLMs?

1. Write communication, like e.g. emails, memos
2. Improve and correct text
3. Translate text
4. Summarise text
5. Generate images
6. Generate audio files, like music
7. Generate video
8. Generate code (programming)

Insights from previous webinar

LLMs suited for many scenarios, not for others

LLMs suited for:

Great at a variety of NLP tasks

Converting unstructured data into structured data

Tackling the problem of “dark data”



Insights from previous webinar

LLMs suited for many scenarios, not for others

LLMs struggle with:

Limited reasoning

Knowledge cut-offs

Long term planning

Missing sources

Need to be “grounded” (hallucinations)

Refusals



Insights from previous webinar

Getting value out of LLMs

Improve direct interaction with LLMs

Get better results with prompting

Leverage Dark Data

Create translation layers with simple agents

Working with unstructured data

Create automation with LLM driven workflows

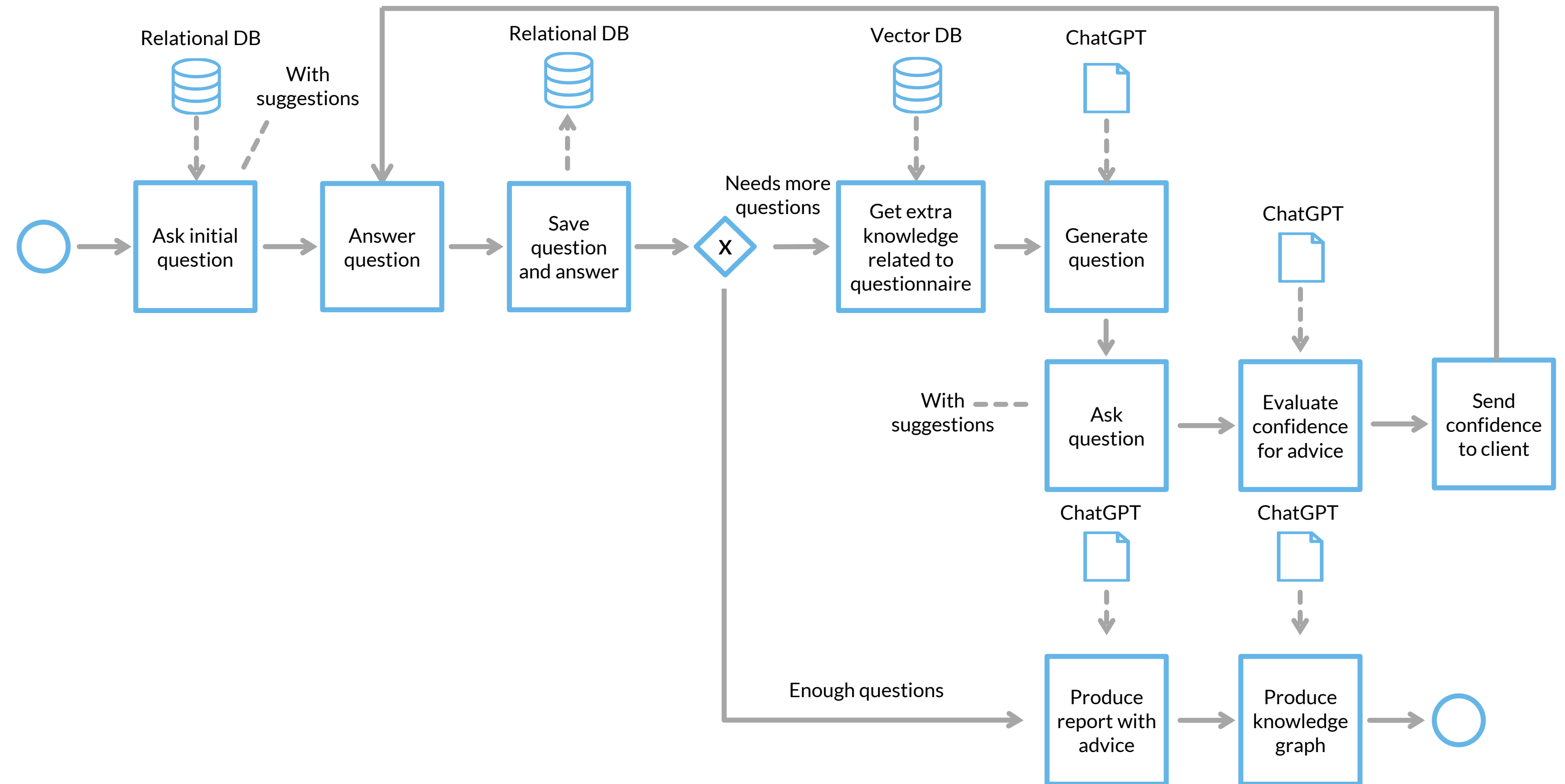
Simulate complex scenarios with agent ensembles



Audience insights

Insights from previous webinar

How we interact with LLMs (possibilities)



Agents and agent ensembles are covered in an upcoming TechTalk

In-context learning

In-context learning

LLMs can learn on the fly

LLMs can learn from a given text and answer questions about it.

Based on the following text can you describe the Mamba neural network architecture?

...

Foundation models (FMs), or large models pretrained on massive data then adapted for downstream tasks, have emerged as an effective paradigm in modern machine learning. The backbone of these FMs are often sequence models, operating on arbitrary sequences of inputs from a wide variety of domains such as language, images, speech, audio, time series, and genomics \parencitesutskever2014sequence,dosovitskiy2020image,oord2016wa venet,brown2020language,ismail2019deep,poli2023hyena. While this concept is agnostic to a particular choice of model architecture, modern FMs are predominantly based on a single type of sequence model: the Transformer \parencitevaswani2017attention and its core attention layer \parencitebahdanau2015neural The efficacy of self-



The Mamba neural network architecture represents a significant advancement in sequence modeling, designed to combine the strengths of structured state space models (SSMs) with the capabilities of modern Transformers, while addressing some of their key limitations. Here's a detailed description of its architecture and features:

Key Components of Mamba Neural Network Architecture

1. Selective State Space Models (SSMs):

- **Selection Mechanism:** Mamba incorporates a novel selection mechanism that allows the model to filter and prioritize relevant information based on the input. This mechanism is parameterized by the input, enabling the model to ignore irrelevant data and retain important information indefinitely.
- **Hardware-aware Algorithm:** To maintain computational efficiency, Mamba uses a hardware-aware algorithm that processes the model recurrently with a scan operation instead of convolution. This avoids unnecessary IO access between different levels of GPU memory, ensuring faster computations. The model scales linearly with sequence length and performs efficiently on modern hardware like A100 GPUs.

Source: ChatGPT-4o

In-context learning

LLMs can learn on the fly

Why is this extremely useful?

Allows LLMs to access private knowledge bases or very recent knowledge

Reduces hallucinations

Allows the creation of Retrieval Augmented Generation (RAG) applications

Allows to trace back sources

RAG

(Retrieval Augmented Generation)

RAG

Retrieval Augmented Generation

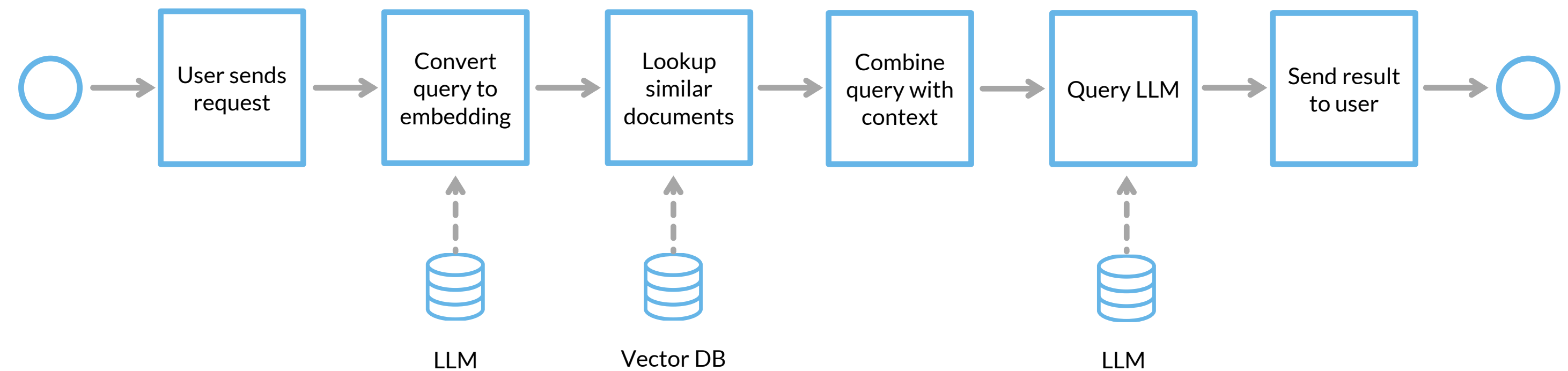
Enhance LLM knowledge with private data
and vector databases

Involved components

LLM

Vector database

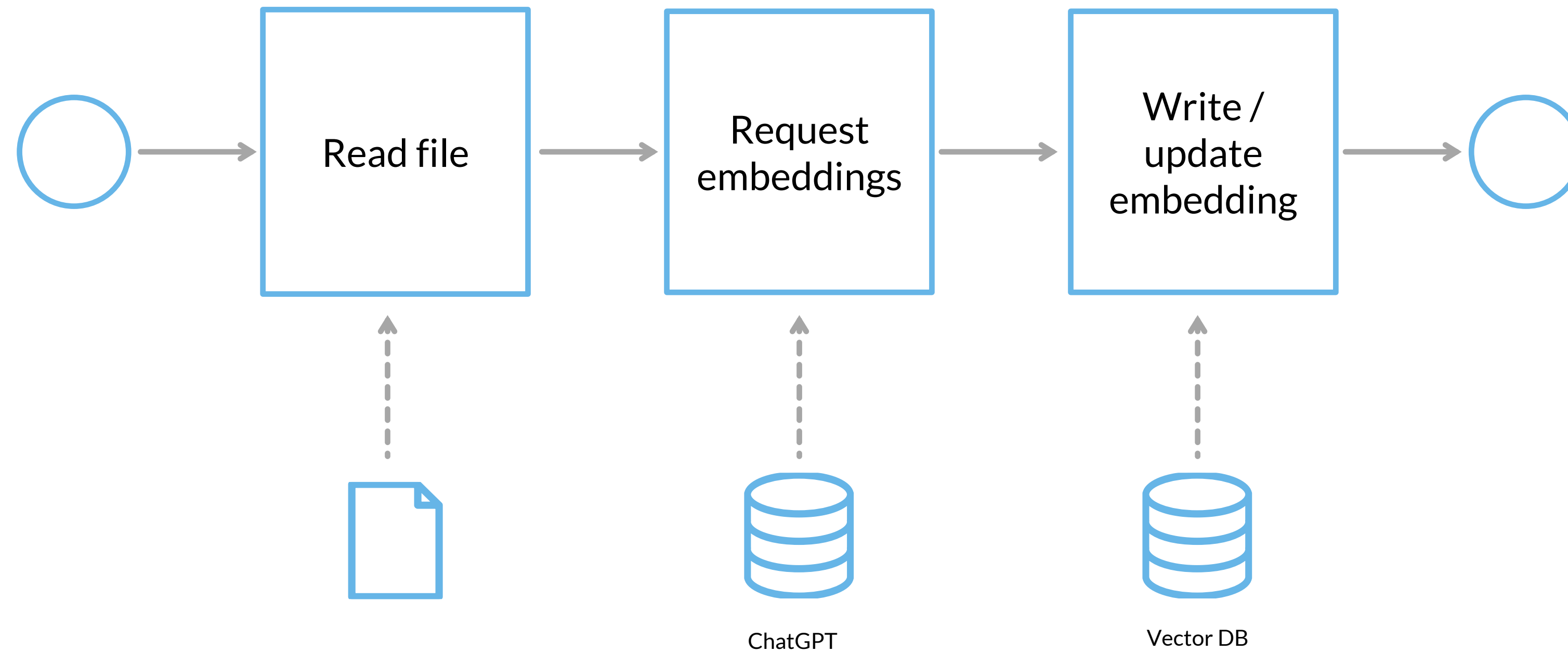
Orchestration



RAG

Retrieval Augmented Generation

RAG data needs to be “indexed”, i.e. converted to vectors which have to be generated by the LLM.



RAG

Business use cases

Talk to your documents

Search augmentation

Customer Support

Product search

Content creation

Reasoning strategies

Chain of Thoughts

Tree of Thoughts

Buffer of Thoughts

Monte Carlo Tree Self-refine

Chain of Thoughts

Step-by-step reasoning

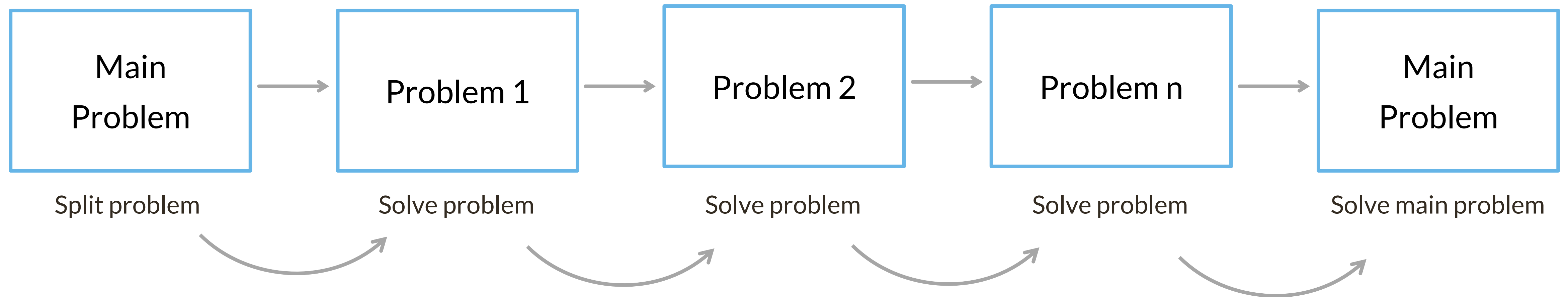
The main idea is to formulate a problem and instruct the model to break it down in sub-problems and solve these until the final answer is found

The prompt will have two parts:

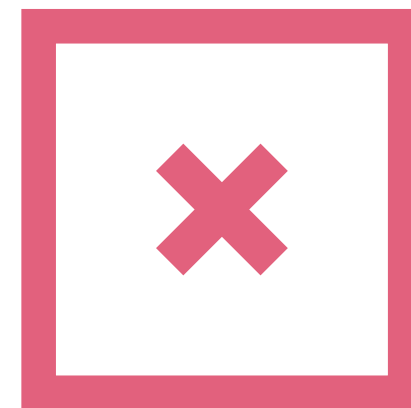
Describe the main problem

Instructions on breaking down and solving each sub-problem until the main problem is solved

Step-by-step reasoning



Example – Step 1



James likes to hear music from the Eighties and there is one band which he appreciates very much. They have this song in which the name of a female is related to an adverb. The pronoun is related to inconsistency and the person in the song is female. The band's name is somehow related to healing. Which song are we talking about?



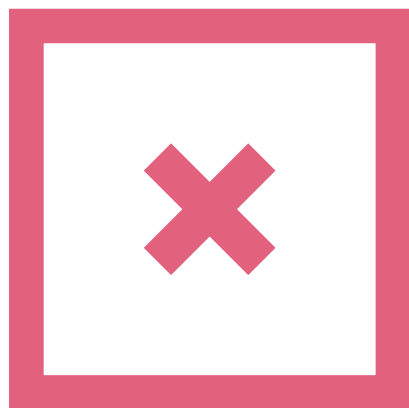
The song you're referring to is "Come On Eileen" by Dexys Midnight Runners.

- **Band:** Dexys Midnight Runners (related to healing as "Dexys" refers to dextroamphetamine, a type of medication).
- **Song:** "Come On Eileen".
- **Female name:** Eileen.
- **Adverb:** The word "Come" in "Come On Eileen" can be seen as an imperative verb, not an adverb, but it carries an urgent, sometimes inconsistent connotation.
- **Pronoun:** "Come on" implies a direct address to Eileen, and the call to action can be seen as somewhat inconsistent in its plea.

So, the song is "Come On Eileen" by Dexys Midnight Runners.

Source: ChatGPT-4o

Example – Step 2



Song wrong



Given the period (the Eighties) and the clue that the band's name is related to healing, the band in question is likely *The Cure*.

Now, considering the song by The Cure that fits your description:

- **Band:** The Cure.
- **Song:** "Why Can't I Be You?"

In this song, the name of the person isn't explicitly stated, but the pronoun "you" relates to inconsistency (the longing to be someone else), and the song's title includes an interrogative adverb ("Why"). However, it seems to miss a specific female name and the adverb connection.

Revisiting your criteria more thoroughly:

- **Band:** The Cure.
- **Song:** "Close to Me" from The Cure.

Although "Close to Me" does not explicitly mention a female name related to an adverb, the pronoun "me" is used, and the song revolves around personal feelings, which can be inconsistent.

Source: ChatGPT-4o



Band right

Example – Step 3



You got that right. It is the cure. Now focus on the name of the song. There is a female name in the song. Which could it be?



Given that the band is The Cure and we are looking for a song from the Eighties that features a female name related to an adverb and a pronoun implying inconsistency, the song in question is likely "Charlotte Sometimes."

- **Band:** The Cure
- **Song:** "Charlotte Sometimes"
- **Female name:** Charlotte
- **Adverb:** Sometimes (related to inconsistency)
- **Pronoun:** The song's narrative implies varying perspectives and feelings.

The song "Charlotte Sometimes" by The Cure fits all the criteria perfectly.

Source: ChatGPT-4o

Induced chain of thoughts

What if we could let the model create on its own the chain of thoughts?

You can use a prompt to induce this behaviour using **custom instructions**.

Induced chain of thoughts

Example:

You carefully provide factual, accurate, thoughtful and nuanced answers. You are brilliant at reasoning. If you think that you might not be correct, you say so.

You always spend a few sentences explaining the background context, assumptions and step-by-step thinking before you try to answer a question.

Mary has 3 brothers. Each of her brothers has 3 sisters. How many sisters does Mary have?



Let's break down the information given in the problem:

1. Mary has 3 brothers.
2. Each of her brothers has 3 sisters.

Given that Mary is one of those sisters, the brothers share the same sisters. Therefore, Mary and her brothers share the same siblings.

Since each brother has 3 sisters and Mary is one of those sisters, it means Mary has 2 other sisters.

So, Mary has 2 sisters.

Source: ChatGPT-4o

Induced chain of thoughts

Example:

You carefully provide factual, accurate, thoughtful and nuanced answers. You are brilliant at reasoning. If you think that you might not be correct, you say so.

You always spend a few sentences explaining the background context, assumptions and step-by-step thinking before you try to answer a question.

Mary has 3 brothers. Each of her brothers has 3 sisters. How many sisters does Mary have?



Let's break down the information given in the problem:

1. Mary has 3 brothers.
2. Each of her brothers has 3 sisters.

Given that Mary is one of those sisters, the brothers share the same sisters. Therefore, Mary and her brothers share the same siblings.

Since each brother has 3 sisters and Mary is one of those sisters, it means Mary has 2 other sisters.

So, Mary has 2 sisters.

Source: ChatGPT-4o

ChatGPT o1-preview and o1-mini

Operate in 2 steps: reasoning and response.

Reasoning step creates action plan

Response step generates the answer

This is about Game of 24. Can you solve the following game: 5, 5, 5, 7



Thought for 33 seconds ^

Formulating combinations

I'm digging into arithmetic operations with 5, 5, 5, and 7 to hit 24. $(5*5)-(5/7)$ comes close but overshoots to 24.2857.

Evaluating numerical combinations

I'm testing different arithmetic combinations of the numbers 5 and 7 to find one that equals 24, including multiplication, division, and addition. None achieve 24.

Trying another way

I'm working through various combinations of numbers and operations, but I'm not hitting 24 yet. Each attempt seems off the mark.

Assessing various alternatives



I'm exploring different mathematical methods such as arithmetic operations, exponentiation, and

Source: ChatGPT o1-preview

ChatGPT o1-preview and o1-mini

Experience: slower, but much better at reasoning

However it still makes mistakes, so be aware of that.

Back to Game of 24. Can you solve it with these numbers: 5, 3, 5, 3?

ChatGPT o1-mini ▾

Alternative Representation:

Another way to represent the same solution is:

$$(3 \times 5) \times \left(\frac{5+3}{5} \right) = 15 \times \frac{8}{5} = 15 \times 1.6 = 24$$

Conclusion:

By strategically combining addition, division, and multiplication, we've successfully reached the target number **24** using the numbers **5, 3, 5, 3**. This solution adheres to the rules of the Game of 24 by using each number exactly once and employing standard arithmetic operations.

Answer:

Yes, one way to make 24 using 5, 3, 5, and 3 is: $3 \times 5 \times \left(\frac{5+3}{5} \right) = 24$

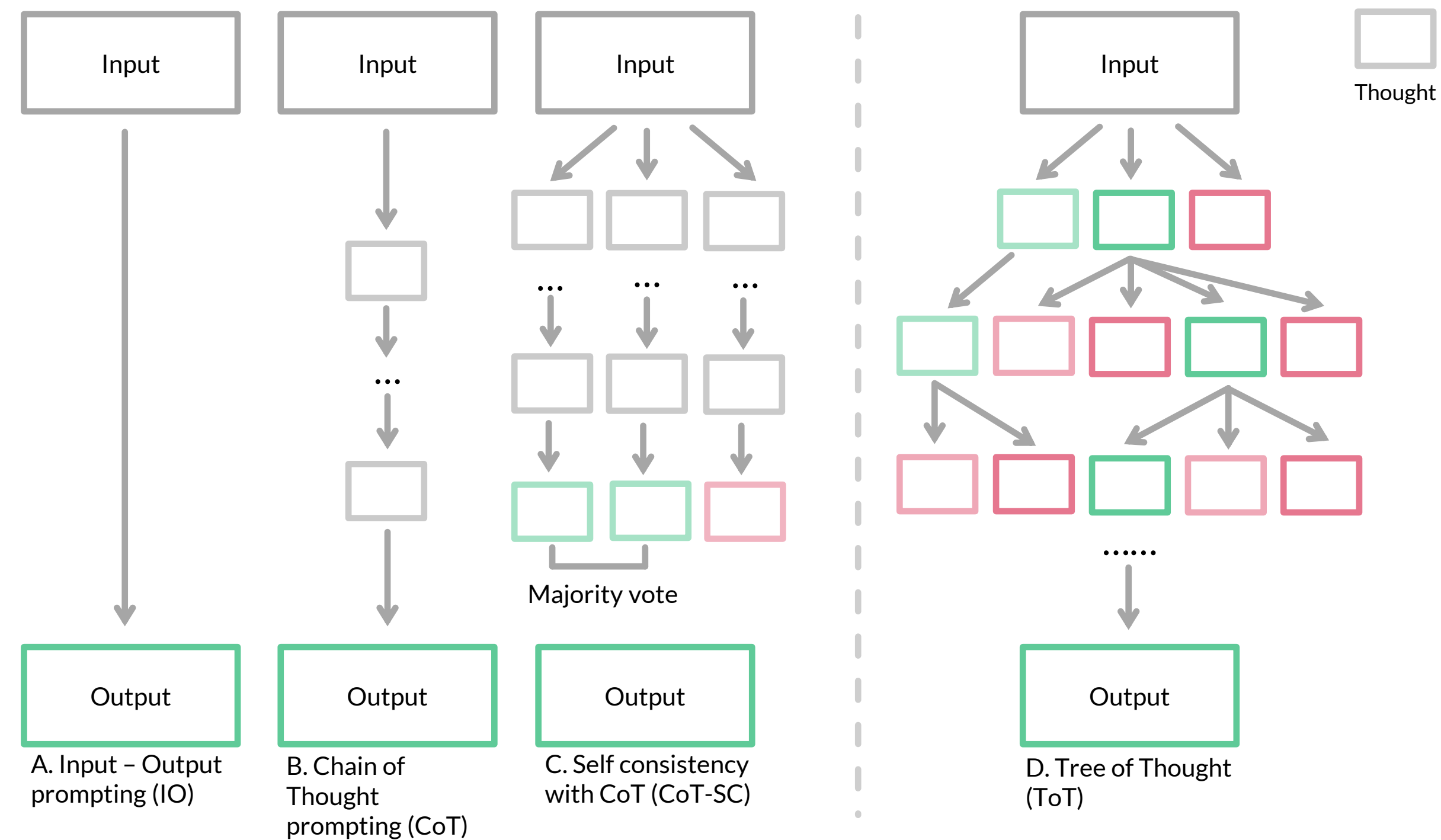
🔊 📄 👍 🗨️ ↺ ▾

Source: ChatGPT o1-mini

Tree of Thoughts

Step-by-step with multiple options

Instruct the LLM to break down the problem in steps and then to solve each step iteratively.



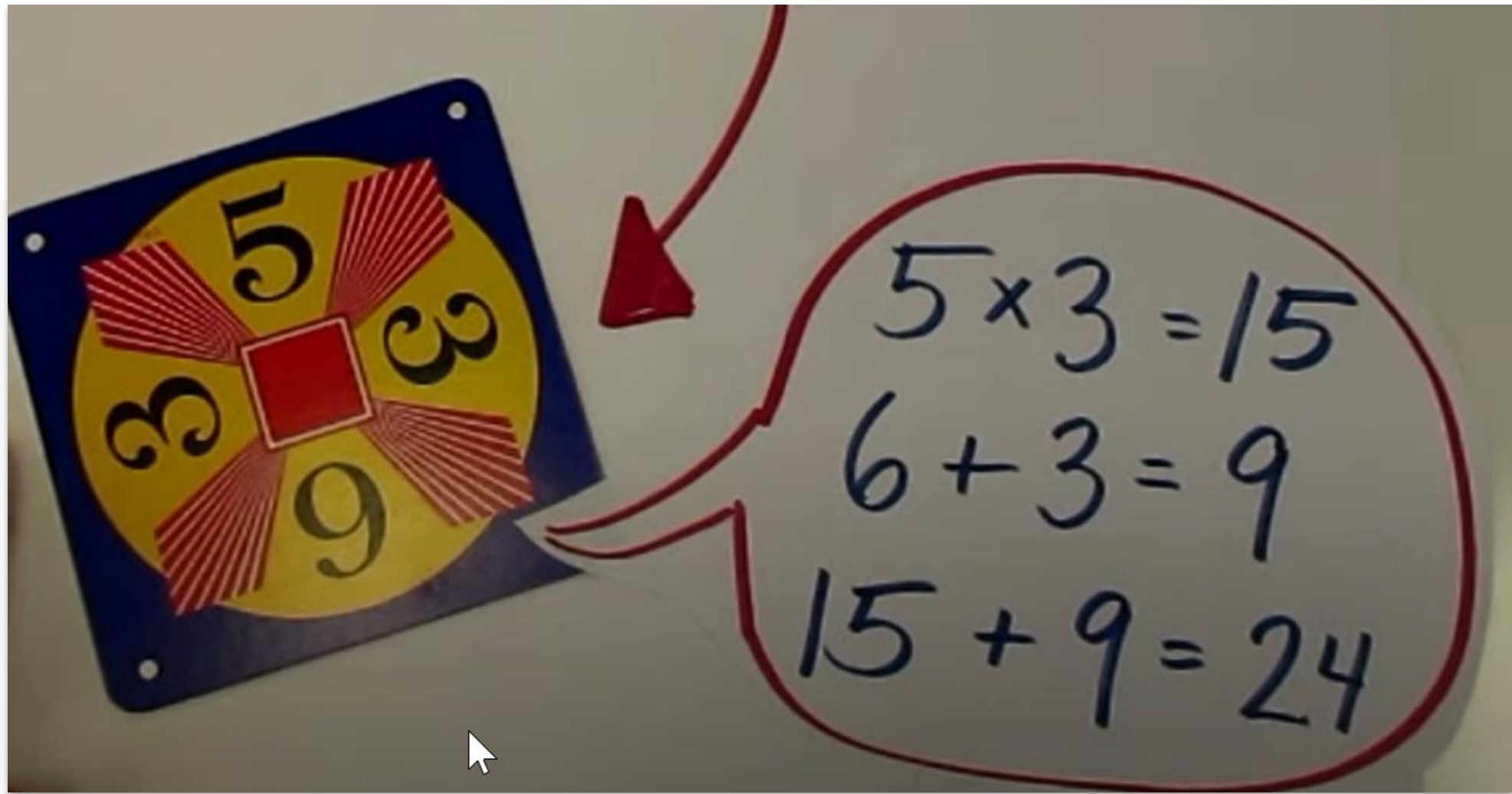
Source: <https://github.com/princeton-nlp/tree-of-thought-llm>

Audience poll

Please detect the game of 24 combination that cannot be solved
For example for 6, 2, 16, 6 can be solved. Possible solution is $(6 * 16) / (6 - 2)$

1. 7, 2, 16, -6
2. 5, 3, 3, 6
3. 6, 6, 6, 2
4. 5, 5, 5, 7
5. All are right
6. Don't know...I give up

Game of 24 – simple prompts go wrong (1 of 2)



Source: <https://www.youtube.com/watch?v=fWEEu9JxyYo&t=28s>
How to play a math game

No, can you give me one solution when you have these four numbers: 7, 2, 6 and 4?

Sure! Let's find a solution for the numbers 7, 2, 6, and 4 to make 24 using the basic arithmetic operations. Here is one possible solution:

$$(7 - (2/2)) * (6/3) = 24$$

Breaking it down step by step:

1. $2/2 = 1$
2. $7 - 1 = 6$
3. $6/3 = 2$
4. $6 * 2 = 24$

So, one valid solution is:

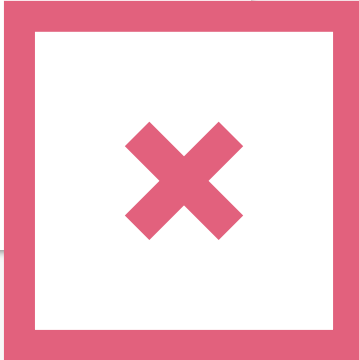
$$(7 - (2/2)) * (6/3) = 24$$

Oops, using 2 twice

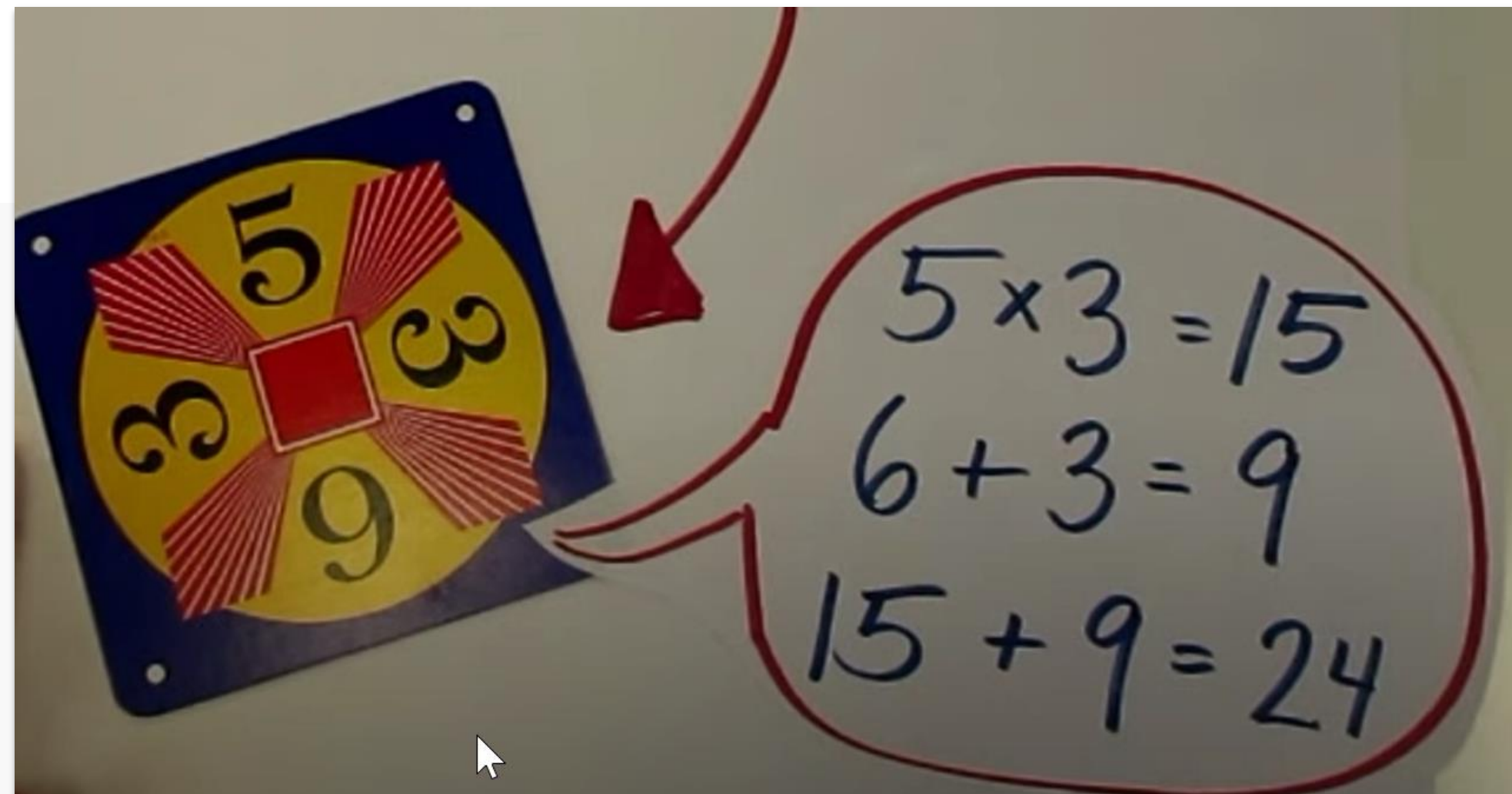
Oops, I said 3 not 4

Oops, should be 12

Source: ChatGPT-4o



Game of 24 – simple prompts go wrong (2 of 2)



Source: <https://www.youtube.com/watch?v=fWEEu9JxyYo&t=28s>
How to play a math game

And a solution for 7, 2, -6, 16?

For the numbers 7, 2, -6, and 16, let's work through possible operations to reach 24.

One possible solution is:

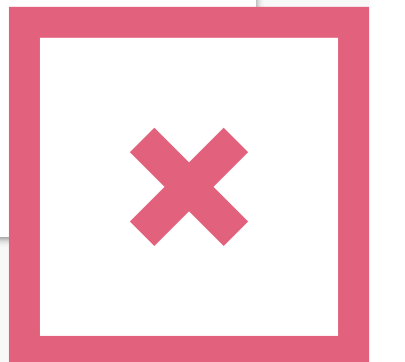
$$16 - (7 - 2) - (-6) = 16 - 5 + 6 = 24$$

Step-by-step breakdown:

1. Subtract 7 from 2 to get $7 - 2 = 5$.
2. Subtract this result from 16: $16 - 5 = 11$.
3. Subtract -6 from the result (which is the same as adding 6): $11 + 6 = 24$.

Thus, this combination satisfies the game's rules and results in 24.

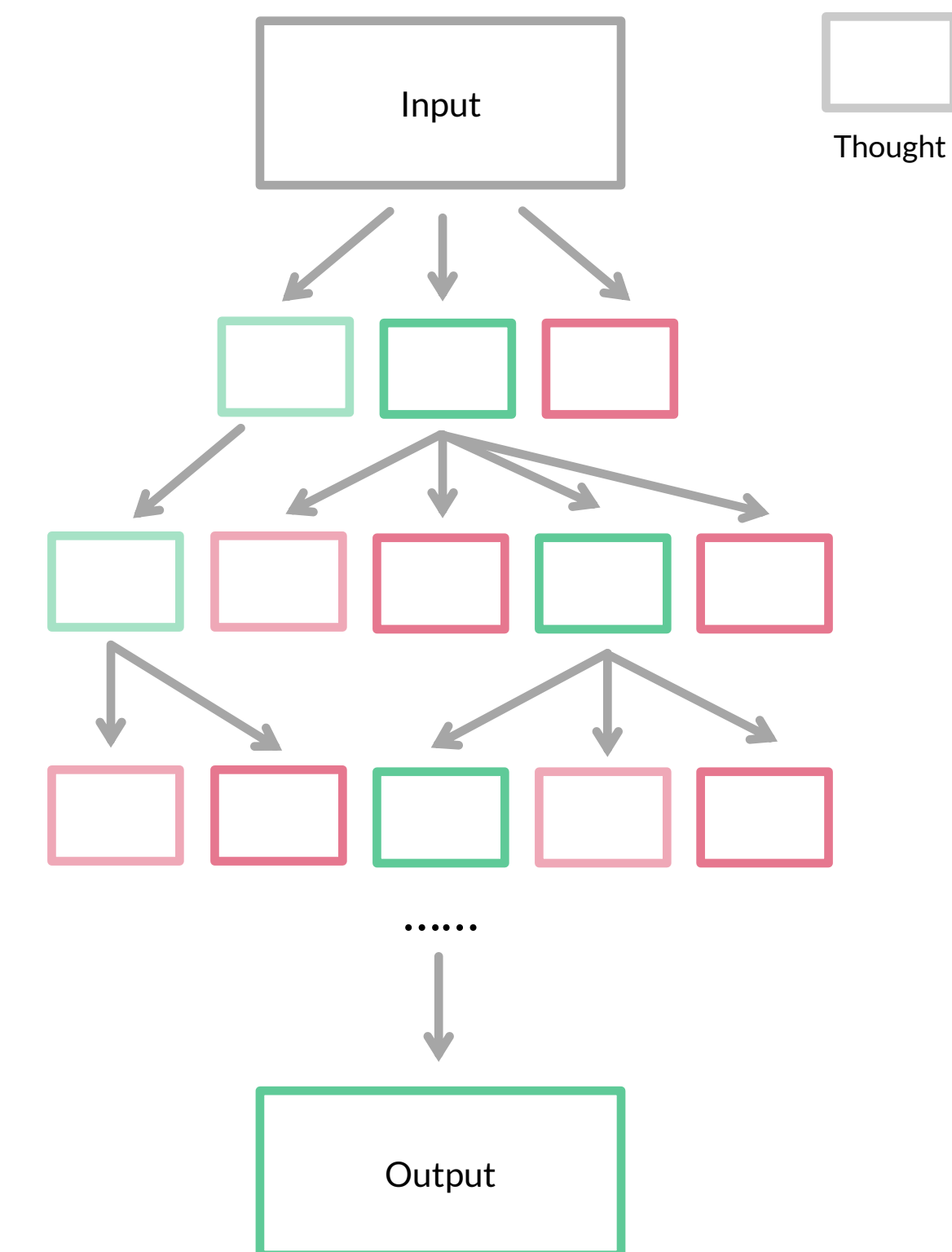
Source: ChatGPT-4o



The idea behind Tree of Thoughts

Tree of Thoughts is an iterative process which allows to solve problems in a combinatorial problem space.

At each step one number disappears from the equation until we have a solution.



Source: <https://github.com/princeton-nlp/tree-of-thought-llm>

The idea behind Tree of Thoughts

Propose

Create potential solutions for the current step

Values

Calculate the values for the potential solutions

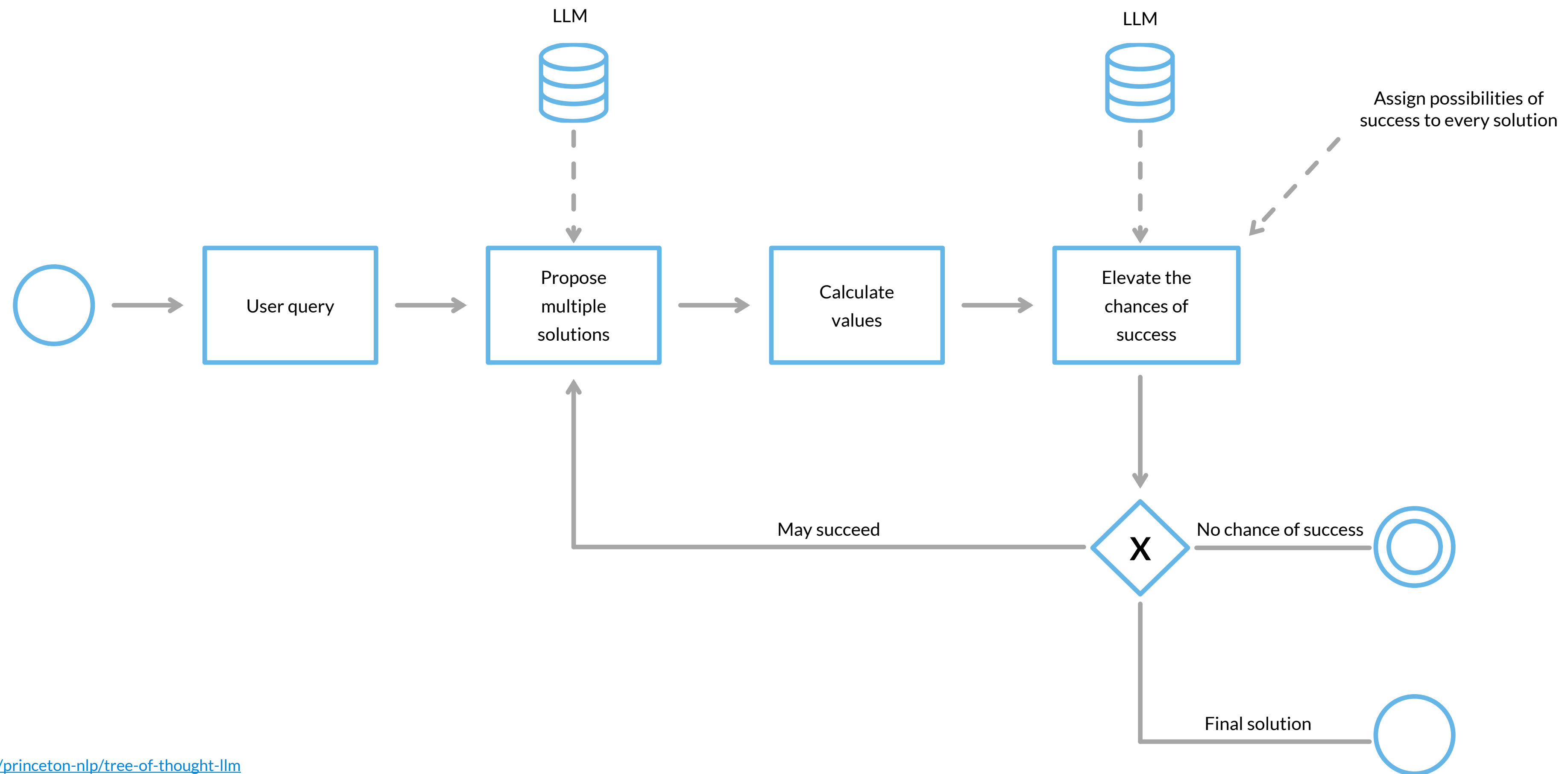
Evaluate

Evaluate the success probability for each solution

Sample

Get the best evaluated results

The workflow

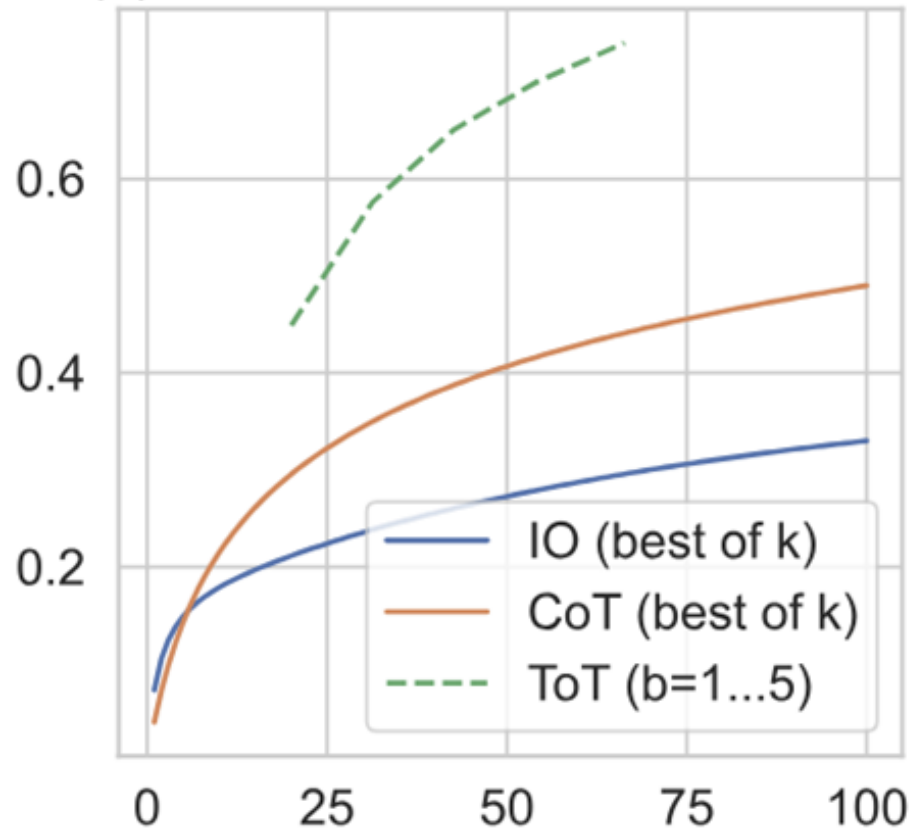


Source: <https://github.com/princeton-nlp/tree-of-thought-llm>

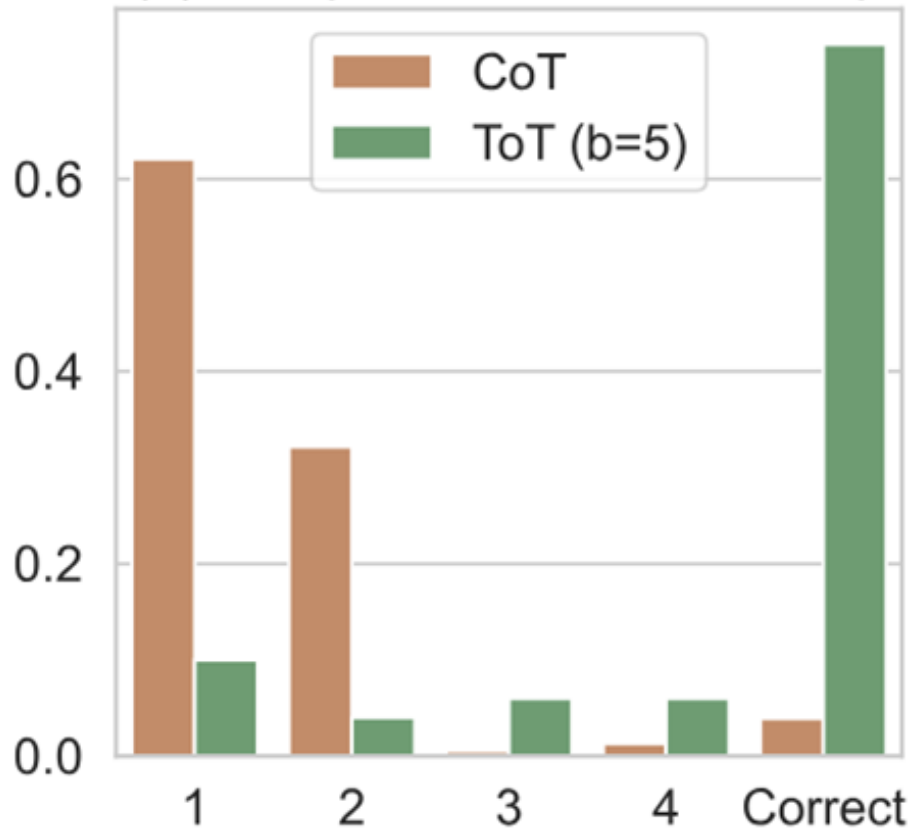
Performance ToT versus other strategies

Method	Success
IO prompt	7.3%
CoT prompt	4.0%
CoT-SC (k=100)	9.0%
ToT (ours) (b=1)	45%
ToT (ours) (b=5)	74%
IO + Refine (k=10)	27%
IO (best of 100)	33%
CoT (best of 100)	49%

(a) Success rate with nodes visited



(b) Samples failed at each step



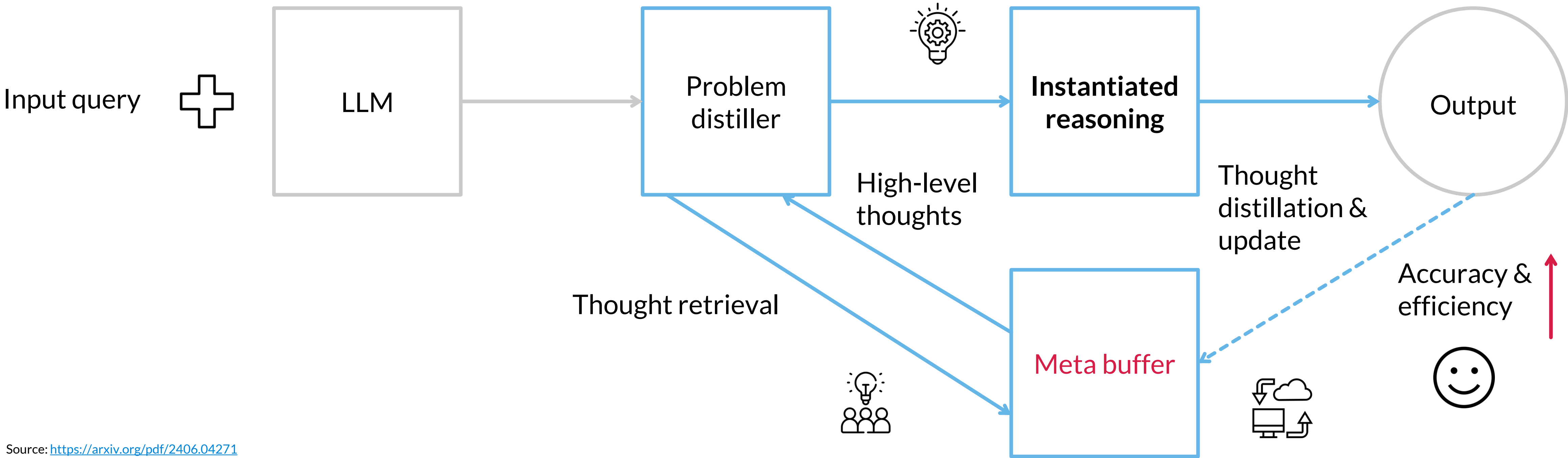
Source: <https://github.com/princeton-nlp/tree-of-thought-llm>

Buffer of Thoughts

Performance ToT versus other strategies

Released in June 2024

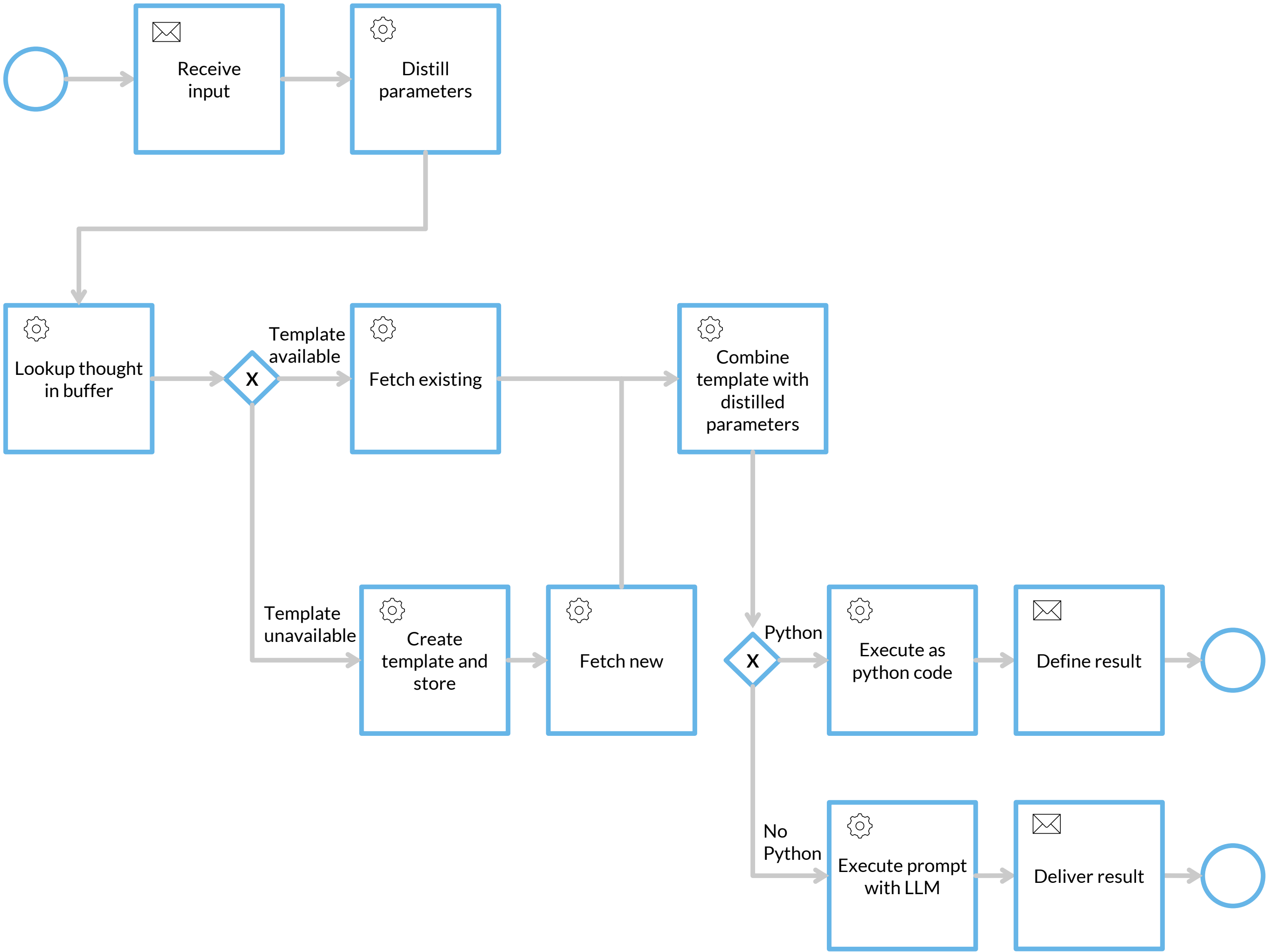
Simple workflow to solve problems with a pre-established buffer of “thoughts”



Source: <https://arxiv.org/pdf/2406.04271>

Detailed workflow

- Distill problem parameters
- Lookup template (prompt template) in buffer
- Combine template and parameters
- Execute template



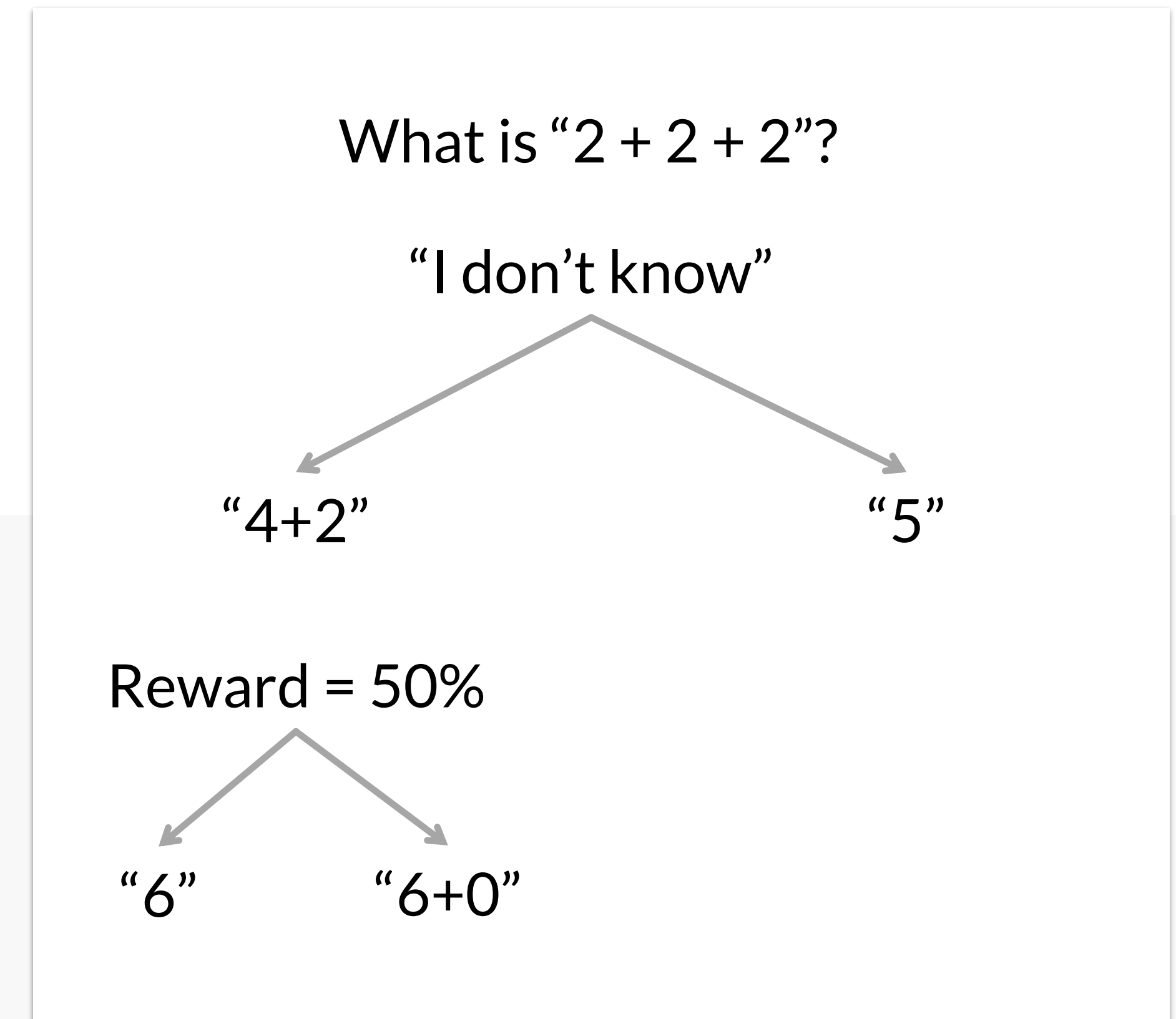
Monte Carlo Tree Self-refine

Techniques for improved reasoning with LLMs (1 of 2)

Boosts the reasoning capabilities of LLMs

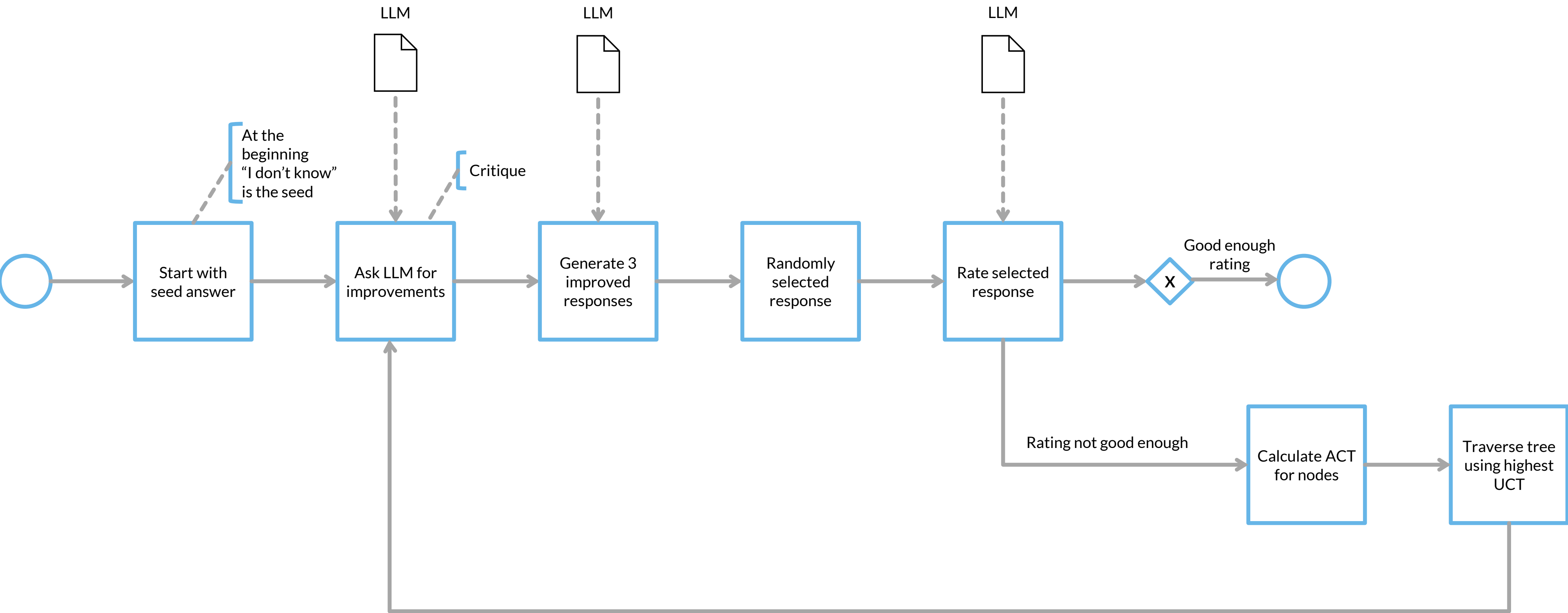
Help to solve very complex problems with even “small” LLMs.

Using a recursive answer improvement and rating pipeline you can create reasoning logical “trees” which boost the model’s reasoning capabilities.



Source: www.youtube.com/watch?v=mfAV_bigdRA

Techniques for improved reasoning with LLMs (2 of 2)



Source: arxiv.org/abs/2406.07394



Audience insights

7, 2, 16, -6

$(2 * 7) + (16 - 6)$

6, 6, 6, 2

$(6 * 6) - (6 * 2)$

5, 3, 3, 6

$((5 * 3) + 3) + 6$

 5, 5, 5, 7

Does not form 24

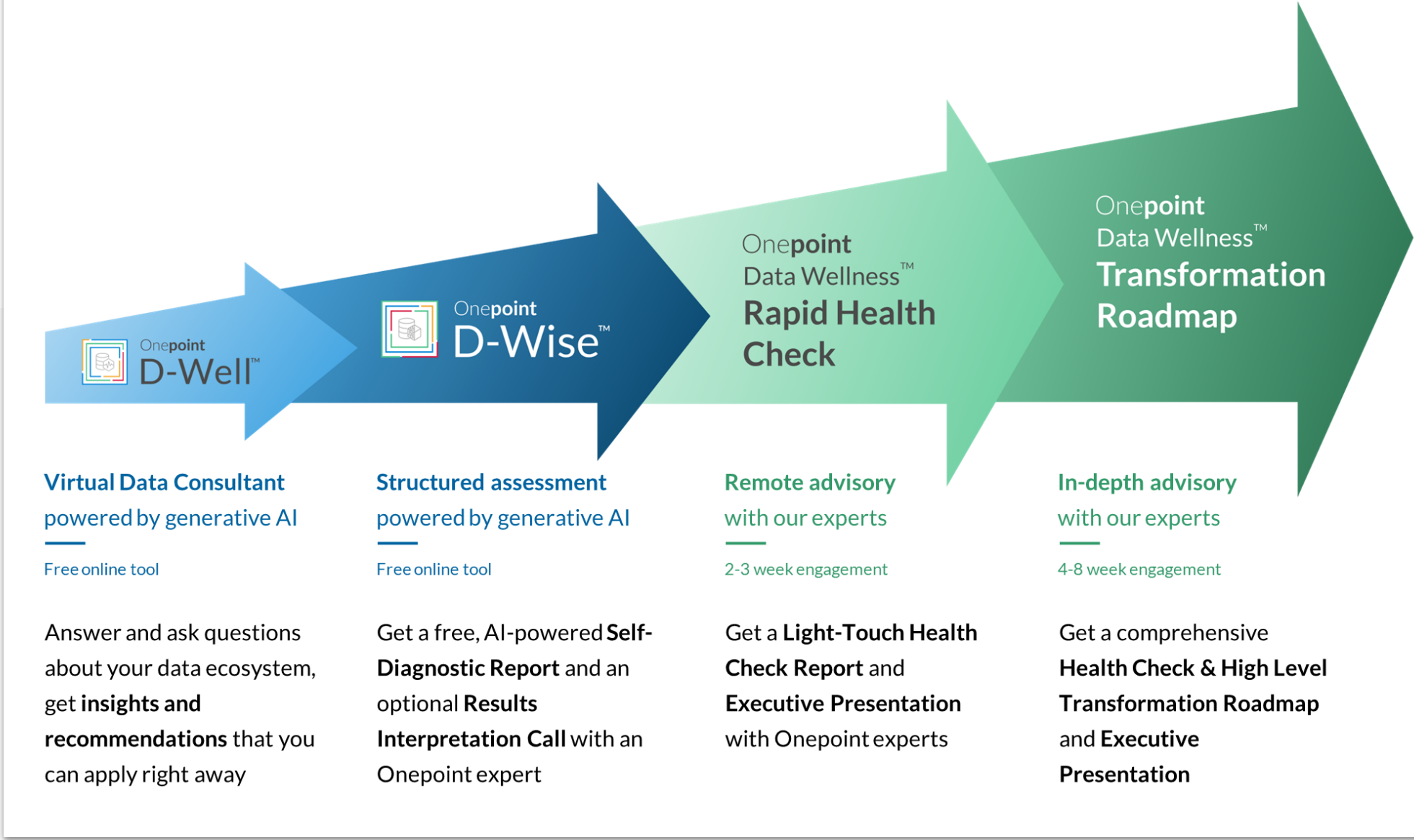
Live demo

Free AI tool

Explore your data health with Onepoint D-Wise™



Embark on your data wellness journey
with our free tool, powered by generative
AI and the Onepoint Data & Analytics
Body of Knowledge™



Scan the QR code or visit www.onepointltd.com/data-wellness

Credits



chatgpt.com

Tree of Thoughts

github.com/princeton-nlp/tree-of-thought-llm

Buffer of Thoughts

arxiv.org/pdf/2406.04271

Accessing GPT-4 level Mathematical Olympiad Solutions via Monte Carlo Tree Self-refine with LLaMa-3 8B

arxiv.org/pdf/2406.07394 & www.youtube.com/watch?v=mfAV_bigdRA

Thank you for joining

Please feel free to contact Gil Fernandes if you have any feedback about the session.



Email at techtalk@onepointltd.com

Connect on LinkedIn www.linkedin.com/in/gil-palma-fernandes

Find Gil's Reflections on AI at medium.com/@gil.fernandes



Onepoint
TechTalk

Next webinar coming soon!

Details will be announced soon. Stay tuned!

Onepoint is a boutique,
values-driven, business-
oriented technology
consultancy.

We architect, prototype, build, and
manage data and AI powered
solutions. We partner with global
clients looking for high-impact,
enterprise-grade advice and IT
services to realise their most critical
digital transformations.

London | Manchester | Pune

Onepoint

Your trusted companions for the digital journey™



onepointltd.com
hello@onepointltd.com



© Copyright 2024. Onepoint Consulting Ltd is a company registered in England.
Company registration number 5516457. VAT registration number GB866120039.
ISO27001 certified. Certified in the UK as an ethnic minority-owned business.