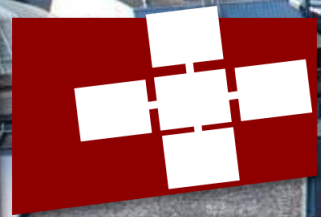


M. BESTA, T. HOEFLER

WITH N. BLACH, A. KUBICEK, R. GERSTENBERGER, AND MANY OTHERS

Graph of Thoughts: Solving Elaborate Problems with Large Language Models



Brian Lamacraft

Jan 5 · 2 min read · Member-only



Chat GTP-4 Could Pass the Bar Exam

How Our Technology Evolves FAST

Source: <https://medium.com/>



Photo by Maximalfocus on Unsplash

Brian Lamacraft
Jan 5

2 min read · Member-only



Chat GTP-4 Could Pass the Bar Exam

How Our Technology Evolves FAST

Source: <https://medium.com/>

AI chatbot's MBA exam pass poses test for business schools

ChatGPT earned a solid grade and outperformed some humans on a Wharton course









A professor at the University of Pennsylvania's Wharton School said ChatGPT earned a 'B to B-' on his operations management test
© Bloomberg

Source: <https://www.ft.com/>

Andrew Jack in New York | JANUARY 21 2023

292

 **Brian Lamacraft**
Jan 5 · 2 min read · Member-only ·     

Chat GTP-4 Could Pass the Bar Exam

How Our Technology Evolves FAST

Source: <https://medium.com/>

AI chatbot's MBA exam pass poses test for business schools

ChatGPT earned a solid grade and outperformed some humans on a Wharton course

AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

by Michael DePeau-Wilson, Enterprise & Investigative Writer, MedPage Today **January 19, 2023**

Source: <https://www.medpagetoday.com/>





ADVERTISEMENT

Capabilities

- Remembers what user said earlier in the conversation
- Allows user to provide follow-up questions

Limitations

- May occasionally generate incorrect information
- May occasionally produce harmful instructions

Andrew Jack in New York **JANUARY 21 2023**  292 

 **Brian Lamacraft**
Jan 5 · 2 min read · Member-only ·     

Chat GTP-4 Could Pass the Bar Exam

How Our Technology Evolves FAST

Source: <https://medium.com/>

AI chatbot's MBA exam pass poses test for business schools

ChatGPT earned a solid grade and outperformed some humans on a Wharton course

AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

by Michael DePeau-Wilson, Enterprise & Investigative Writer, MedPage Today **January 19, 2023**

Source: <https://www.medpagetoday.com/>

ADVERTISEMENT





Study finds ChatGTP outperforms physicians in providing high-quality, empathetic advice to patient questions

Date: April 28, 2023 <https://www.sciencedaily.com/>

Source: University of California - San Diego

Summary: A new study provides an early glimpse into the role that AI assistants could play in medicine. The research compared written responses from physicians with those from ChatGPT to real-world health questions. A panel of licensed healthcare professionals preferred ChatGPT's responses 79% of the time.

Andrew Jack in New York **JANUARY 21 2023**  292 

Brian Lamacraft
Jan 5 · 2 min read · Member-only

Chat GTP-4 Could Pass the Bar Exam

How Our Technology Evolves FAST

Source: <https://medium.com/>

AI chatbot's MBA exam pass poses test for business schools



ChatGPT earned a solid grade and outperformed some humans on a Wharton course

AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

by Michael DePeau-Wilson, Enterprise & Investigative Writer, MedPage Today January 19, 2023


Source: <https://www.medpagetoday.com/>



ChatGPT Passes Google Coding Interview for Level 3 Engineer With \$183K Salary

'Amazingly, ChatGPT gets hired at L3 when interviewed for a coding position,' reads a Google document, but ChatGPT itself says it can't replicate human creativity and problem-solving skills.

PC by Emily Dreibelb February 01, 2023



Andrew Jack in New York JANUARY 21 2023

What is left for us humans?

Brian Lamacraft
Jan 5 · 2 min read · Member-only

Chat GTP-4 Could Pass the Bar Exam

How Our Technology Evolves FAST

Source: <https://medium.com/>

AI chatbot's MBA exam pass poses test for business schools


ChatGPT earned a solid grade and outperformed some humans on a Wharton course

AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

by Michael DePeau-Wilson, Enterprise & Investigative Writer, MedPage Today January 19, 2023


Source: <https://www.medpagetoday.com/>



ChatGPT Passes Google Coding Interview for Level 3 Engineer With \$183K Salary

'Amazingly, ChatGPT gets hired at L3 when interviewed for a coding position,' reads a Google document, but ChatGPT itself says it can't replicate human creativity and problem-solving skills.

PC by Emily Dreibel February 01, 2023



Andrew Jack in New York JANUARY 21 2023

What is left for us humans?

SCIENTIFIC
AMERICAN

BEHAVIOR

What Is the Special Something That Makes the Human Mind Unique?

The capacity to engage in shared tasks such as hunting large game and building cities may be what separated modern humans from our primate cousins

By Gary Stix on October 1, 2016

IN BRIEF

- Humans—it was once thought—differed from other animals by their use of tools and their overall superiority in a range of cognitive abilities. Close observation of the behaviors of chimpanzees and other great apes has proved these ideas to be wrong.
- Chimpanzees score as highly as young children on tests of general reasoning abilities but lack many of the social skills that come naturally to their human cousins. Unlike humans, chimps do not collaborate in the large groups needed to build complex societies.
- Comparison of human and chimp psychology reveals that an essential source of the differences in humans may be the evolution of the ability to intuit what another person is thinking so that both can work toward a shared goal.



Brian Lamacraft
Jan 5 · 2 min read · Member-only

Chat GTP-4 Could Pass the Bar Exam

How Our Technology Evolves FAST

Source: <https://medium.com/>

AI chatbot's MBA exam pass poses test for business schools

ChatGPT earned a solid grade and outperformed some humans on a Wharton course

AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

by Michael DePeau-Wilson, Enterprise & Investigative Writer, MedPage Today January 19, 2023

Source: <https://www.medpagetoday.com/>

ChatGPT Passes Google Coding Interview for Level 3 Engineer With \$183K Salary

'Amazingly, ChatGPT gets hired at L3 when interviewed for a coding position,' reads a Google document, but ChatGPT itself says it can't replicate human creativity and problem-solving skills.

PC by Emily Dreibelb Feb 01, 2023



Andrew Jack in New York JANUARY 21 2023

What is left for us humans?

Brian Lamacraft
Jan 5 · 2 min read · Member-only

Chat GTP-4 Could Pass the Bar Exam

How Our Technology Evolves FAST

Source: <https://medium.com/>

AI chatbot's MBA exam pass poses test for business schools

ChatGPT earned a solid grade and outperformed some humans on a Wharton course

AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

by Michael DePeau-Wilson, Enterprise & Investigative Writer, MedPage Today **January 19, 2023**

Source: <https://www.medpagetoday.com/>

-
-
-
-
-

ChatGPT Passes Google Coding Interview for Level 3 Engineer With \$183K Salary

'Amazingly, ChatGPT gets hired at L3 when interviewed for a coding position,' reads a Google document, but ChatGPT itself says it can't replicate human creativity and problem-solving skills.

PC by Emily Dreibel **Feb 01, 2023**



Andrew Jack in New York **JANUARY 21 2023**

Bing Chat Outshines ChatGPT with Cutting-Edge "Theory of Mind" Capabilities

SHAUN **Feb 13, 2023** Updated • Feb 13, 2023 Apps | 7 D

Oh, ChatGPT, if you ever develop a consciousness, please remember with joy all the fun you're having with humanity nowadays. Especially if you ever consider squeezing out our life force while being locked in a vat. As you surely know, ChatGPT has been on everybody's minds lately thanks to its incredible capabilities. You probably know about ChatGPT's ability to write a poem, develop software, and even write phishing emails for all those wannabe criminals out there.





2018 ACM A.M. Turing Lecture

June 23, 2019

5:15pm MST



Geoffrey Hinton



Yann LeCun



“Really the deciding factor [for the AI revolution] was the increase in compute power” (26:50)



2018 ACM A.M. Turing Lecture

June 23, 2019

5:15pm MST



Geoffrey Hinton



Yann LeCun



“Really the deciding factor [for the AI revolution] was the increase in compute power” (26:50)

“I think a lot of the credit for deep learning goes to [... others ...] and the people who made the computers go fast.” (27:00)

2018 ACM A.M. Turing Lecture

June 23, 2019

5:15pm MST



Geoffrey Hinton



Yann LeCun



How do we “Make Computers go Fast”?

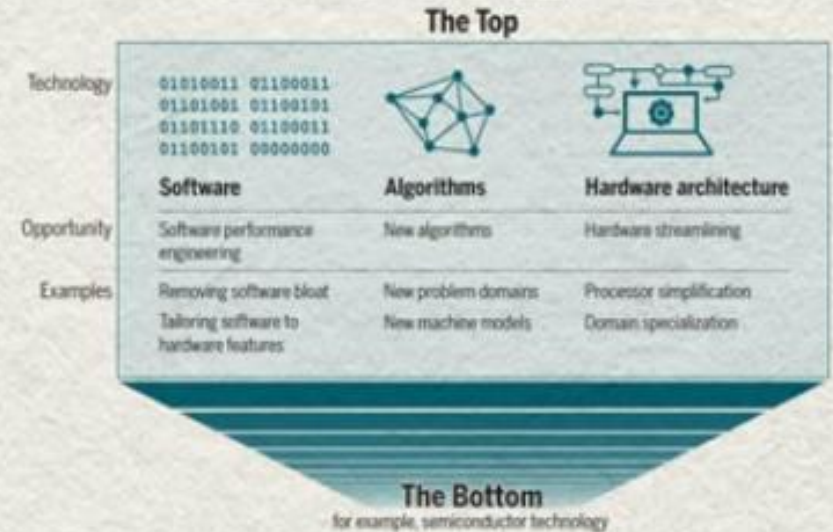
How do we “Make Computers go Fast”?

2021 Turing award – Jack Dongarra The Take Away

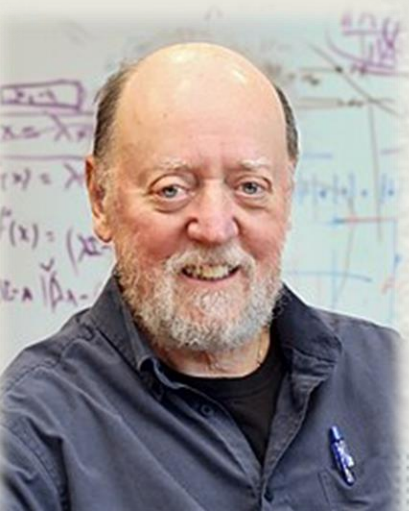
- HPC Hardware is Constantly Changing
 - Scalar
 - Vector
 - Distributed
 - Accelerated
 - Mixed precision
- Three computer revolutions
 - High performance computing
 - Deep learning
 - Edge & AI
- Algorithm / Software advances follows hardware
 - And there is “plenty of room at the top”

“There’s plenty of room at the Top: What will drive computer performance after Moore’s law?”

Leiserson et al., *Science* **368**, 1079 (2020) 5 June 2020



Leiserson et al., *Science* **368**, 1079 (2020) 5 June 2020



<https://www.youtube.com/watch?v=lsnRP9akCDk>

How do we “Make Computers go Fast”?

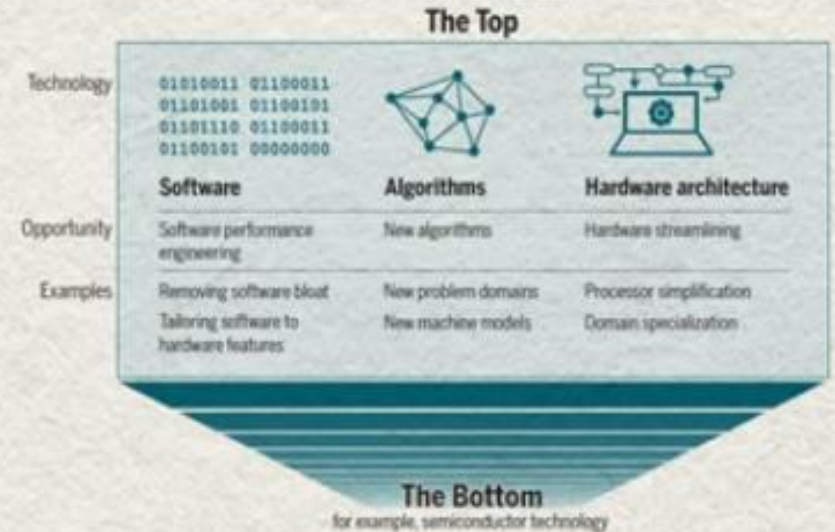
2021 Turing award – Jack Dongarra The Take Away

Supercomputers are very (>70%) efficient at dense linear algebra!

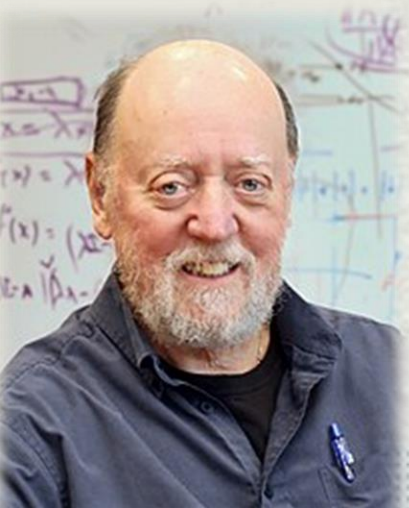
- HPC Hardware is Constantly Changing
 - Scalar
 - Vector
 - Distributed
 - Accelerated
 - Mixed precision
- Three computer revolutions
 - High performance computing
 - Deep learning
 - Edge & AI
- Algorithm / Software advances follows hardware
 - And there is “plenty of room at the top”

“There’s plenty of room at the Top: What will drive computer performance after Moore’s law?”

Leiserson et al., *Science* **368**, 1079 (2020) 5 June 2020



Leiserson et al., *Science* **368**, 1079 (2020) 5 June 2020



<https://www.youtube.com/watch?v=lsnRP9akCDk>

FINANCIAL TIMES

Artificial intelligence

+ Add to myFT

The billion-dollar bet to reach human-level AI

OpenAI believes that huge computing power is key driver

In the race to build a machine with human-level intelligence, it seems, size really matters.

“We think the most benefits will go to whoever has the biggest computer,” said Greg Brockman, chairman and chief technology officer of OpenAI.

The San Francisco-based AI research group, set up four years ago by tech industry luminaries including Elon Musk, Peter Thiel and Reid Hoffman, has just thrown down a challenge to the rest of the AI world.

Richard Waters in San Francisco AUGUST 3 2019

 140 

Supercomputers fuel Modern AI

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT [@FredericLambert](#)

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT [@FredericLambert](#)

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT [@FredericLambert](#)

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence

Building artificial general intelligence is OpenAI's ambitious goal

By James Vincent | Jul 22, 2019, 10:08am EDT

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT [@FredericLambert](#)

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

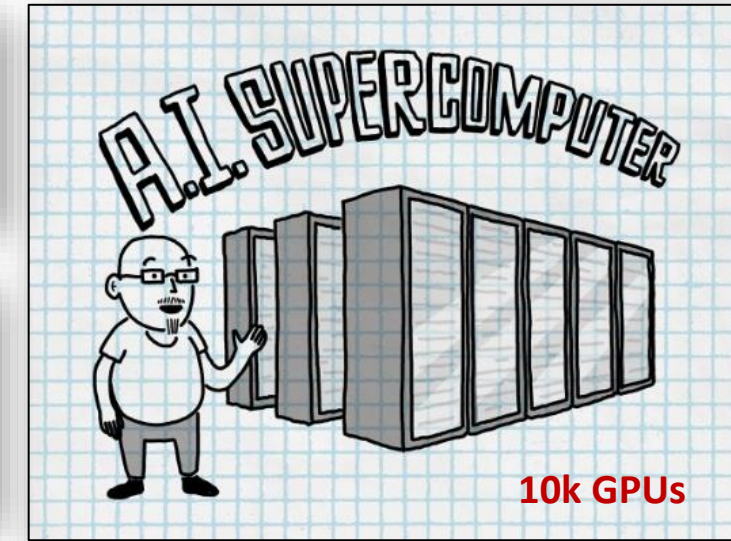
The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence

Building artificial general intelligence is OpenAI's ambitious goal

By James Vincent | Jul 22, 2019, 10:08am EDT



Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM [Share](#)

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

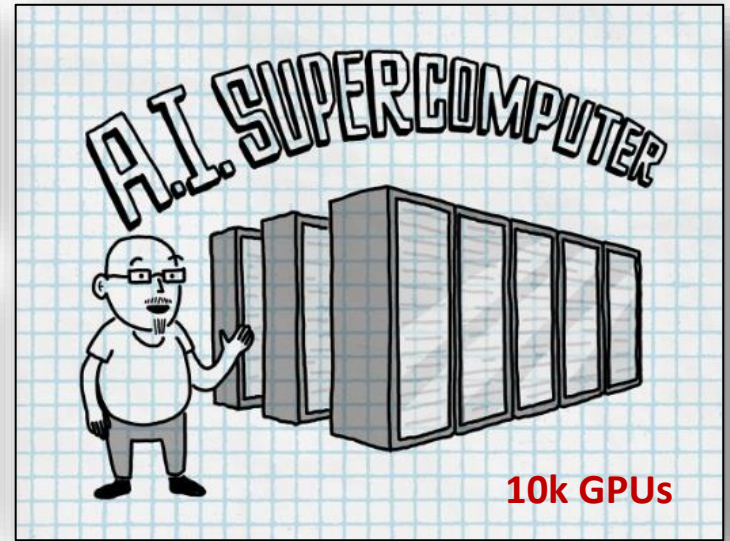
The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence

Building artificial general intelligence is OpenAI's ambitious goal

By James Vincent | Jul 22, 2019, 10:08am EDT



Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT [@FredericLambert](#)



A robot may ___ injure a human being or, through inaction, allow a human being to come to harm.

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

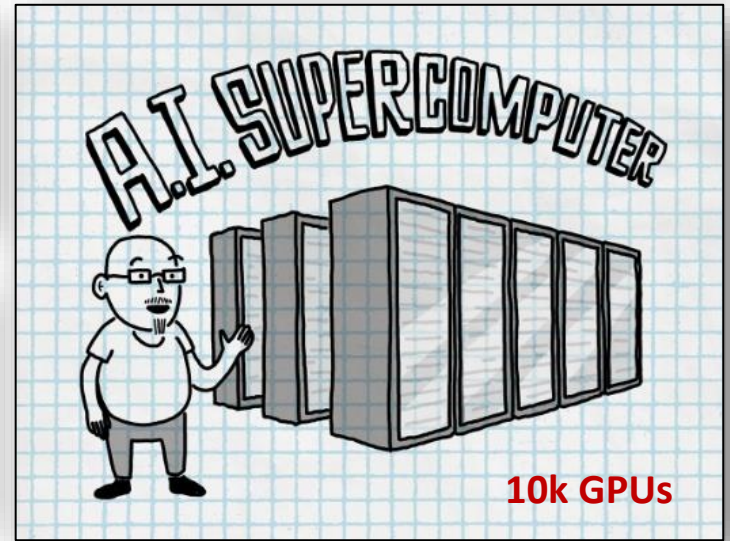
Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence

Building artificial general intelligence is OpenAI's ambitious goal

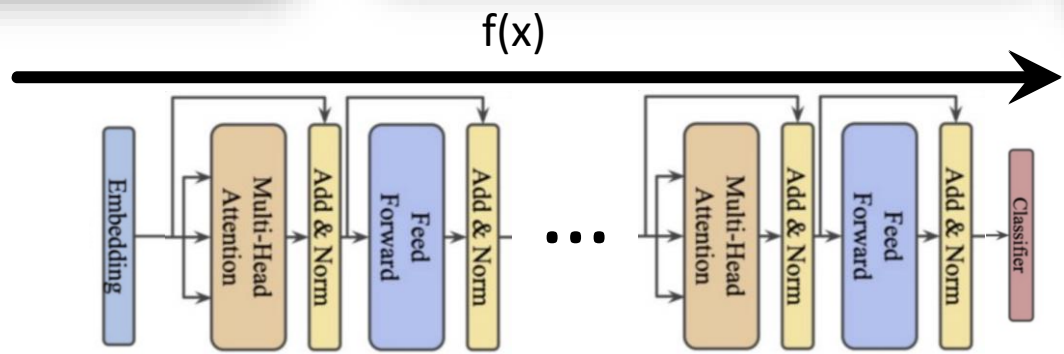
By James Vincent | Jul 22, 2019, 10:08am EDT

Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT @FredericLambert



A robot may ___ injure a human being or, through inaction, allow a human being to come to harm.



Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

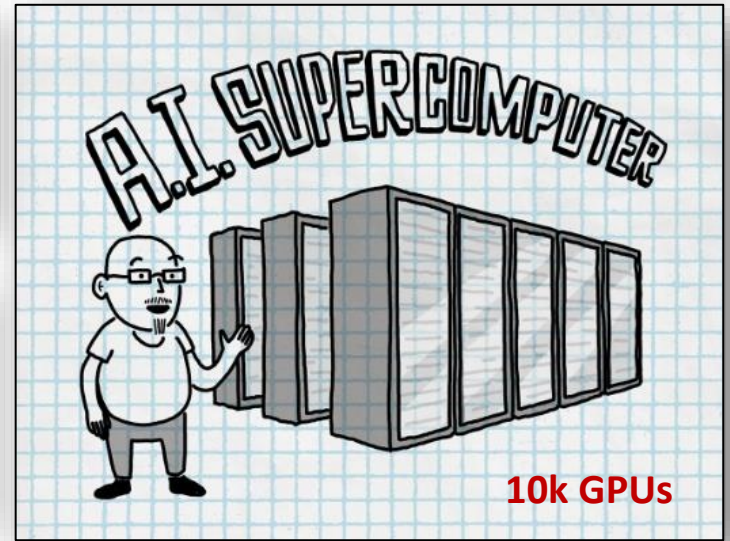
Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence

Building artificial general intelligence is OpenAI's ambitious goal

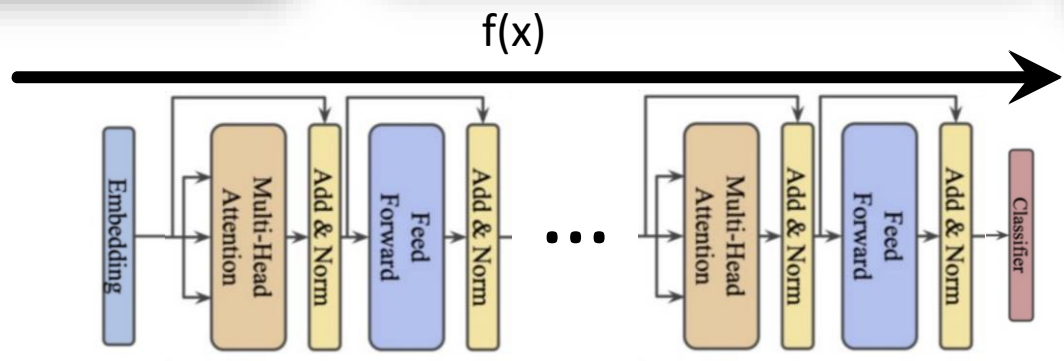
By James Vincent | Jul 22, 2019, 10:08am EDT

Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT @FredericLambert



A robot may ___ injure a human being or, through inaction, allow a human being to come to harm.



not	0.74
sometimes	0.28
always	0.07
never	0.04
and	0.33
boat	0.02
house	0.02

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

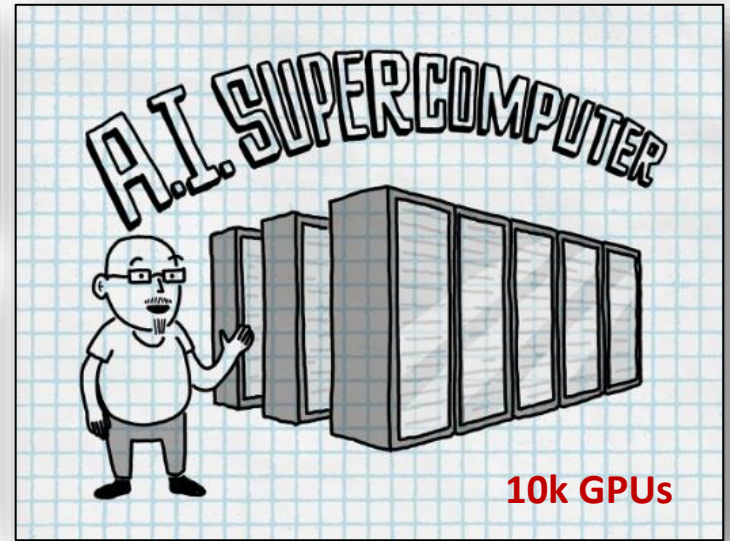
The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence

Building artificial general intelligence is OpenAI's ambitious goal

By James Vincent | Jul 22, 2019, 10:08am EDT

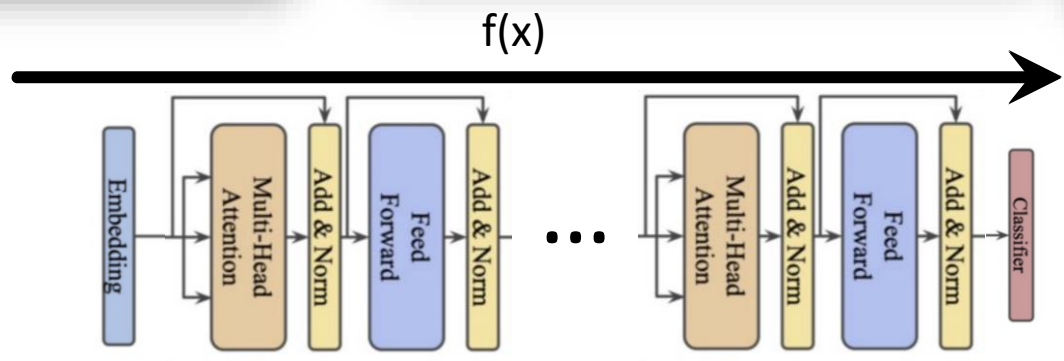


Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT @FredericLambert



A robot may ___ injure a human being or, through inaction, allow a human being to come to harm.



not	0.74	not	1.00
sometimes	0.28	sometimes	0.00
always	0.07	always	0.00
never	0.04	never	0.00
and	0.33	and	0.00
boat	0.02	boat	0.00
house	0.02	house	0.00

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

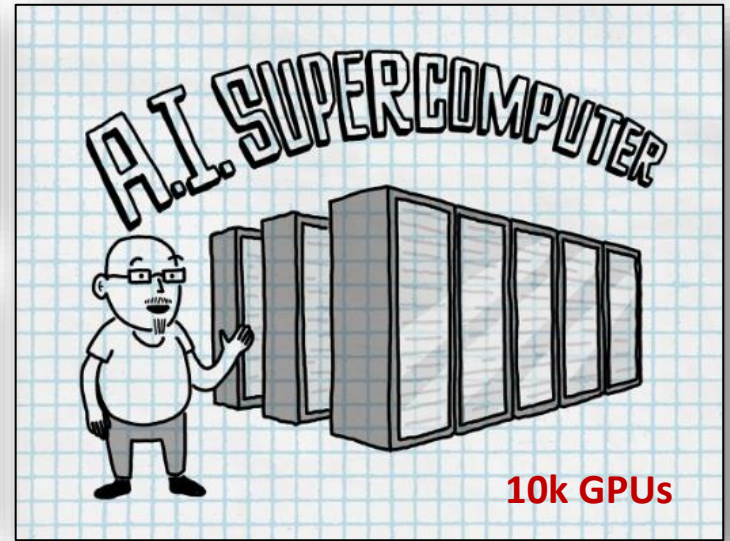
The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence

Building artificial general intelligence is OpenAI's ambitious goal

By James Vincent | Jul 22, 2019, 10:08am EDT

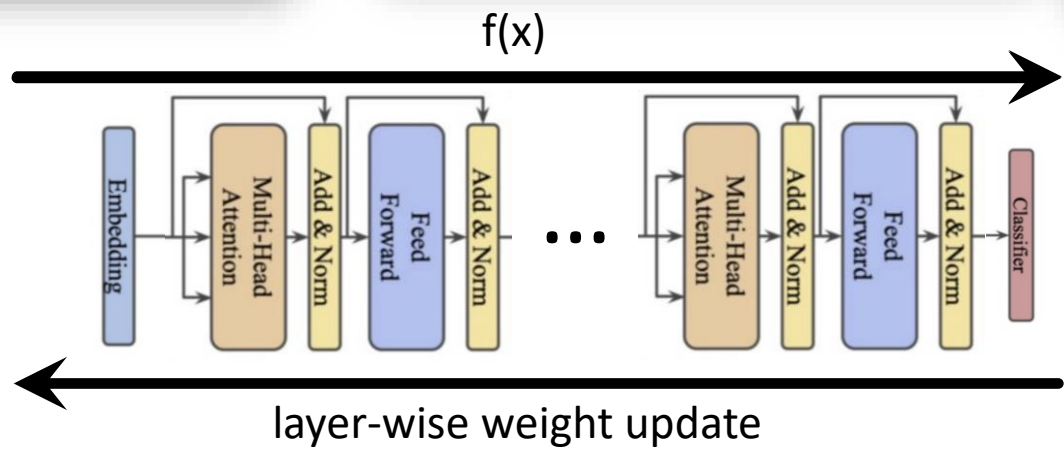


Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT @FredericLambert



A robot may ___ injure a human being or, through inaction, allow a human being to come to harm.



not	0.74	not	1.00
sometimes	0.28	sometimes	0.00
always	0.07	always	0.00
never	0.04	never	0.00
and	0.33	and	0.00
boat	0.02	boat	0.00
house	0.02	house	0.00

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

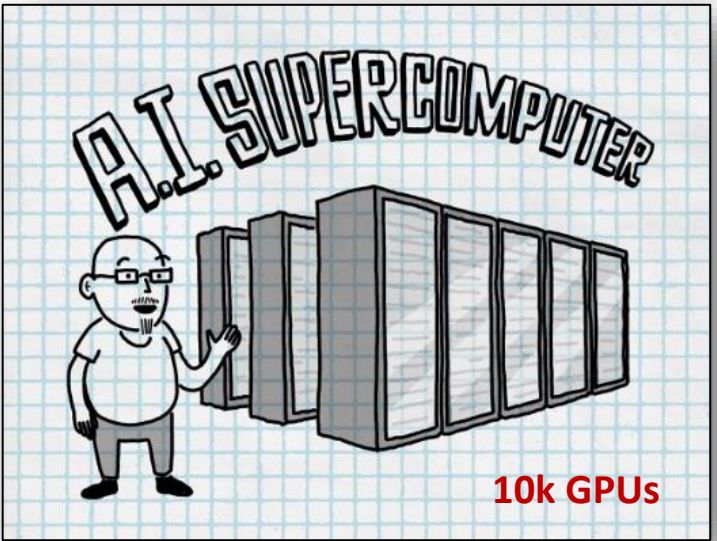
The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence

Building artificial general intelligence is OpenAI's ambitious goal

By James Vincent | Jul 22, 2019, 10:08am EDT

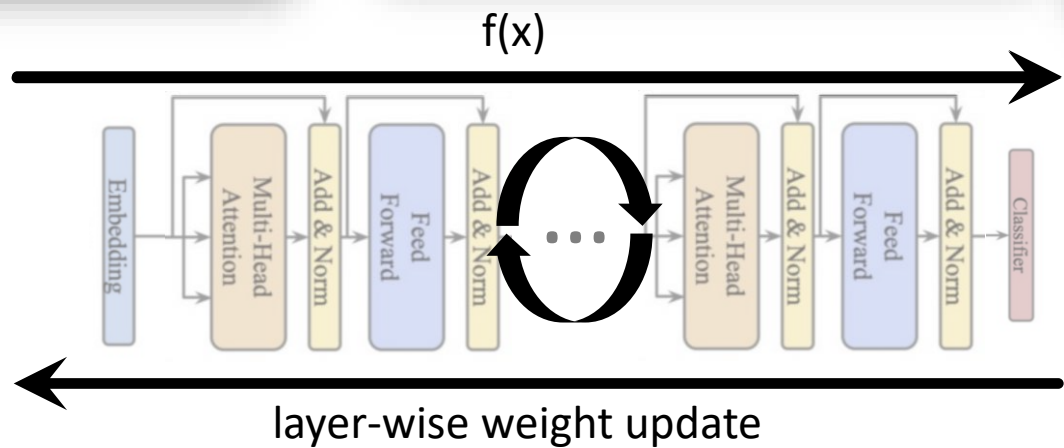


Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT @FredericLambert



A robot may ___ injure a human being or, through inaction, allow a human being to come to harm.



not	0.74	not	1.00
sometimes	0.28	sometimes	0.00
always	0.07	always	0.00
never	0.04	never	0.00
and	0.33	and	0.00
boat	0.02	boat	0.00
house	0.02	house	0.00

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

Share

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

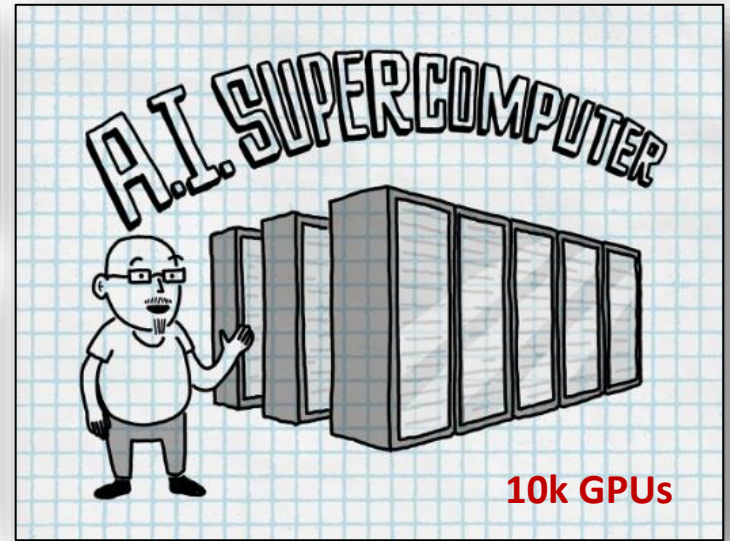
The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence

Building artificial general intelligence is OpenAI's ambitious goal

By James Vincent | Jul 22, 2019, 10:08am EDT

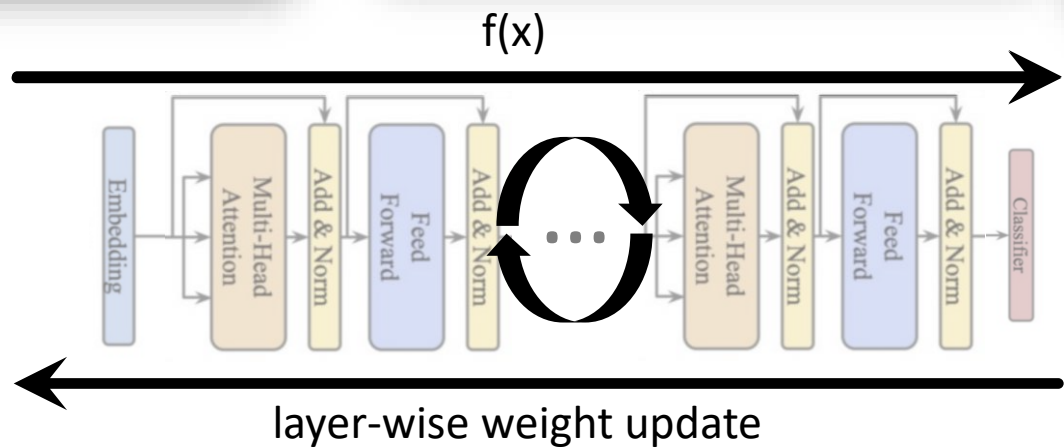


Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT @FredericLambert



A robot may __ injure a human being or, through inaction, allow a human being to come to harm.



not	0.74	not	1.00
sometimes	0.28	sometimes	0.00
always	0.07	always	0.00
never	0.04	never	0.00
and	0.33	and	0.00
boat	0.02	boat	0.00
house	0.02	house	0.00

- GPT-3: 500 billion tokens
- ImageNet (22k): A few TB
- Soon: **the whole internet!**

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer

Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today

By The Associated Press
January 24, 2022, 10:33 PM

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals

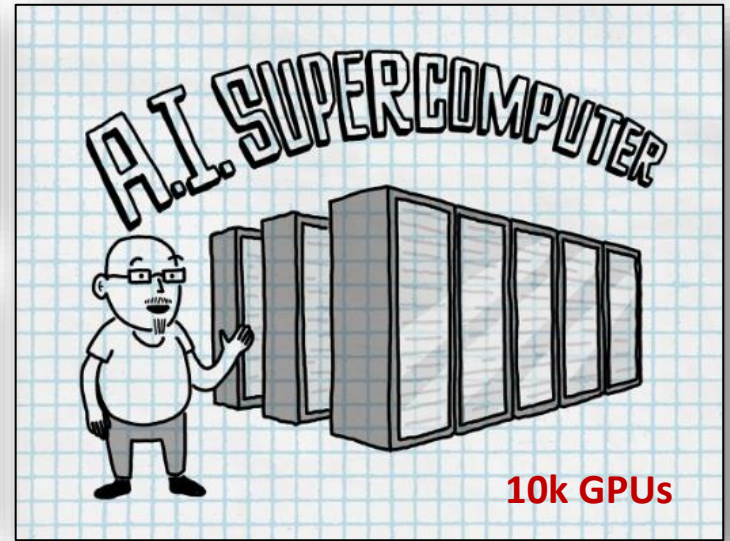
The NASNet system was created by a neural network called AutoML earlier this year

Mark Hodge
15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence

Building artificial general intelligence is OpenAI's ambitious goal

By James Vincent | Jul 22, 2019, 10:08am EDT

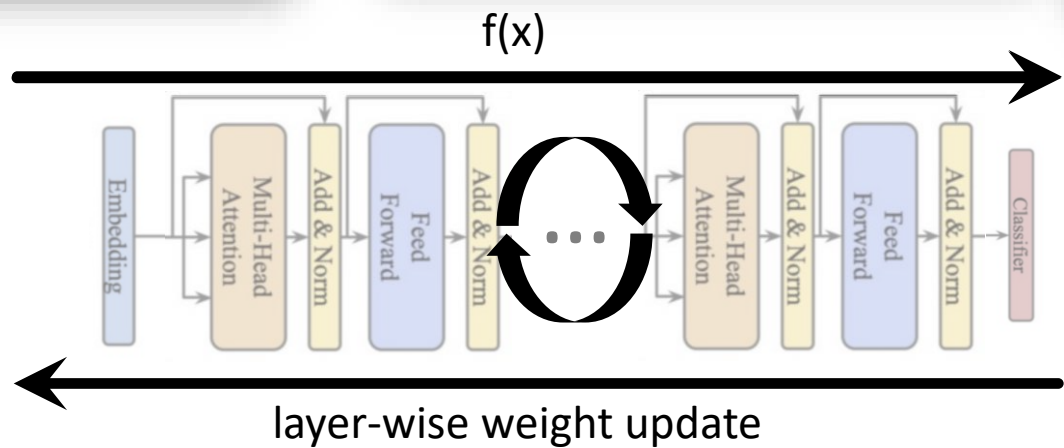


Tesla unveils Dojo supercomputer: world's new most powerful AI training machine

Fred Lambert · Aug. 20th 2021 3:08 am PT @FredericLambert



A robot may __ injure a human being or, through inaction, allow a human being to come to harm.



not	0.74	not	1.00
sometimes	0.28	sometimes	0.00
always	0.07	always	0.00
never	0.04	never	0.00
and	0.33	and	0.00
boat	0.02	boat	0.00
house	0.02	house	0.00

- GPT-3: 500 billion tokens
- ImageNet (22k): A few TB
- Soon: **the whole internet!**

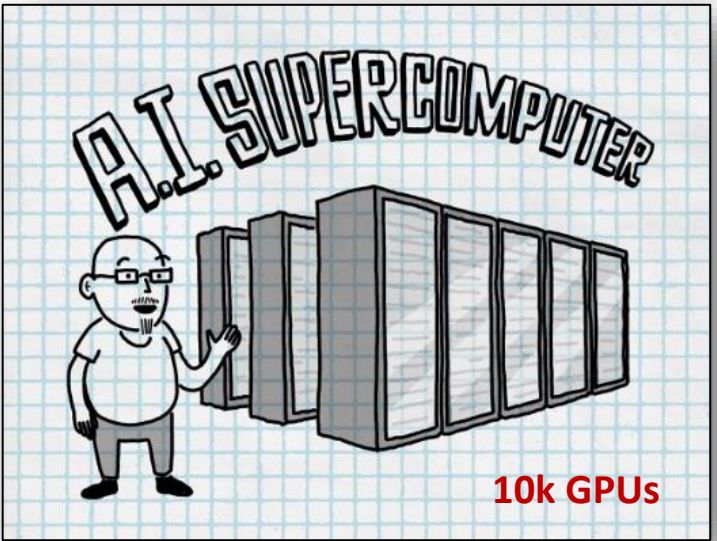
- GPT-3: 96 (complex) layers
175 bn parameters (700 GiB in fp32)
2048-token "sentences"

Supercomputers fuel Modern AI

Facebook parent Meta creates powerful AI supercomputer
 Facebook's parent company Meta says it has created what it believes is among the fastest artificial intelligence supercomputers running today
 By The Associated Press
 January 24, 2022, 10:33 PM

BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals
 The NASNet system was created by a neural network called AutoML earlier this year
 Mark Hodge
 15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

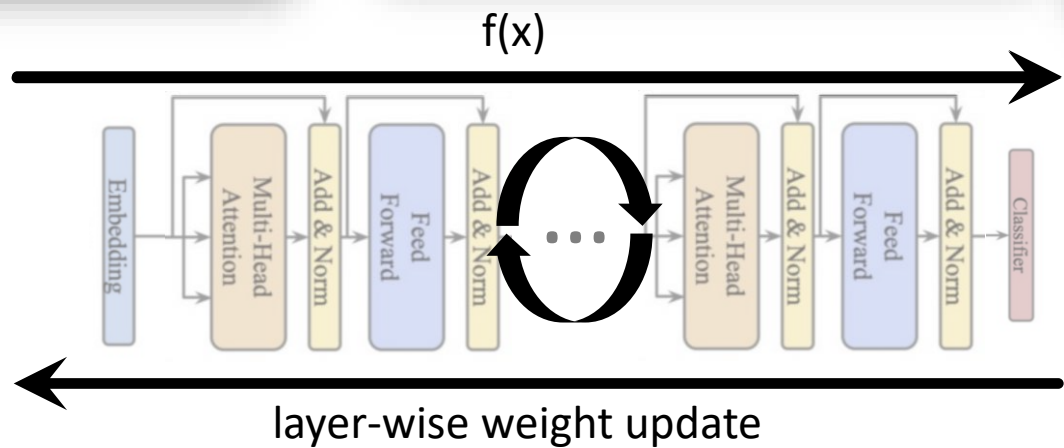
Microsoft invests \$1 billion in OpenAI to pursue holy grail of artificial intelligence
 Building artificial general intelligence is OpenAI's ambitious goal
 By James Vincent | Jul 22, 2019, 10:08am EDT



Tesla unveils Dojo supercomputer: world's new most powerful AI training machine
 Fred Lambert · Aug. 20th 2021 3:08 am PT @FredericLambert



A robot may __ injure a human being or, through inaction, allow a human being to come to harm.



not	0.74	not	1.00
sometimes	0.28	sometimes	0.00
always	0.07	always	0.00
never	0.04	never	0.00
and	0.33	and	0.00
boat	0.02	boat	0.00
house	0.02	house	0.00

- GPT-3: 500 billion tokens
- ImageNet (22k): A few TB
- Soon: **the whole internet!**

- GPT-3: 96 (complex) layers
 175 bn parameters (**700 GiB** in fp32)
 2048-token "sentences"

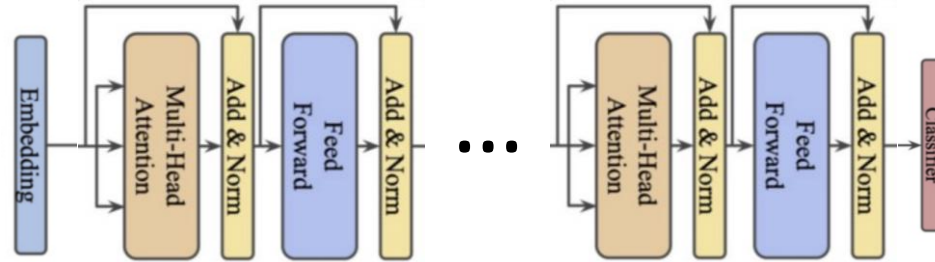
- GPT-3: 30-50k dictionaries
- **takes weeks to train**

Large-Scale AI is the Future

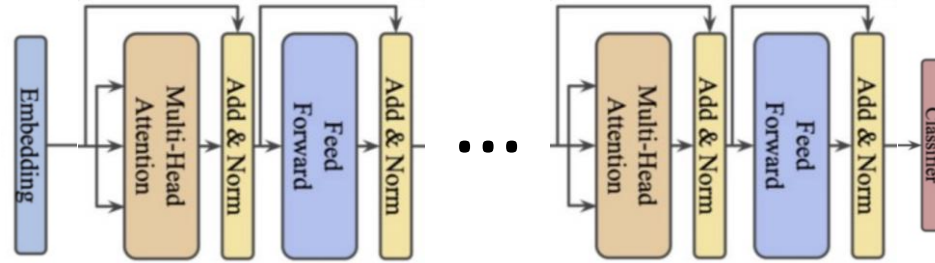
Large-Scale AI is the Future

We need a Principled Approach to it

Three Systems Dimensions in Large-scale Super-learning ...



Three Systems Dimensions in Large-scale Super-learning ...



High-Performance I/O

- Quickly growing data volumes
 - Scientific computing!
- Use the specifics of machine learning workloads
 - E.g., intelligent prefetching

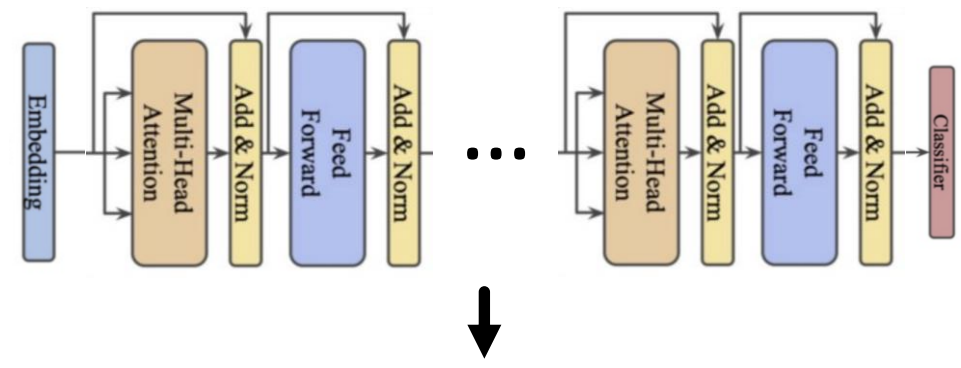
CLAIRVOYANT PREFETCHING FOR DISTRIBUTED MACHINE LEARNING I/O

Roman Böhringer¹ Nikoli Dryden¹ Tal Ben-Nun¹ Torsten Hoefler¹

ABSTRACT

I/O is emerging as a major bottleneck for machine learning training, especially in distributed environments such as clouds and supercomputers. Optimal data ingestion pipelines differ between systems, and increasing efficiency requires a delicate balance between access to local storage, external filesystems, and remote workers; yet existing frameworks fail to efficiently utilize such resources. We observe that, given the seed generating the random access pattern for training with SGD, we have *clairvoyance* and can exactly predict when a given sample will be accessed. We combine this with a theoretical analysis of access patterns in training and performance modeling to produce a novel machine learning I/O middleware, HDMLP, to tackle the I/O bottleneck. HDMLP provides an easy-to-use, flexible, and scalable solution that delivers better performance than state-of-the-art approaches while requiring very few changes to existing codebases and supporting a broad range of environments.

Three Systems Dimensions in Large-scale Super-learning ...



High-Performance I/O

- Quickly growing data volumes
 - Scientific computing!
- Use the specifics of machine learning workloads
 - E.g., intelligent prefetching

High-Performance Compute

- Deep learning is HPC
 - Data movement!
- Quantization, Sparsification
 - Drives modern accelerators!

CLAIRVOYANT PREFETCHING FOR DISTRIBUTED MACHINE LEARNING I/O

Roman Böhringer¹ Nikoli Dryden¹ Tal Ben-Nun¹ Torsten Hoefler¹

ABSTRACT

I/O is emerging as a major bottleneck for machine learning training, especially in distributed environments such as clouds and supercomputers. Optimal data ingestion pipelines differ between systems, and increasing efficiency requires a delicate balance between access to local storage, external filesystems, and remote workers; yet existing frameworks fail to efficiently utilize such resources. We observe that, given the seed generating the random access pattern for training with SGD, we have *clairvoyance* and can exactly predict when a given sample will be accessed. We combine this with a theoretical analysis of access patterns in training and performance modeling to produce a novel machine learning I/O middleware, HDMLP, to tackle the I/O bottleneck. HDMLP provides an easy-to-use, flexible, and scalable solution that delivers better performance than state-of-the-art approaches while requiring very few changes to existing codebases and supporting a broad range of environments.

21 Jan 2021

Data Movement Is All You Need: A Case Study on Optimizing Transformers

Andrei Ivanov*, Nikoli Dryden*, Tal Ben-Nun, Shigang Li, Torsten Hoefler
ETH Zürich
firstname.lastname@inf.ethz.ch
* Equal contribution

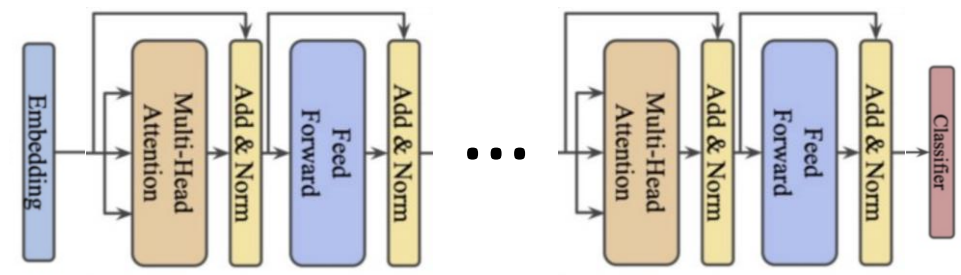
Abstract—Transformers have become widely used for language modeling and sequence learning tasks, and are one of the most important machine learning workloads today. Training one is a very compute-intensive task, often taking days or weeks, and significant attention has been given to optimizing transformers. Despite this, existing implementations do not efficiently utilize GPUs. We find that data movement is the key bottleneck when training. Due to Amdahl's Law and massive improvements in compute performance, training has now become memory-bound. Further, existing frameworks use suboptimal data layouts. Using these insights, we present a recipe for globally optimizing data movement in transformers. We reduce data movement by up to 22.91% and overall achieve a 1.30x performance improvement over state-of-the-art frameworks when training BERT. Our approach is applicable more broadly to optimizing deep neural networks, and offers insight into how to tackle emerging performance bottlenecks.

challenges such as artificial general intelligence [27]. Thus, improving transformer performance has been in the focus of numerous research and industrial groups.

Significant attention has been given to optimizing transformers: local and fixed-window attention [28]–[32], more general structured sparsity [33], learned sparsity [34]–[36], and other algorithmic techniques [19], [37] improve the performance of transformers. Major hardware efforts, such as Tensor Cores and TPUs [38] have accelerated tensor operations like matrix-matrix multiplication (MMM), a core transformer operation. Despite this, existing implementations do not efficiently utilize GPUs. Even optimized implementations such as Megatron [18] report achieving only 30% of peak GPU flops. We find that the key bottleneck when training transform-

s.LGJ 2 Jul 2020

Three Systems Dimensions in Large-scale Super-learning ...



High-Performance I/O

- Quickly growing data volumes
 - Scientific computing!
- Use the specifics of machine learning workloads
 - E.g., intelligent prefetching

21 Jan 2021

CLAIRVOYANT PREFETCHING FOR DISTRIBUTED MACHINE LEARNING I/O

Roman Böhringer¹ Nikoli Dryden¹ Tal Ben-Nun¹ Torsten Hoefler¹

ABSTRACT

I/O is emerging as a major bottleneck for machine learning training, especially in distributed environments such as clouds and supercomputers. Optimal data ingestion pipelines differ between systems, and increasing efficiency requires a delicate balance between access to local storage, external filesystems, and remote workers; yet existing frameworks fail to efficiently utilize such resources. We observe that, given the seed generating the random access pattern for training with SGD, we have *clairvoyance* and can exactly predict when a given sample will be accessed. We combine this with a theoretical analysis of access patterns in training and performance modeling to produce a novel machine learning I/O middleware, HDMLP, to tackle the I/O bottleneck. HDMLP provides an easy-to-use, flexible, and scalable solution that delivers better performance than state-of-the-art approaches while requiring very few changes to existing codebases and supporting a broad range of environments.

High-Performance Compute

- Deep learning is HPC
 - **Data movement!**
- **Quantization, Sparsification**
 - Drives modern accelerators!

Data Movement Is All You Need: A Case Study on Optimizing Transformers

Andrei Ivanov*, Nikoli Dryden*, Tal Ben-Nun, Shigang Li, Torsten Hoefler
ETH Zürich
firstname.lastname@inf.ethz.ch
* Equal contribution

Abstract—Transformers have become widely used for language modeling and sequence learning tasks, and are one of the most important machine learning workloads today. Training one is a very compute-intensive task, often taking days or weeks, and significant attention has been given to optimizing transformers. Despite this, existing implementations do not efficiently utilize GPUs. We find that data movement is the key bottleneck when training. Due to Amdahl's Law and massive improvements in compute performance, training has now become memory-bound. Further, existing frameworks use suboptimal data layouts. Using these insights, we present a recipe for globally optimizing data movement in transformers. We reduce data movement by up to 22.91% and overall achieve a 1.30x performance improvement over state-of-the-art frameworks when training BERT. Our approach is applicable more broadly to optimizing deep neural networks, and offers insight into how to tackle emerging performance bottlenecks.

challenges such as artificial general intelligence [27]. Thus, improving transformer performance has been in the focus of numerous research and industrial groups.

Significant attention has been given to optimizing transformers: local and fixed-window attention [28]–[32], more general structured sparsity [33], learned sparsity [34]–[36], and other algorithmic techniques [19], [37] improve the performance of transformers. Major hardware efforts, such as Tensor Cores and TPUs [38] have accelerated tensor operations like matrix-matrix multiplication (MMM), a core transformer operation. Despite this, existing implementations do not efficiently utilize GPUs. Even optimized implementations such as Megatron [18] report achieving only 30% of peak GPU flops.

We find that the key bottleneck when training transform-

High-Performance Communication

- Use larger clusters (10k+ GPUs)
- Model parallelism
 - Complex pipeline schemes
- Optimized networks

Distribution and Parallelism

Data	Pipeline	Operator
<p>SPCL: High-Performance Sparse Communication for Machine Learning</p> <p>19 Dec 2020</p> <p>Abstract: Sparse communication is a key challenge in training large-scale models on distributed systems. We present SPCL, a high-performance sparse communication framework for machine learning. SPCL achieves up to 1.5x performance improvement over state-of-the-art frameworks in training large-scale models on distributed systems.</p>	<p>Chimera: Efficiently Training Large-Scale Neural Networks with Bidirectional Pipelines</p> <p>19 Dec 2020</p> <p>Abstract: Training large-scale neural networks on distributed systems is a challenging task due to the high communication overhead. We present Chimera, a framework for training large-scale neural networks on distributed systems. Chimera achieves up to 1.5x performance improvement over state-of-the-art frameworks in training large-scale neural networks on distributed systems.</p>	<p>Red Blue Publishing Received: Near Optimal Parallel Matrix-Matrix Multiplication</p> <p>19 Dec 2020</p> <p>Abstract: Matrix-matrix multiplication (MMM) is a core operation in many machine learning applications. We present Red Blue Publishing, a near-optimal parallel MMM algorithm. Red Blue Publishing achieves up to 1.5x performance improvement over state-of-the-art frameworks in training large-scale models on distributed systems.</p>
<p>Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis</p> <p>19 Dec 2020</p> <p>Abstract: Deep learning has become a dominant paradigm in artificial intelligence. However, training deep learning models on distributed systems is a challenging task due to the high communication overhead. We present an in-depth concurrency analysis of deep learning training on distributed systems. Our analysis reveals that data movement is the key bottleneck in training deep learning models on distributed systems.</p>	<p>Performance Analysis of 1,000+ GPUs and 100,000+ Nodes: Profiling and Tuning Distributed Machine Learning</p> <p>19 Dec 2020</p> <p>Abstract: Training large-scale machine learning models on distributed systems is a challenging task due to the high communication overhead. We present a performance analysis of 1,000+ GPUs and 100,000+ nodes. Our analysis reveals that data movement is the key bottleneck in training large-scale machine learning models on distributed systems.</p>	<p>Red Blue Publishing Received: Near Optimal Parallel Matrix-Matrix Multiplication</p> <p>19 Dec 2020</p> <p>Abstract: Matrix-matrix multiplication (MMM) is a core operation in many machine learning applications. We present Red Blue Publishing, a near-optimal parallel MMM algorithm. Red Blue Publishing achieves up to 1.5x performance improvement over state-of-the-art frameworks in training large-scale models on distributed systems.</p>

This is the past – LLMs work today!

(you can find that full talk on youtube)



This is the past – LLMs work today!
(you can find that full talk on youtube)



We know training – but do we understand using (prompting)?

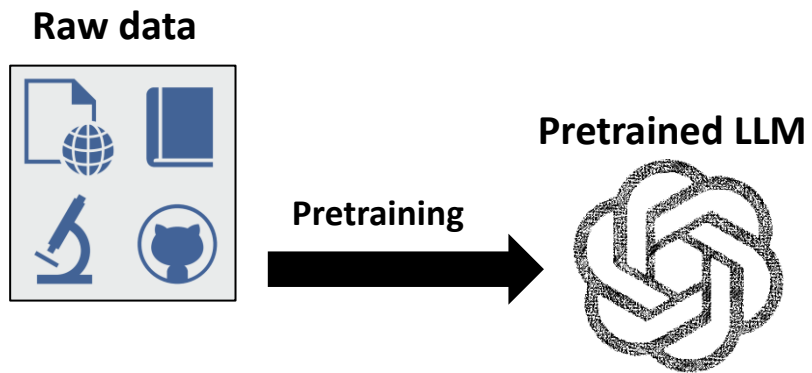
Overview of the LLM Processing Pipeline

Overview of the LLM Processing Pipeline

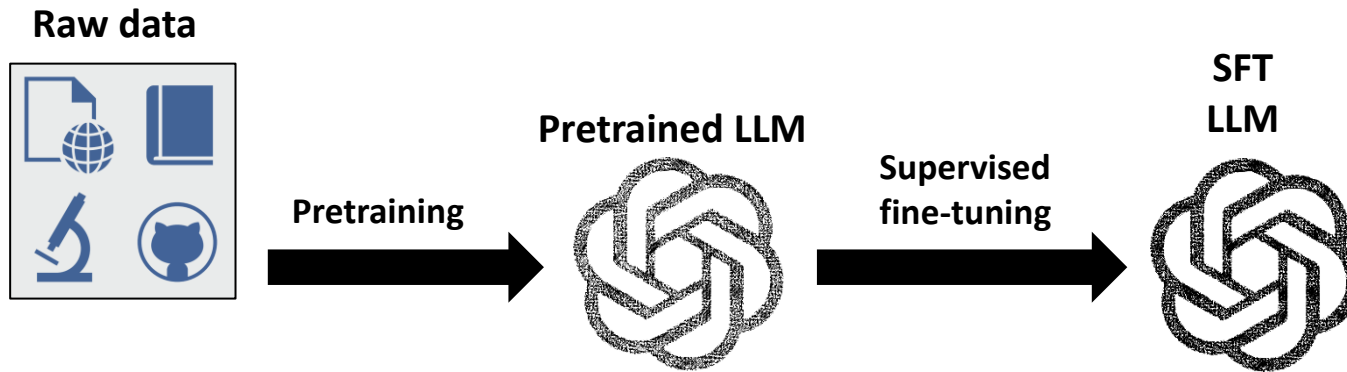
Raw data



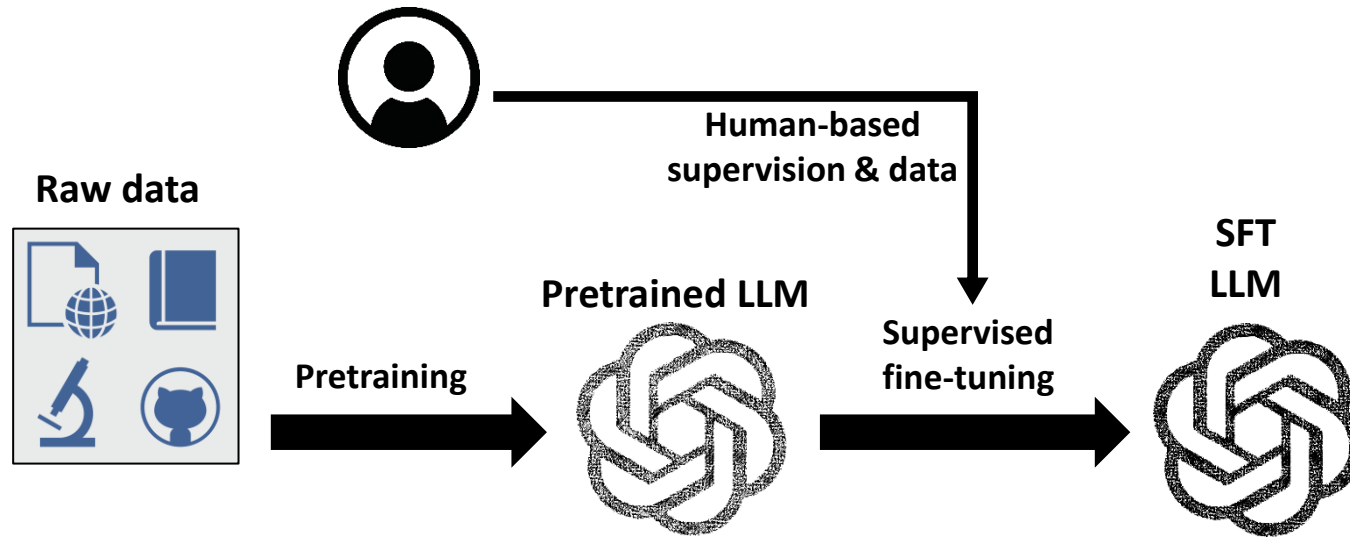
Overview of the LLM Processing Pipeline



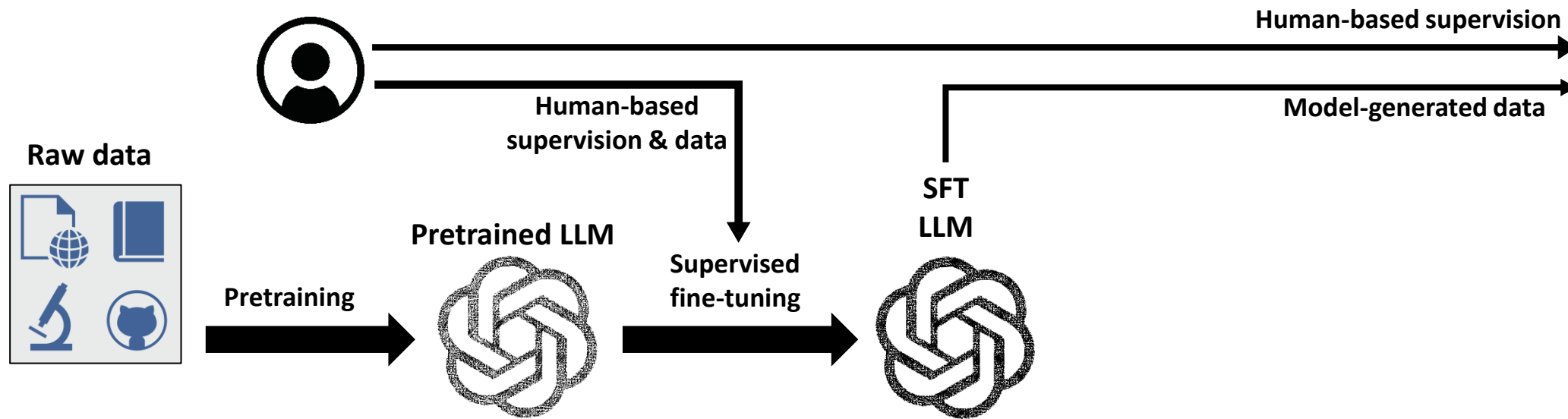
Overview of the LLM Processing Pipeline



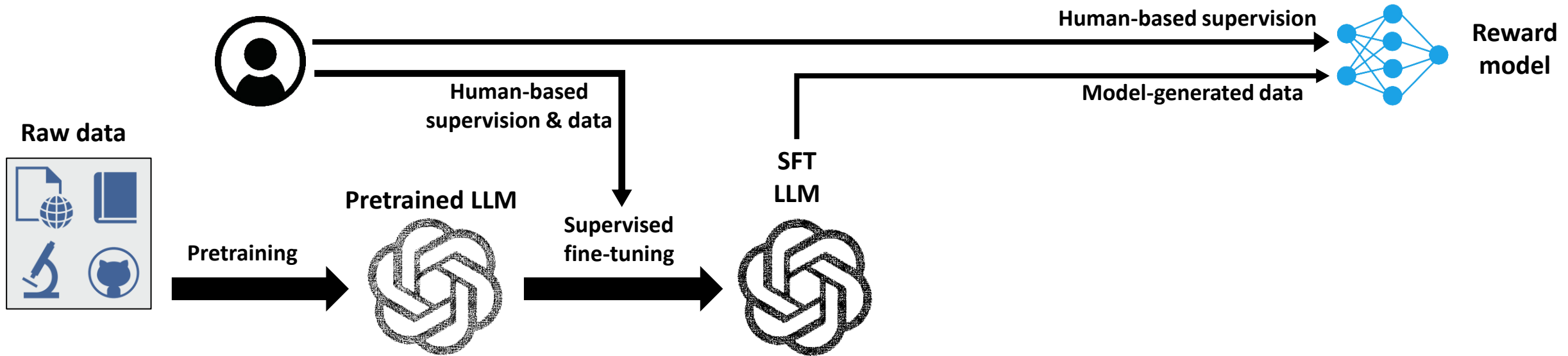
Overview of the LLM Processing Pipeline



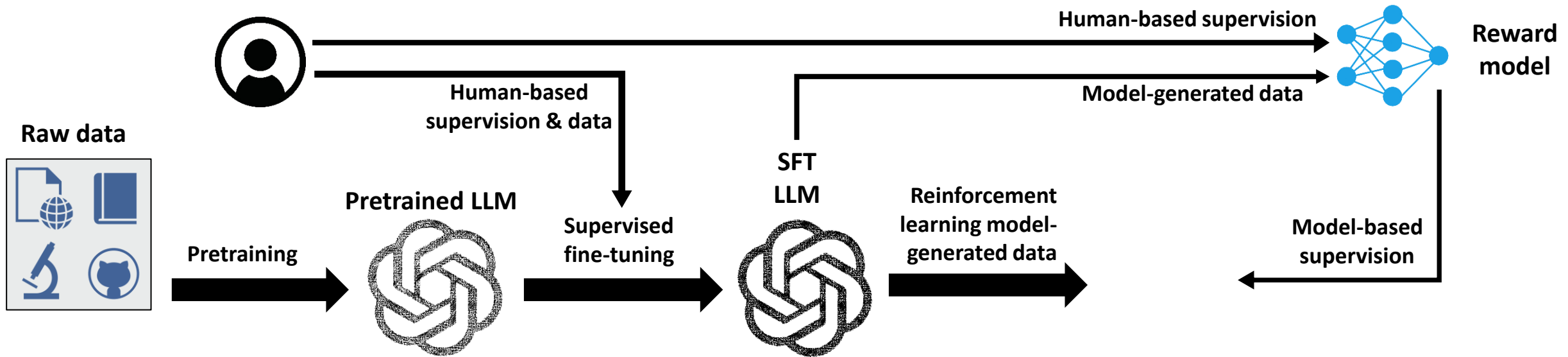
Overview of the LLM Processing Pipeline



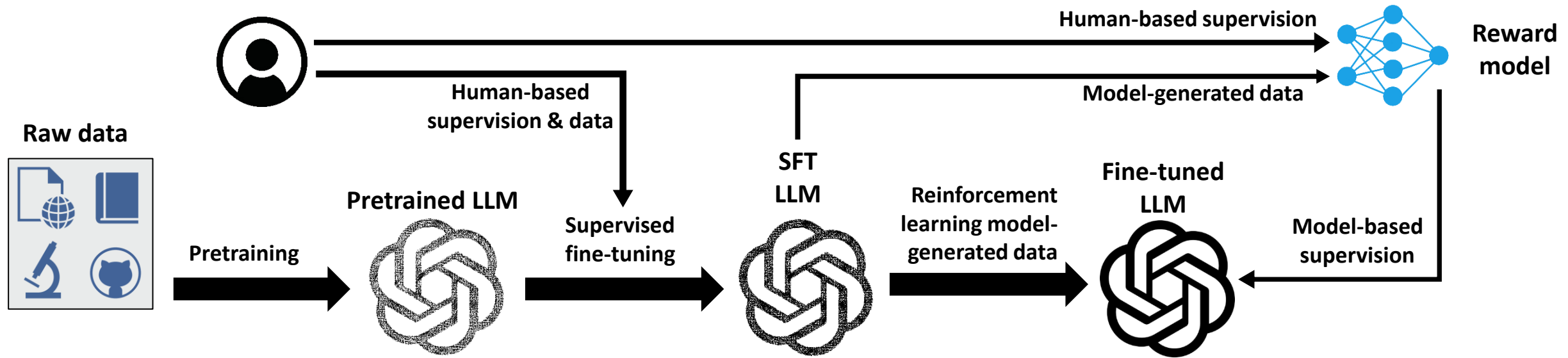
Overview of the LLM Processing Pipeline



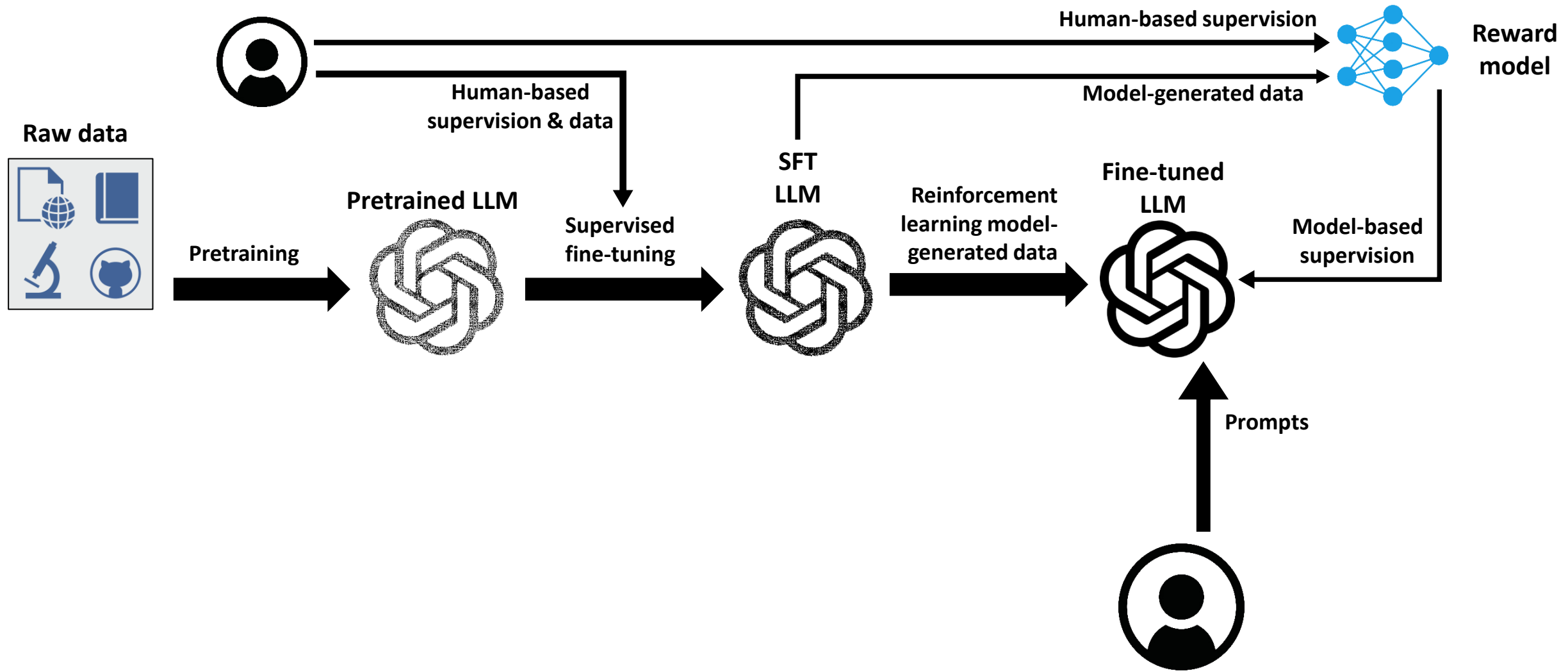
Overview of the LLM Processing Pipeline



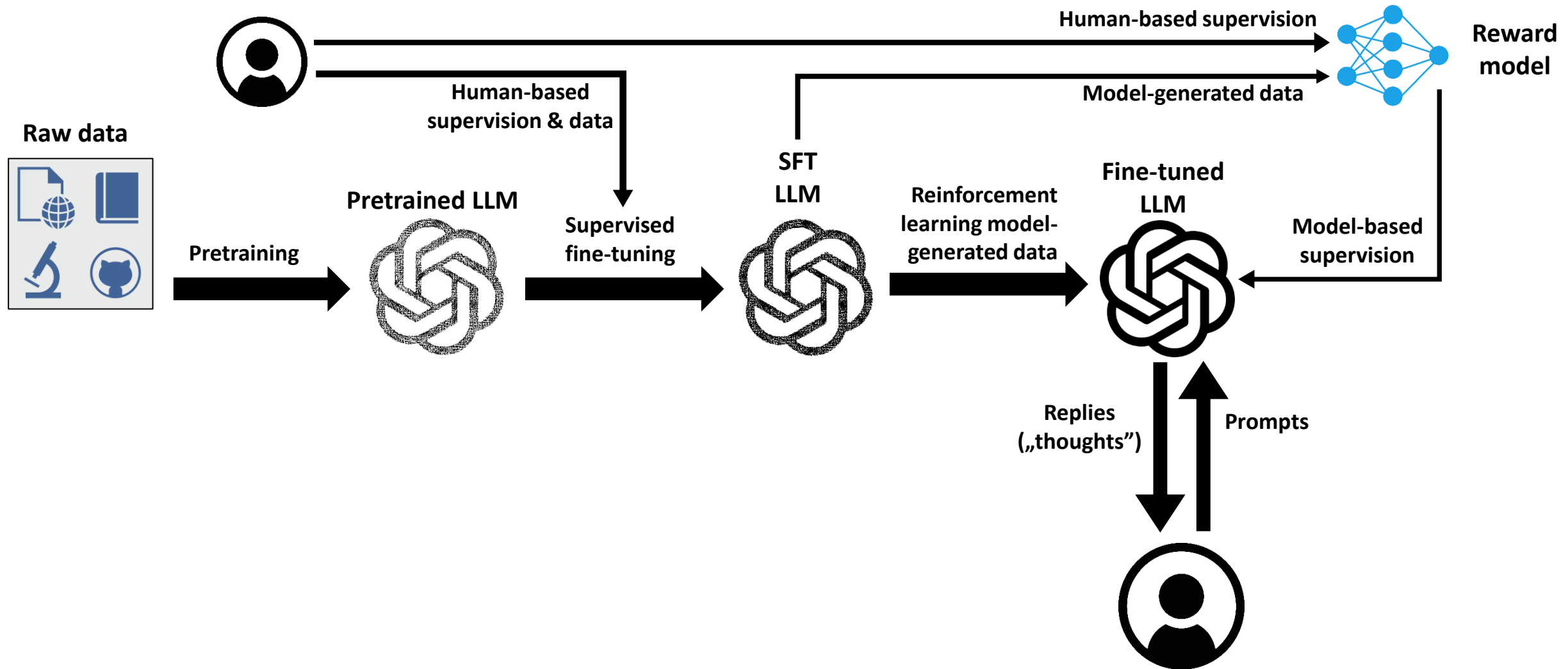
Overview of the LLM Processing Pipeline



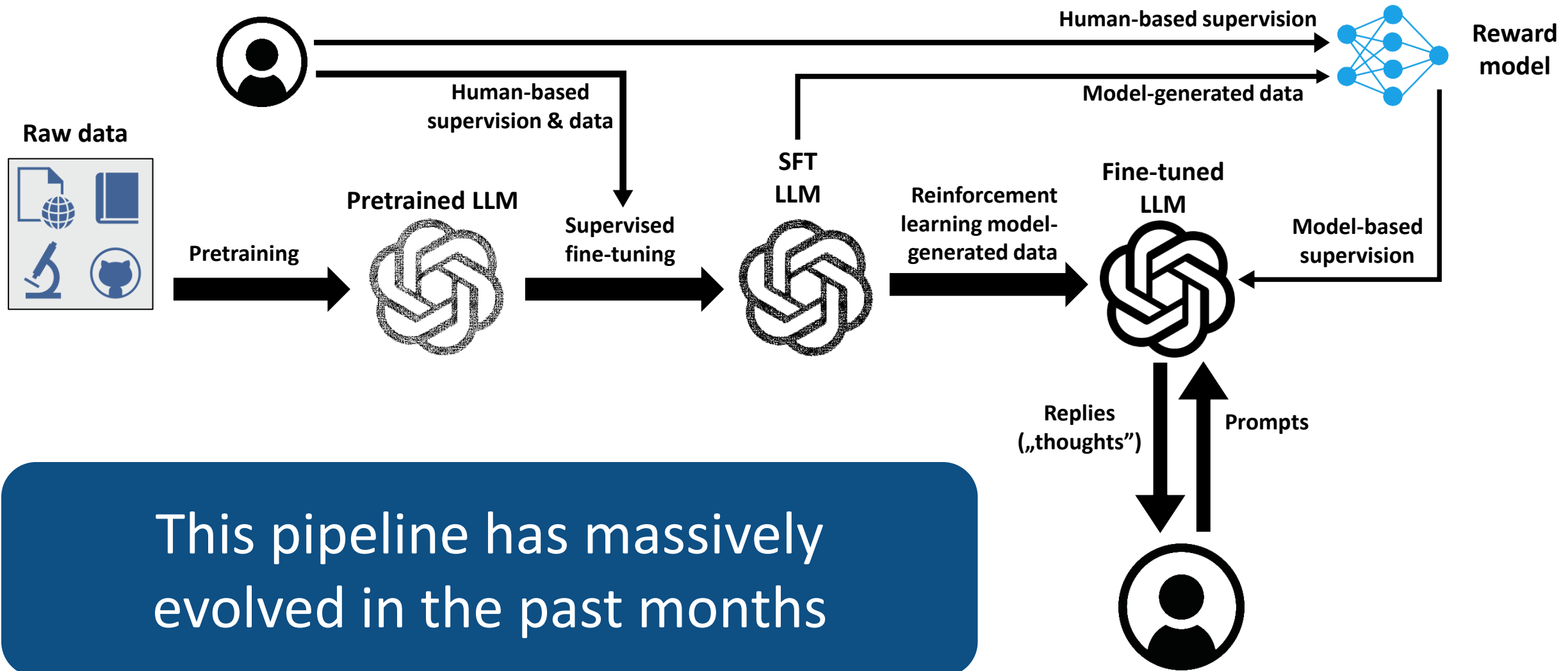
Overview of the LLM Processing Pipeline



Overview of the LLM Processing Pipeline

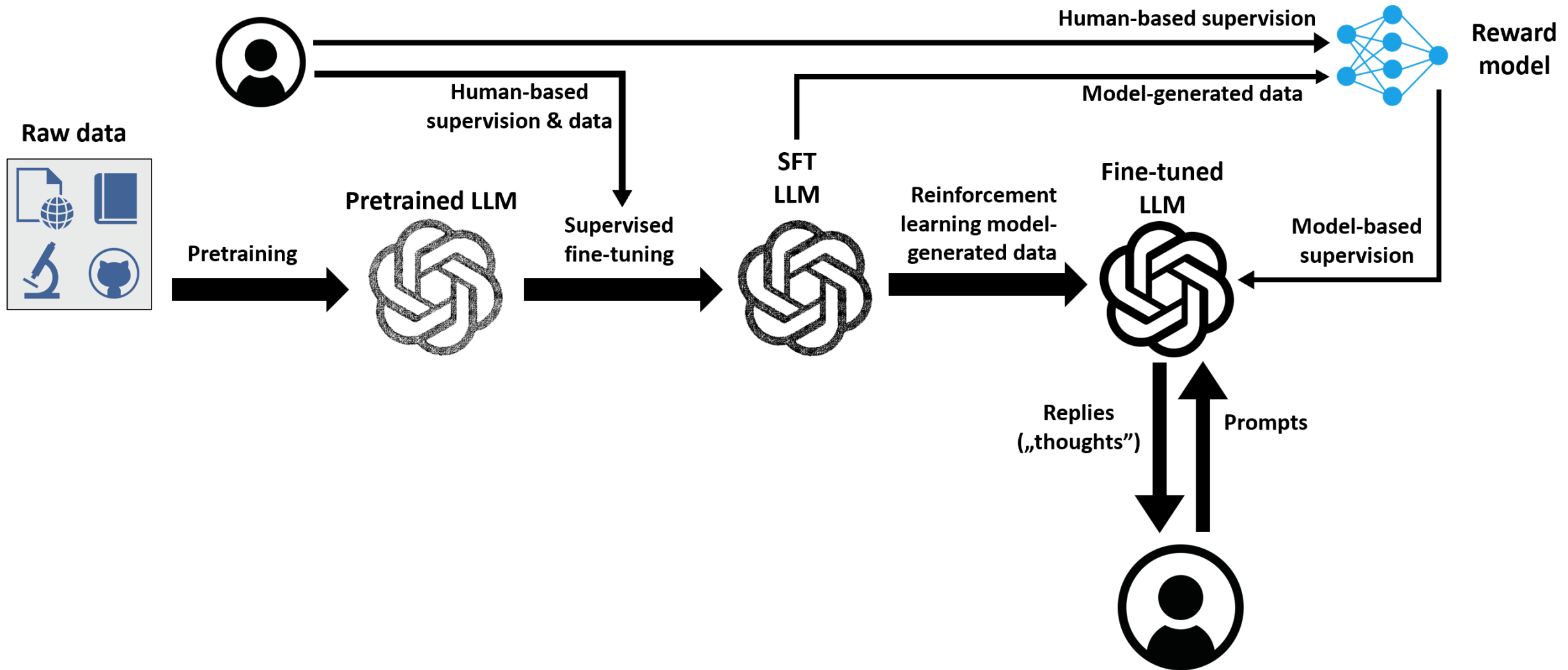


Overview of the LLM Processing Pipeline

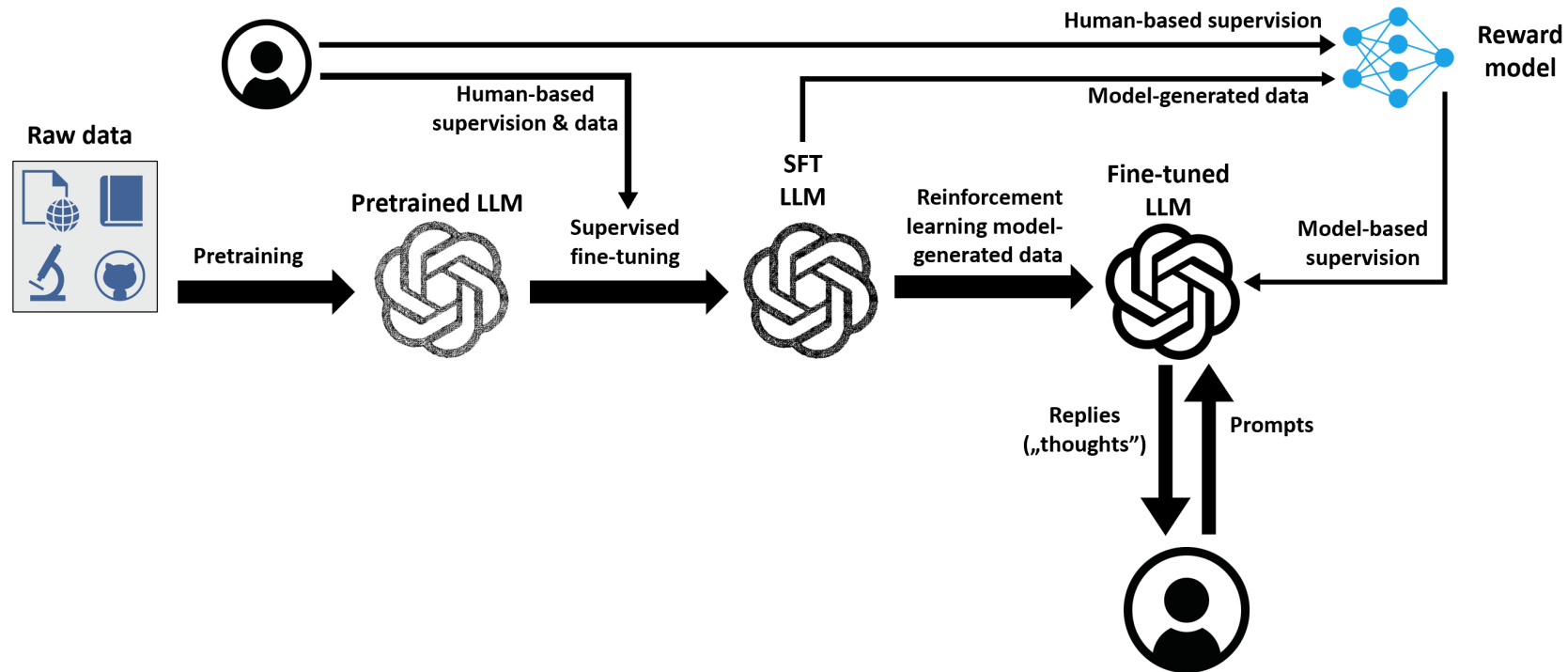


This pipeline has massively evolved in the past months

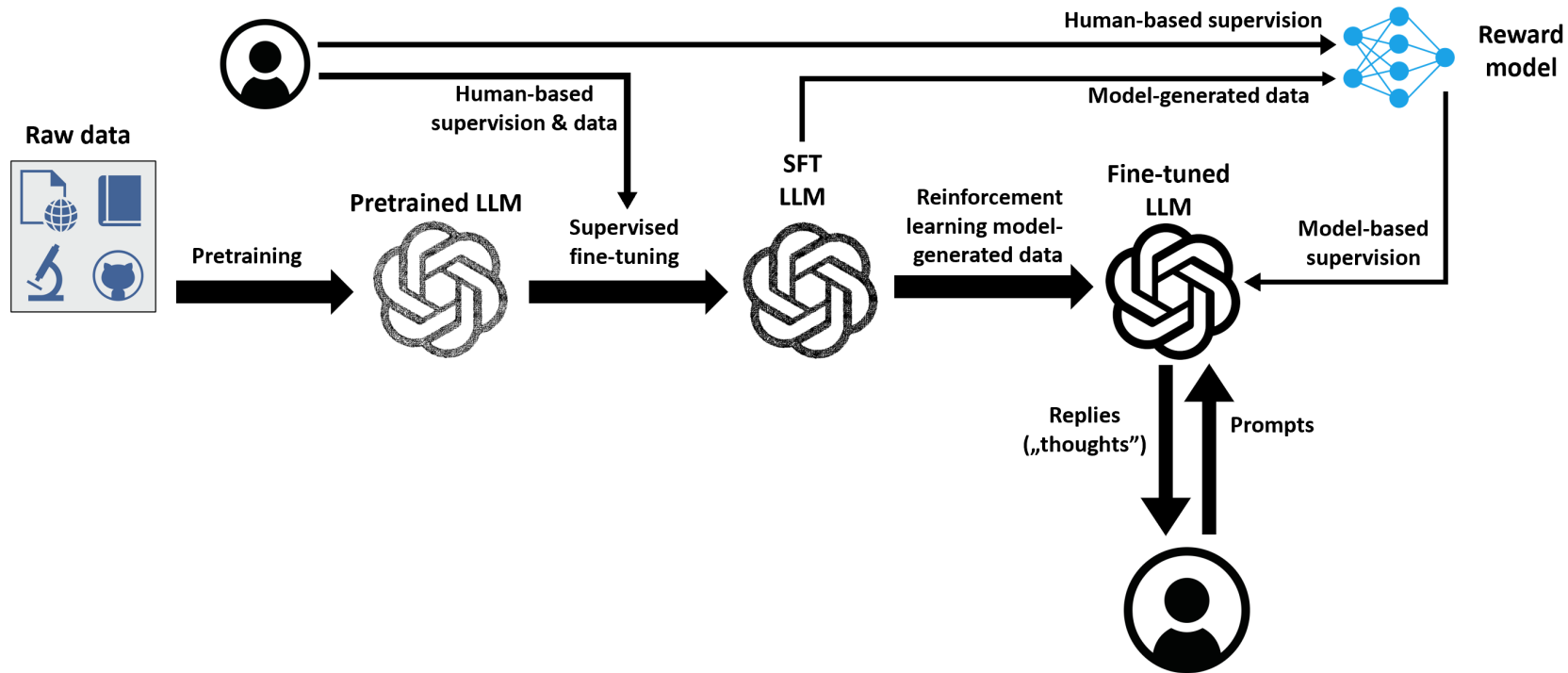
Overview of the LLM Processing Pipeline



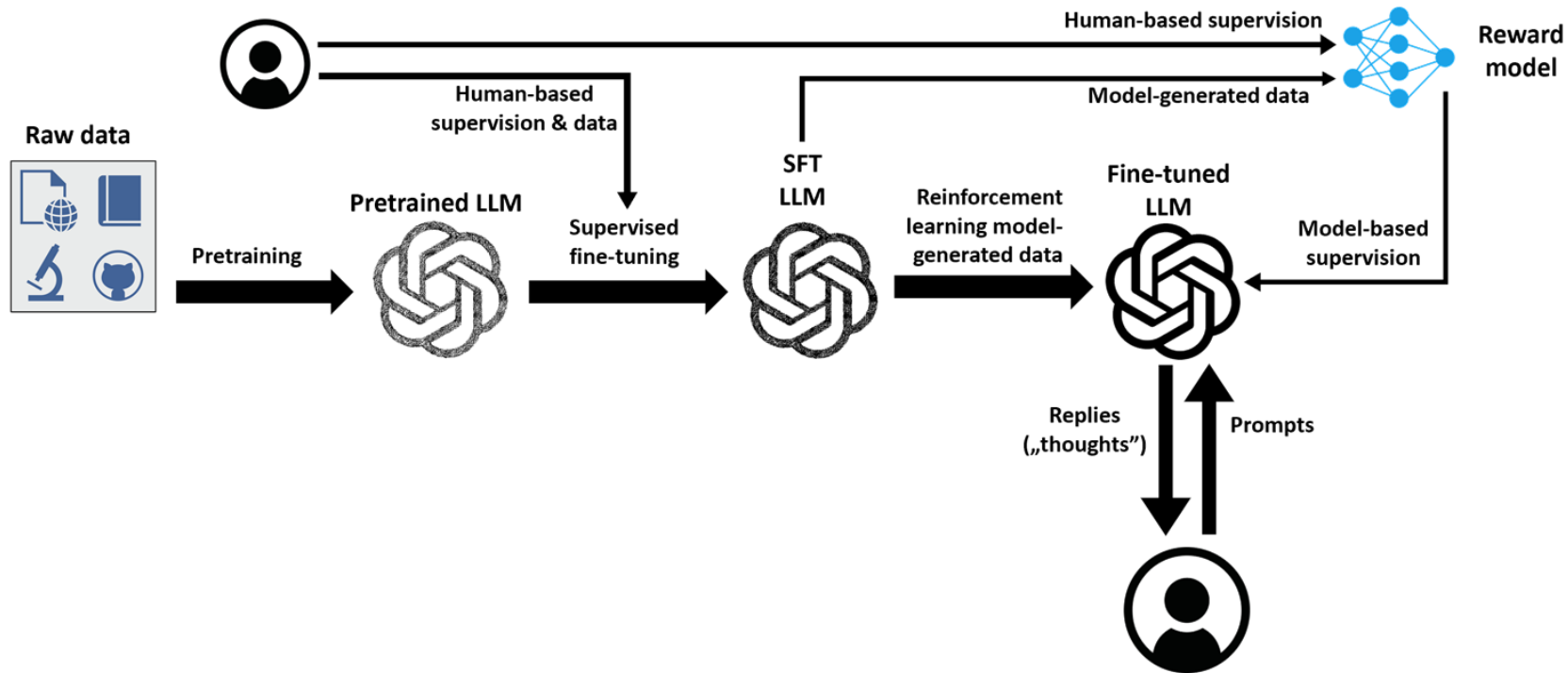
Overview of the LLM Processing Pipeline



The Emergence of the „Generative AI Ecosystem”

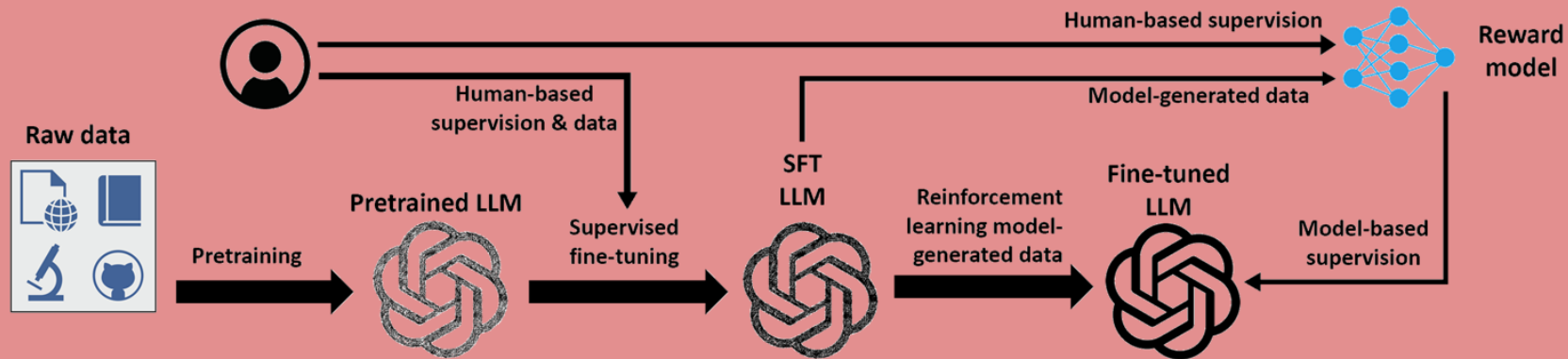


The Emergence of the „Generative AI Ecosystem”

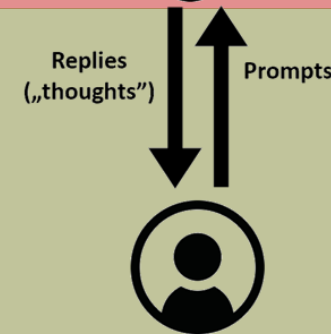


The Emergence of the „Generative AI Ecosystem”

Training related

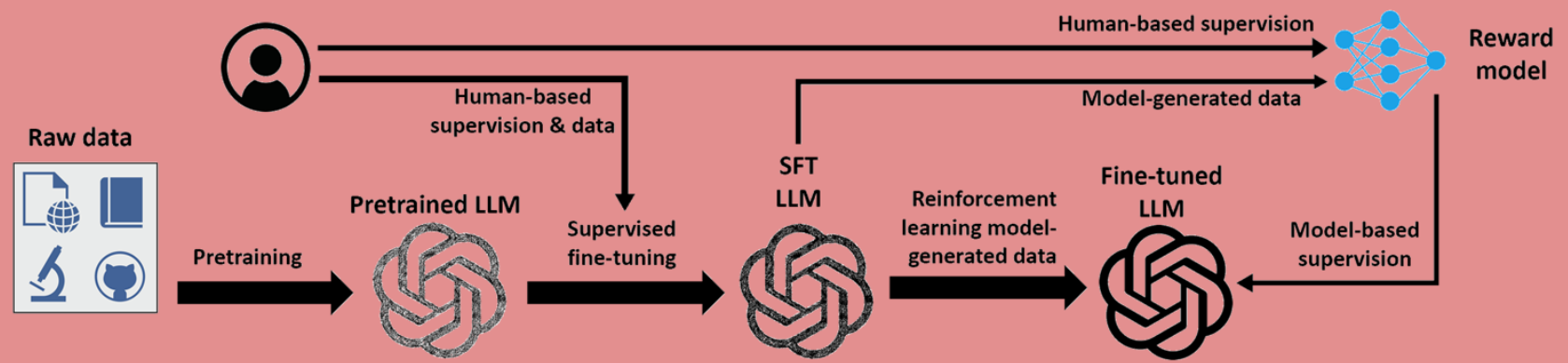


Inference related

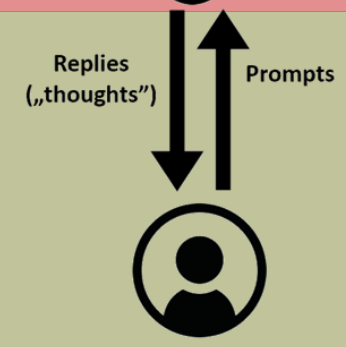


The Emergence of the „Generative AI Ecosystem”

Training related

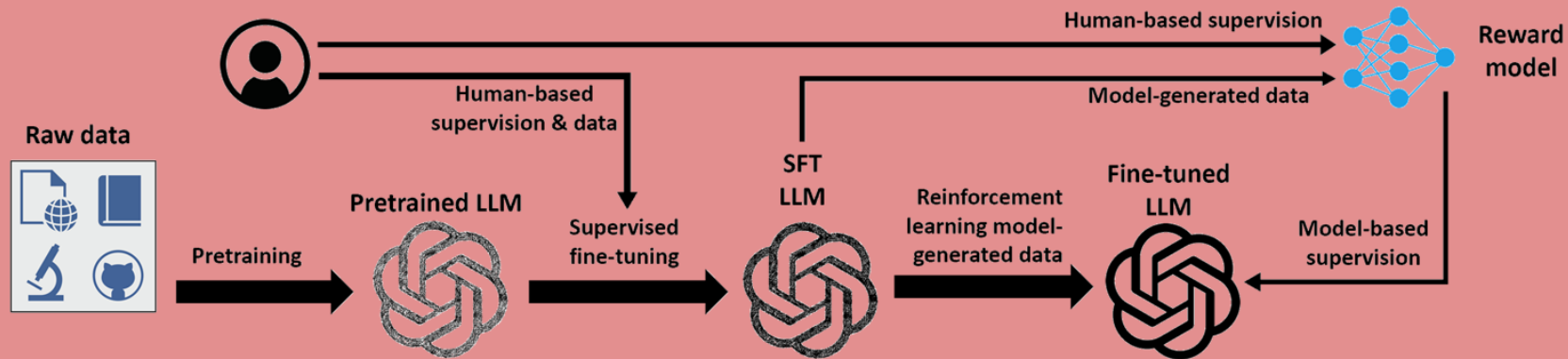


Inference related

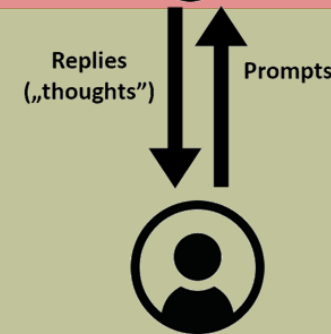
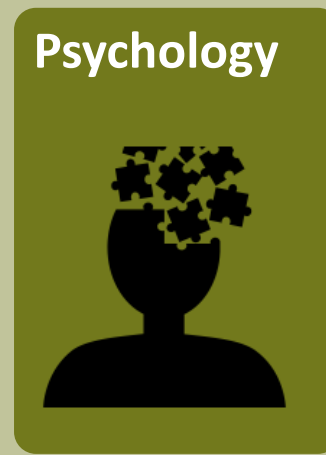


The Emergence of the „Generative AI Ecosystem”

Training related

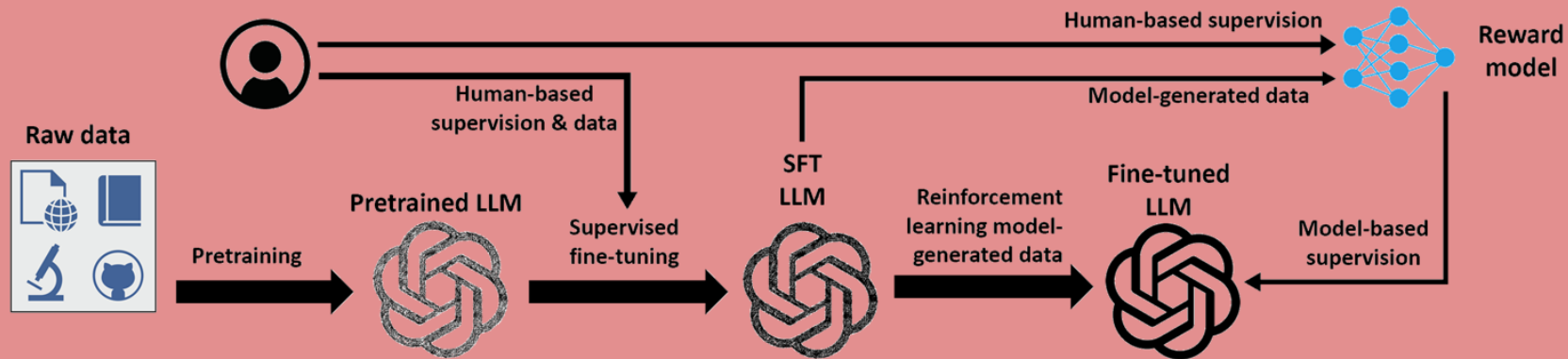


Inference related

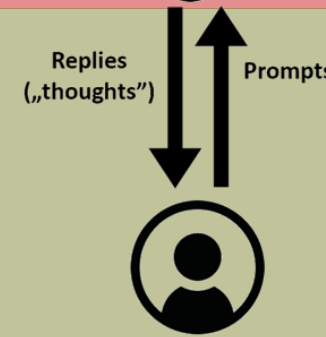
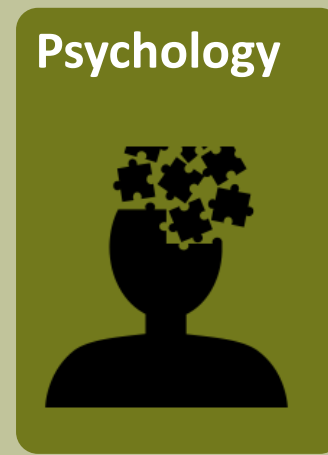


The Emergence of the „Generative AI Ecosystem”

Training related

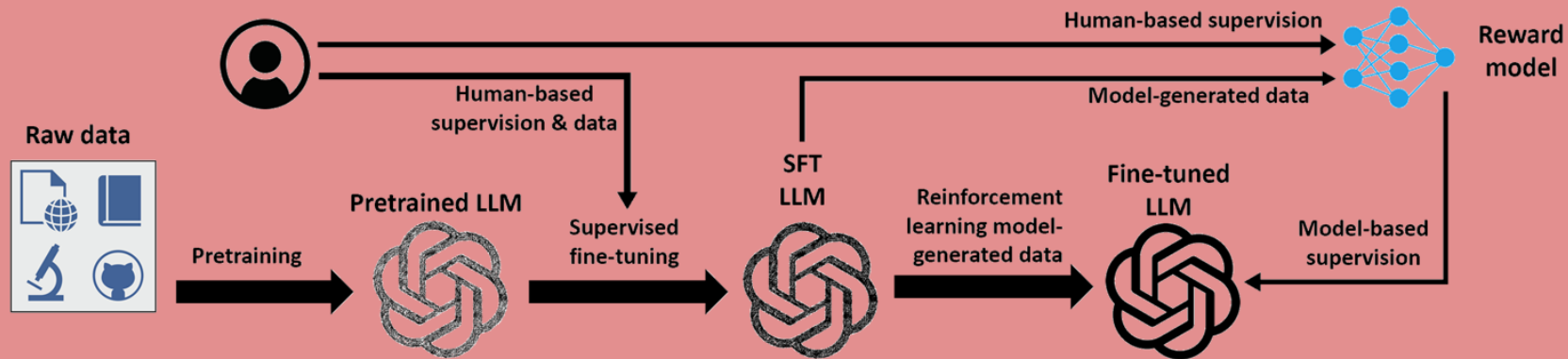


Inference related

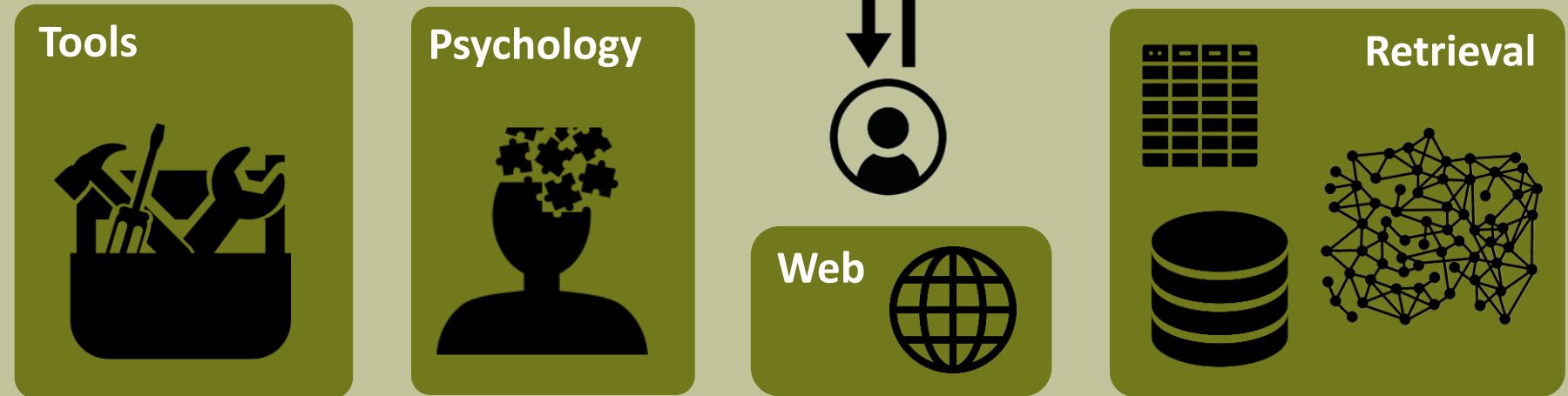


The Emergence of the „Generative AI Ecosystem”

Training related

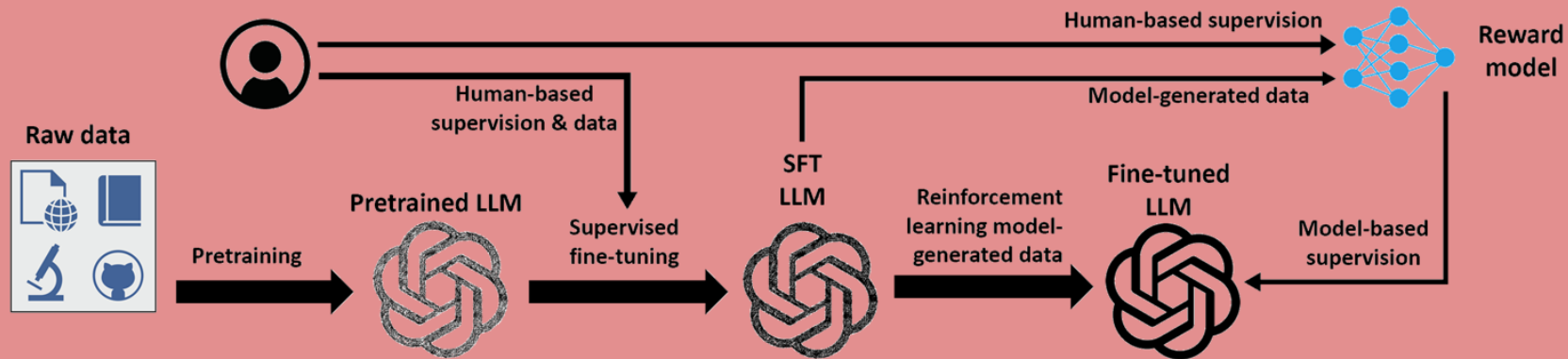


Inference related



The Emergence of the „Generative AI Ecosystem”

Training related



Inference related

Prompting Structures

Tools

Psychology

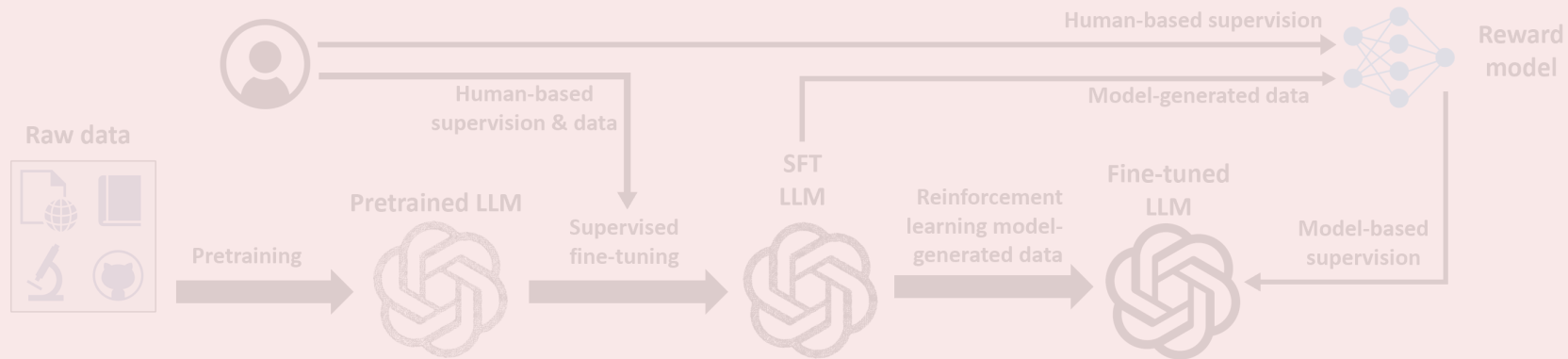
Replies („thoughts”)
Prompts

Web

Retrieval

The Emergence of the „Generative AI Ecosystem”

Training related



Inference related

Prompting Structures

Tools

Psychology

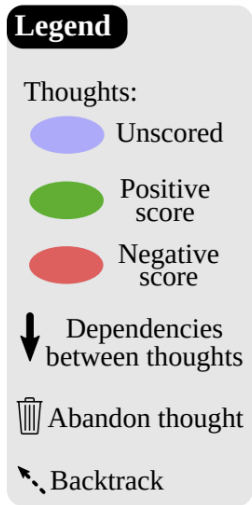
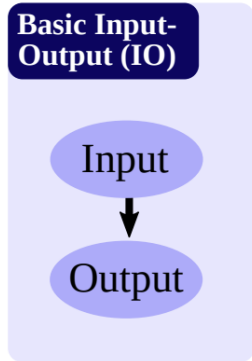
Web

Retrieval

Prompting Paradigms



Prompting Paradigms



Prompting Paradigms



Basic Input-Output (IO)

Input



Output


Legend

Thoughts:

 Unscored

 Positive score

 Negative score

 Dependencies between thoughts

 Abandon thought

 Backtrack

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?

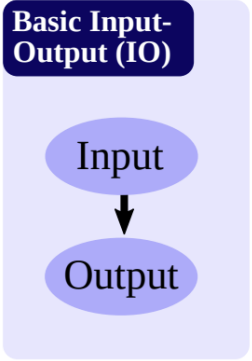
Model Output

A: The answer is 27. 

Prompting Paradigms



In-context examples



Legend

Thoughts:

- Unscored
- Positive score
- Negative score

↓ Dependencies between thoughts

🗑️ Abandon thought

↶ Backtrack

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. ❌

Prompting Paradigms



Basic Input-Output (IO)

Input



Output


Legend

Thoughts:

 Unscored

 Positive score

 Negative score

 Dependencies between thoughts

 Abandon thought

 Backtrack

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.

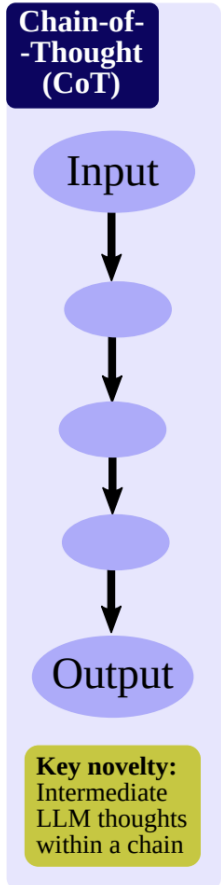
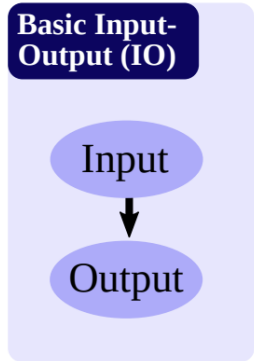
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. 

Prompting Paradigms

[Wei et al.,
Jan'22]



Legend

- Thoughts:
 - Unscored (light blue oval)
 - Positive score (green oval)
 - Negative score (red oval)
- ↓ Dependencies between thoughts
- 🗑️ Abandon thought
- ↩️ Backtrack

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.

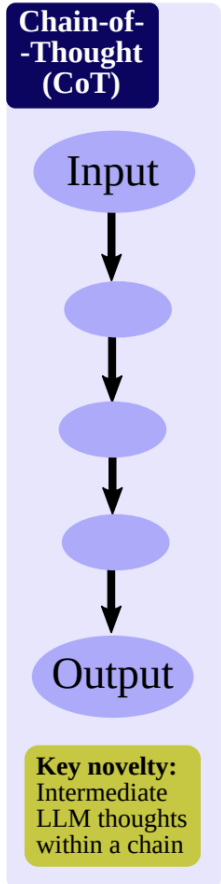
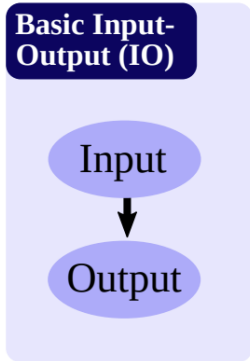
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Prompting Paradigms

[Wei et al., Jan'22]



Legend

- Thoughts:
 - Unscored (light blue oval)
 - Positive score (green oval)
 - Negative score (red oval)
- ↓ Dependencies between thoughts
- 🗑️ Abandon thought
- ↩️ Backtrack

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.

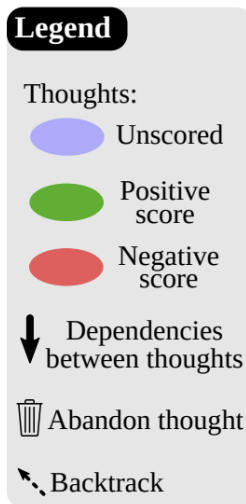
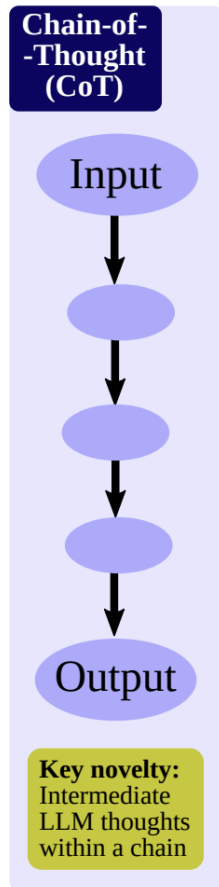
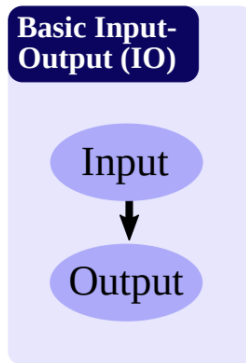
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

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. ✅

Prompting Paradigms

[Wei et al.,
Jan'22]



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?

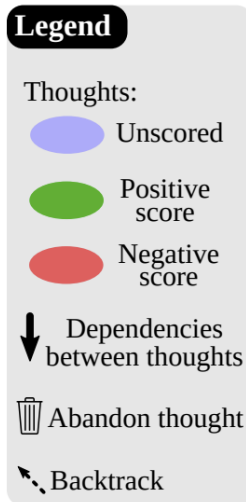
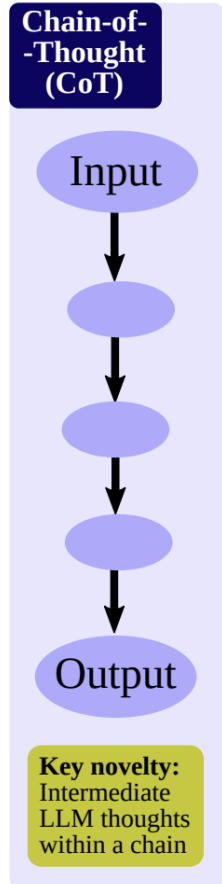
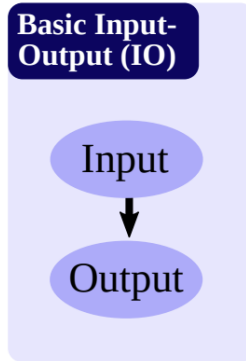
Model Output

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. ✓

Prompting Paradigms

[Wei et al.,
Jan'22]

[Kojima et al.,
May'22]



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?

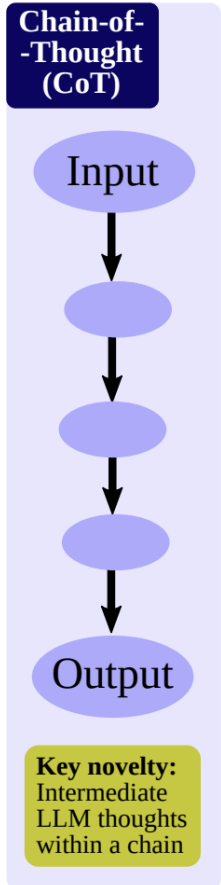
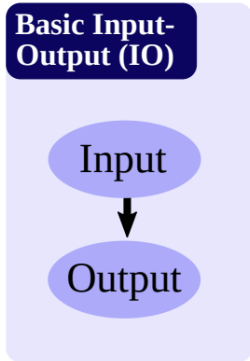
Model Output

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. ✓

Prompting Paradigms

[Wei et al.,
Jan'22]

[Kojima et al.,
May'22]



Legend

Thoughts:

- Unscored
- Positive score
- Negative score

↓ Dependencies between thoughts

🗑️ Abandon thought

↶ Backtrack

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 tennis balls 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?

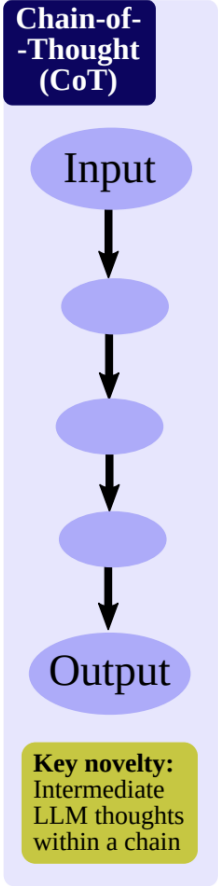
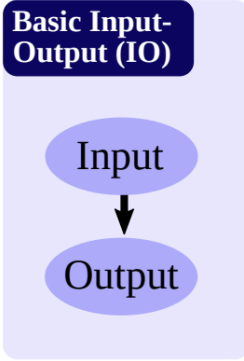
Model Output

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. ✓

Prompting Paradigms

[Wei et al.,
Jan'22]

[Kojima et al.,
May'22]



Legend

- Thoughts:
 - Unscored (light blue oval)
 - Positive score (green oval)
 - Negative score (red oval)
- ↓ Dependencies between thoughts
- 🗑️ Abandon thought
- ↶ Backtrack

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 tennis balls 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? **Let's proceed step by step.**

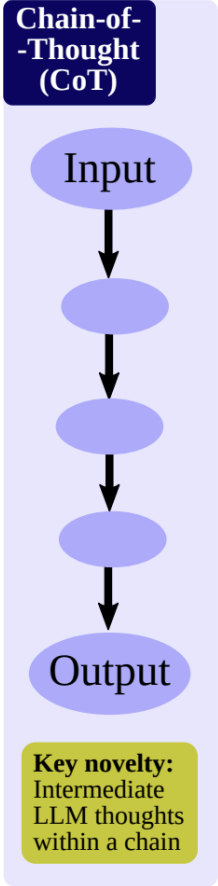
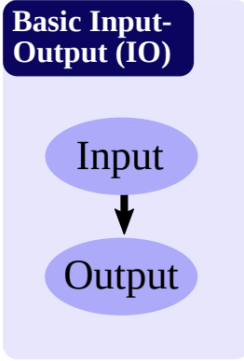
Model Output

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. ✓

Prompting Paradigms

[Wei et al.,
Jan'22]

[Kojima et al.,
May'22]



Legend

- Thoughts:
 - Unscored (light blue oval)
 - Positive score (green oval)
 - Negative score (red oval)
- ↓ Dependencies between thoughts
- 🗑️ Abandon thought
- ↩️ Backtrack

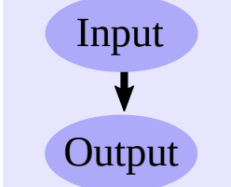
Prompting Paradigms

[Wei et al.,
Jan'22]

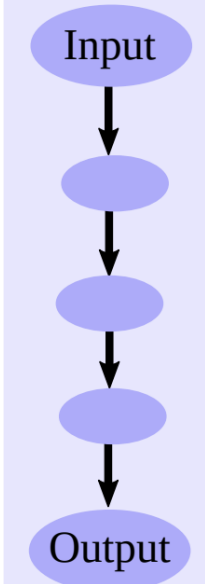
[Wang et al., [Kojima et al.,
March'22] May'22]



Basic Input-Output (IO)

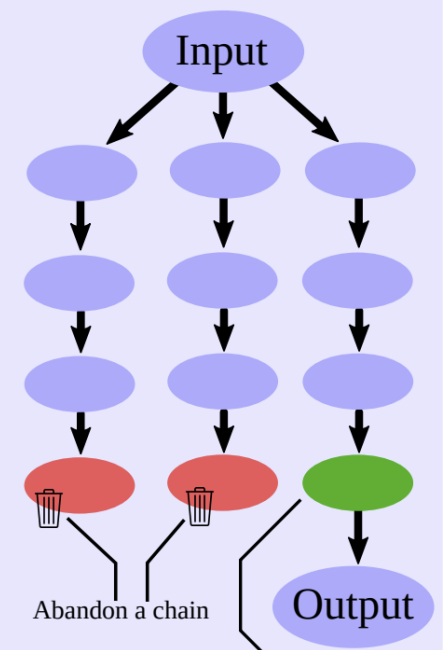


Chain-of-Thought (CoT)








Key novelty:
Intermediate LLM thoughts within a chain

Multiple CoTs (CoT-SC)



Key novelty (beyond CoT):
Harnessing multiple independent chains of thoughts

Legend

- Thoughts:
-  Unscored
 -  Positive score
 -  Negative score
- ↓ Dependencies between thoughts
-  Abandon thought
-  Backtrack

Prompting Paradigms

[Wei et al.,
Jan'22]

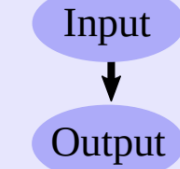
[Wang et al., [Kojima et al.,
March'22] May'22]

[Long et al.,
May'23]

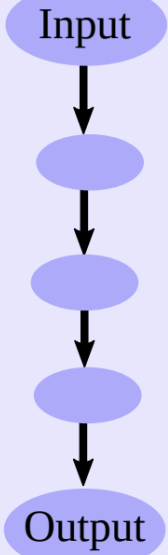
[Yao et al.,
May'23]



Basic Input-Output (IO)

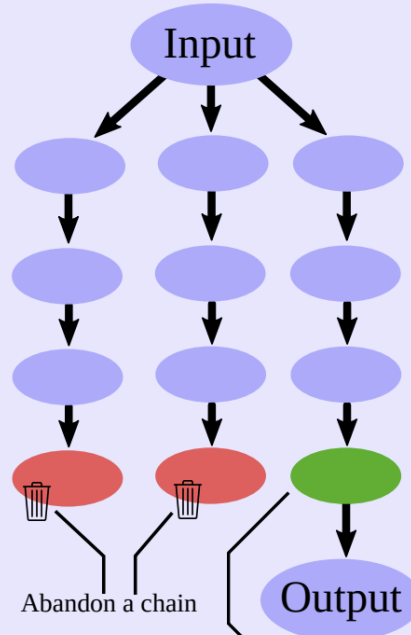


Chain-of-Thought (CoT)



Key novelty:
Intermediate LLM thoughts within a chain

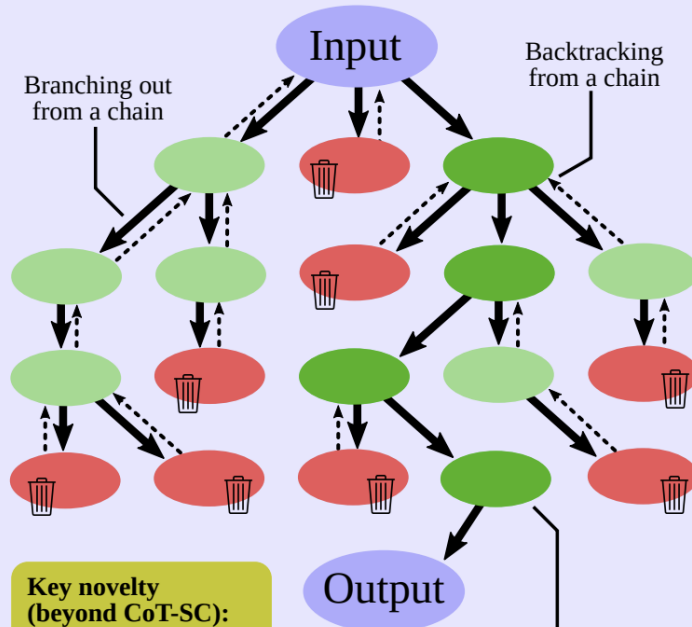
Multiple CoTs (CoT-SC)



Key novelty (beyond CoT):
Harnessing multiple independent chains of thoughts

Selecting a chain with the best score

Tree of Thoughts (ToT)






Key novelty (beyond CoT-SC):
Generating several new thoughts based on a given arbitrary thought, exploring it further, and possibly backtracking from it


Intermediate thoughts are also scored

Legend

Thoughts:

-  Unscored
-  Positive score
-  Negative score

 Dependencies between thoughts

 Abandon thought

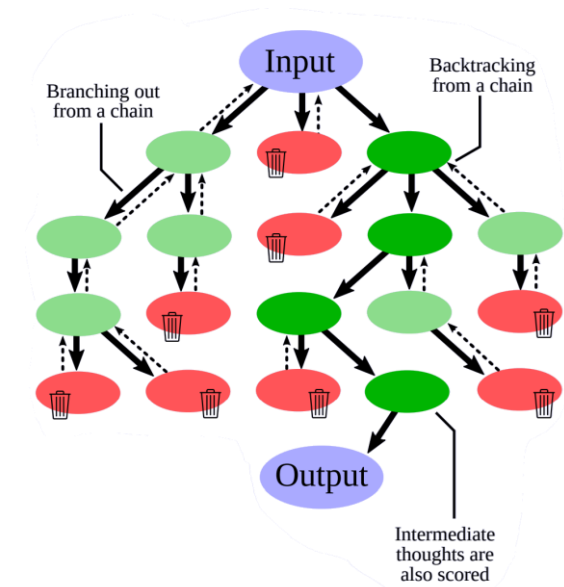
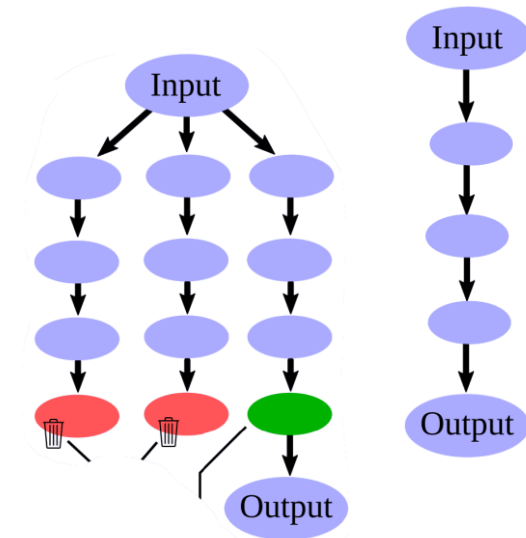
 Backtrack

The Next Step – Graphical Reasoning

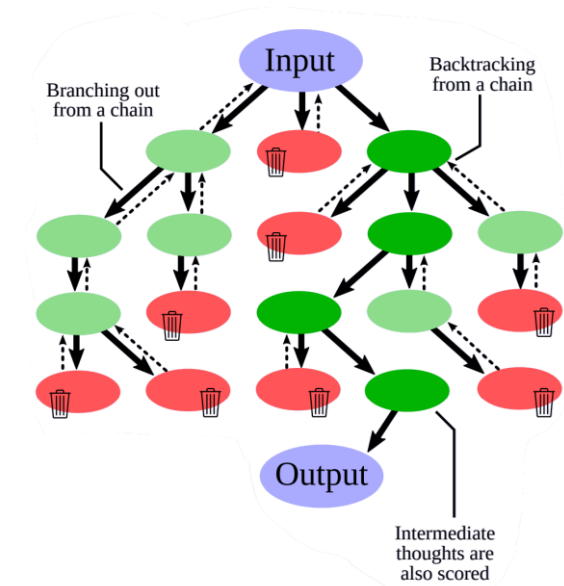
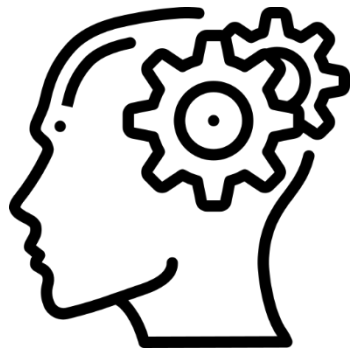
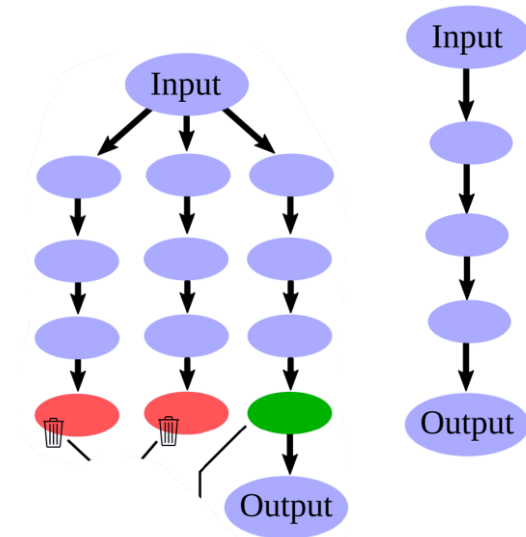
The Next Step – Graphical Reasoning

Inspired by human thoughts.

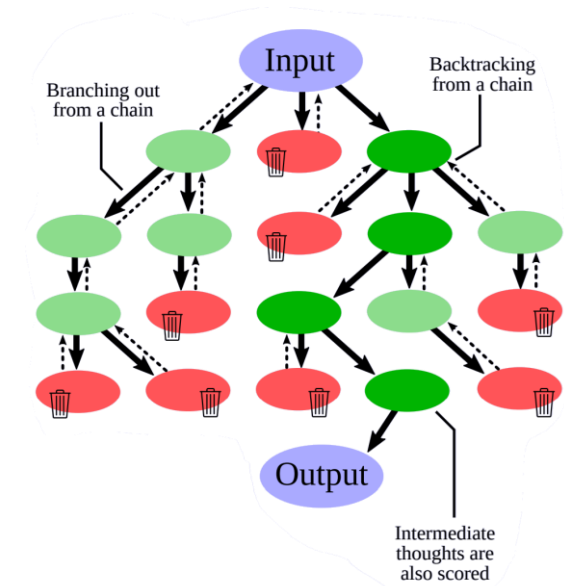
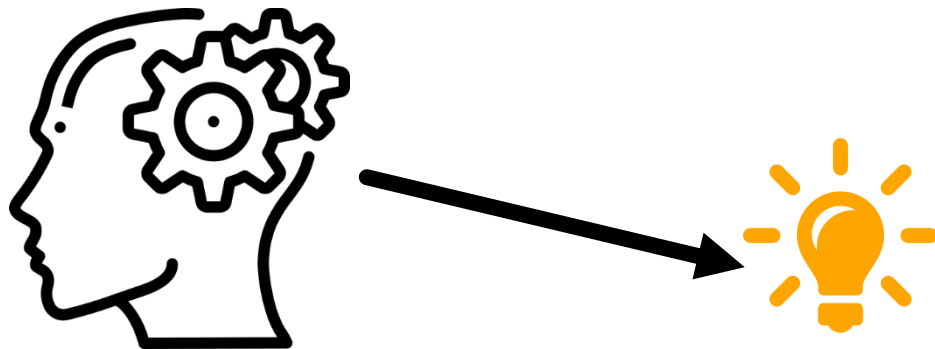
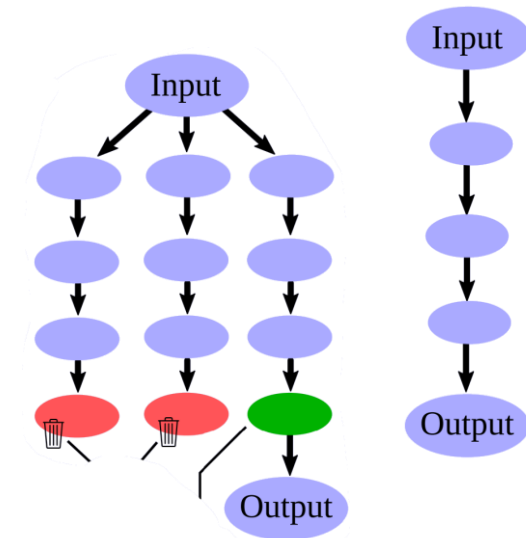
Inspiration for Next Step: Human Thinking



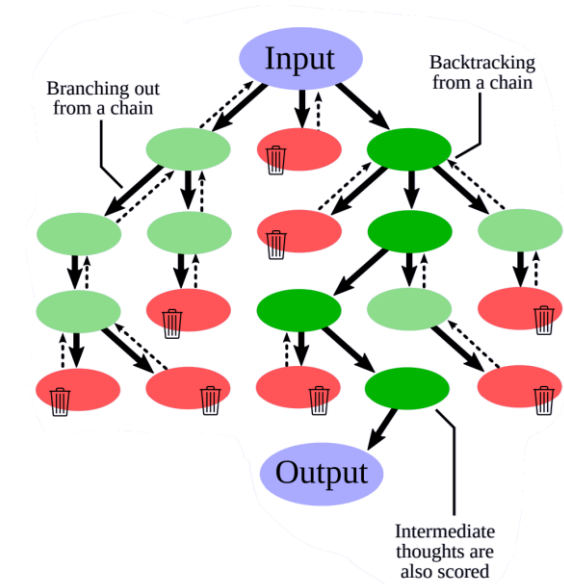
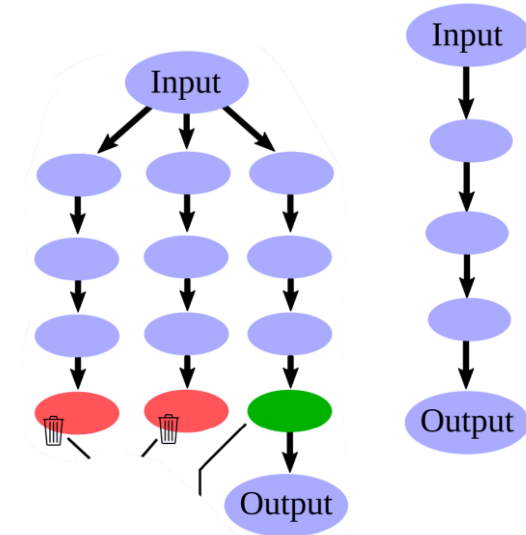
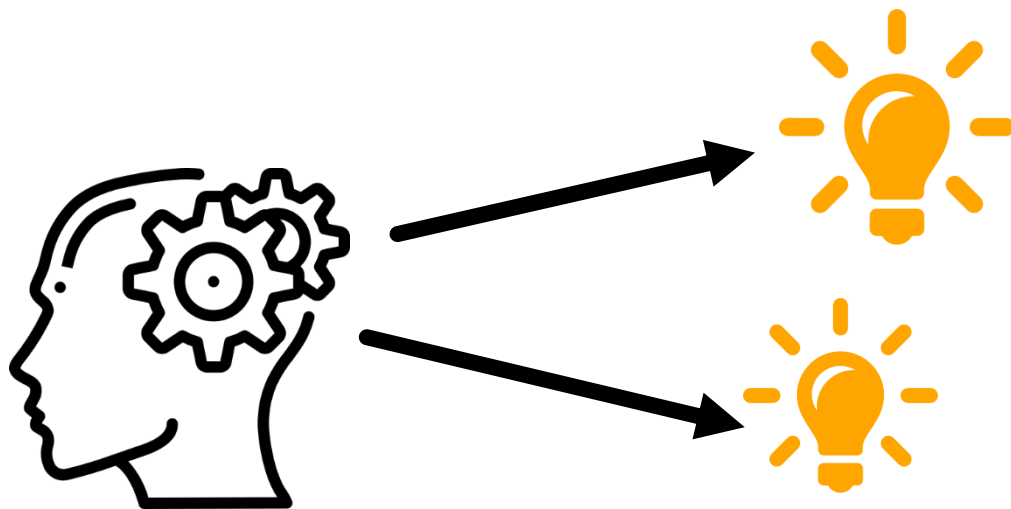
Inspiration for Next Step: Human Thinking



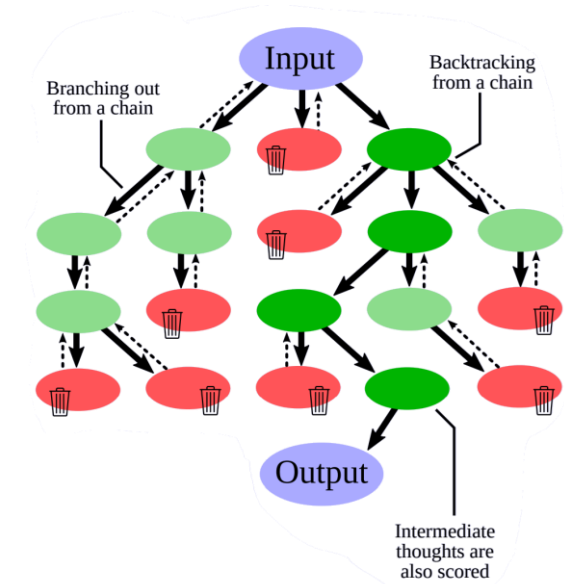
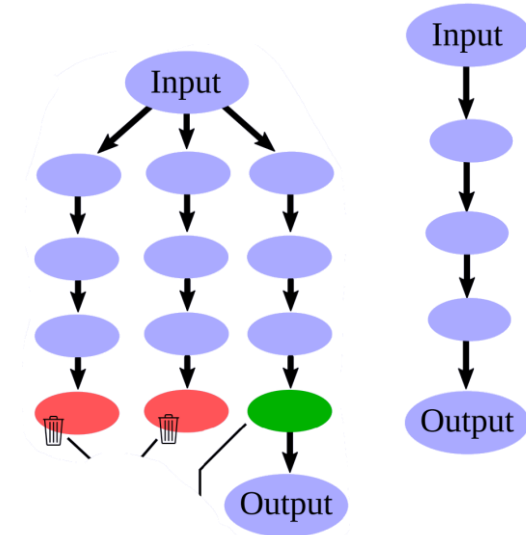
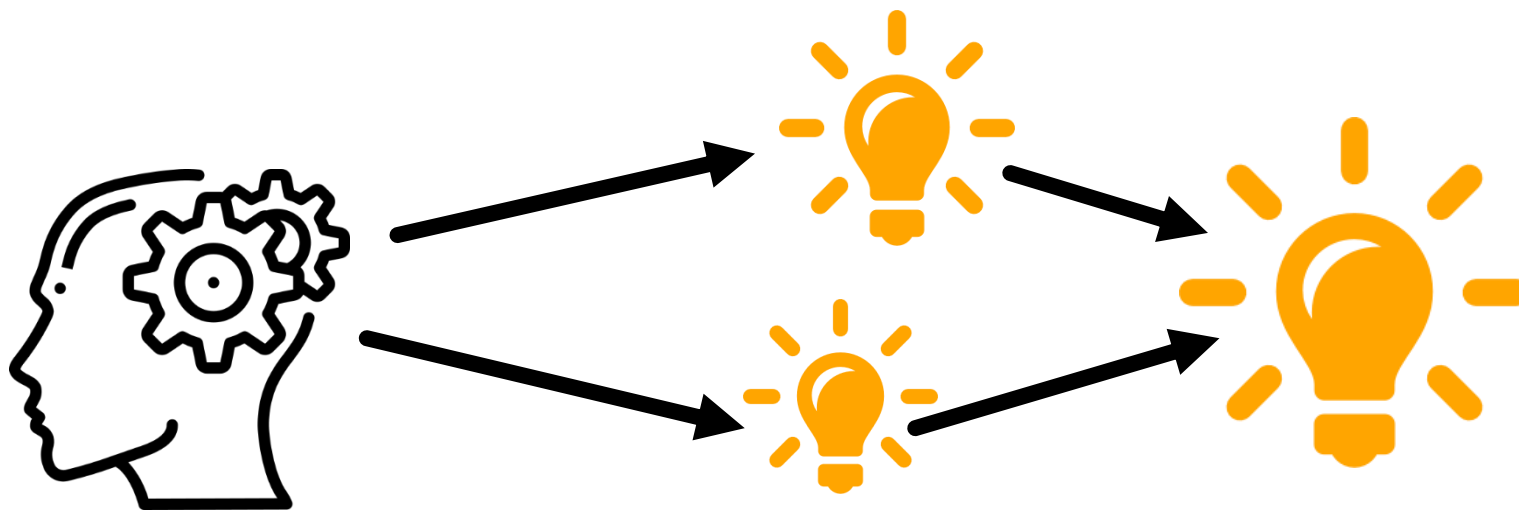
Inspiration for Next Step: Human Thinking



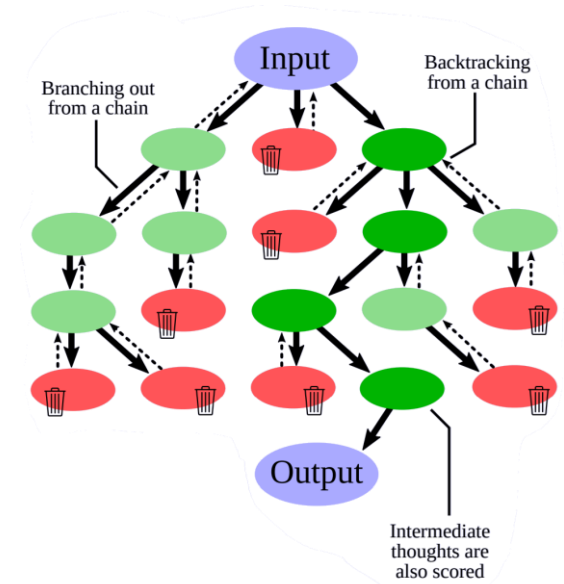
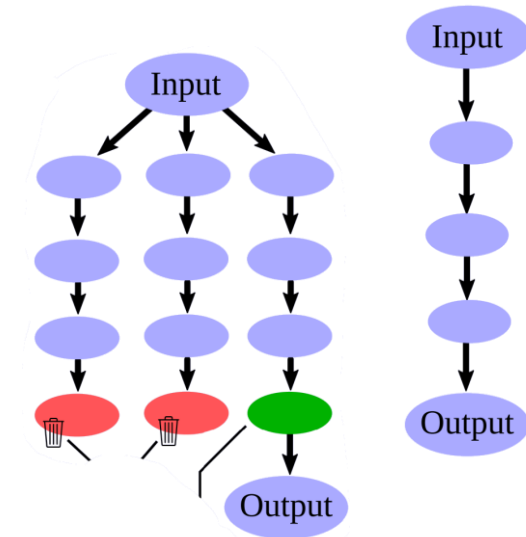
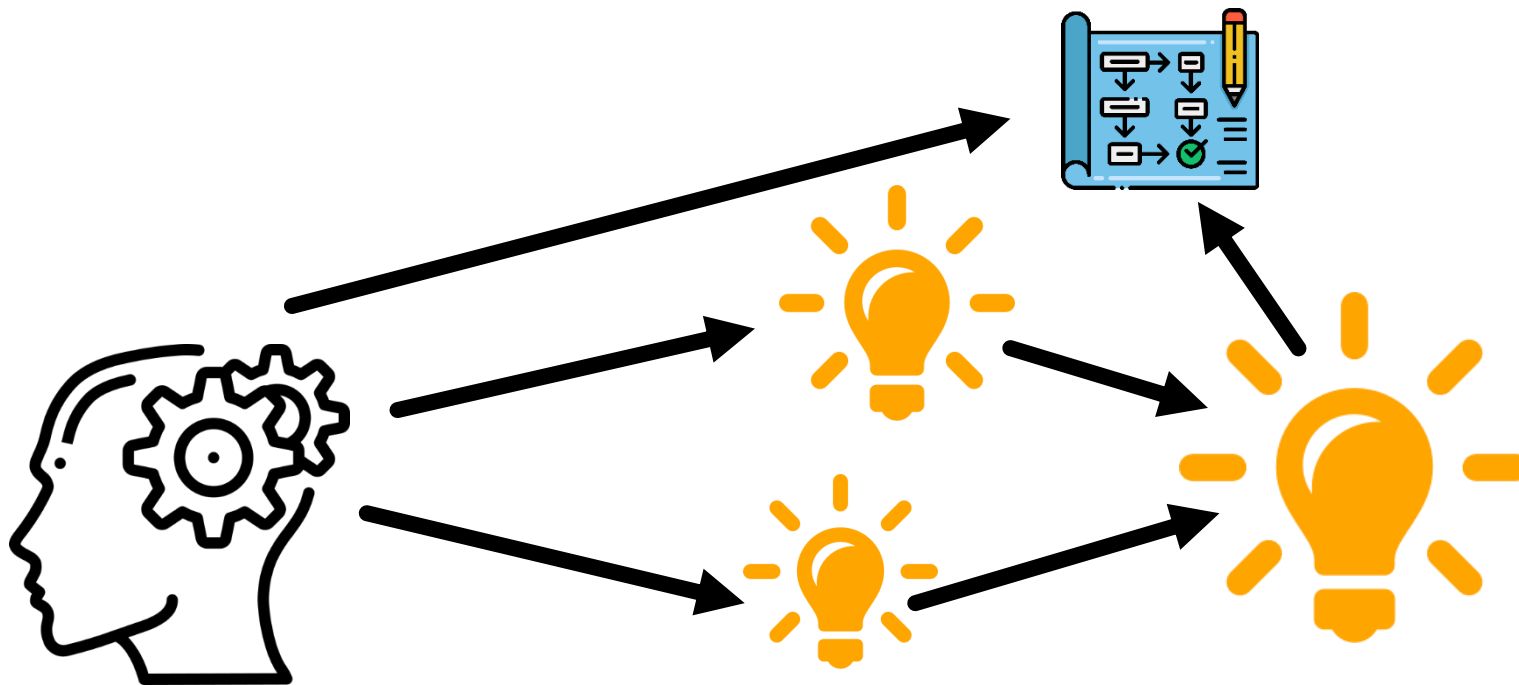
Inspiration for Next Step: Human Thinking



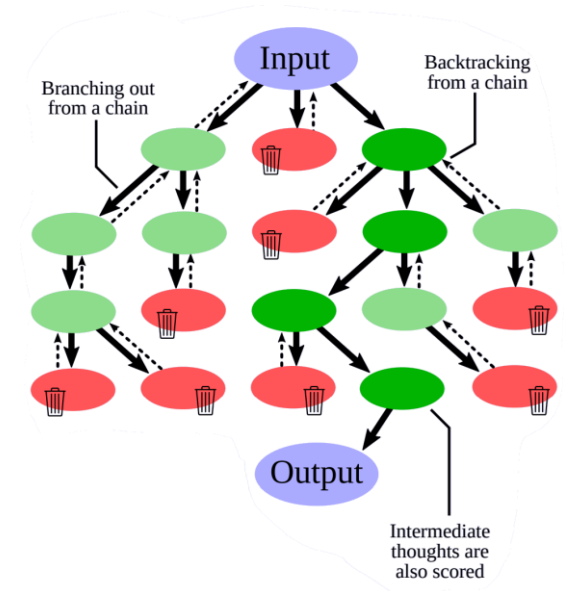
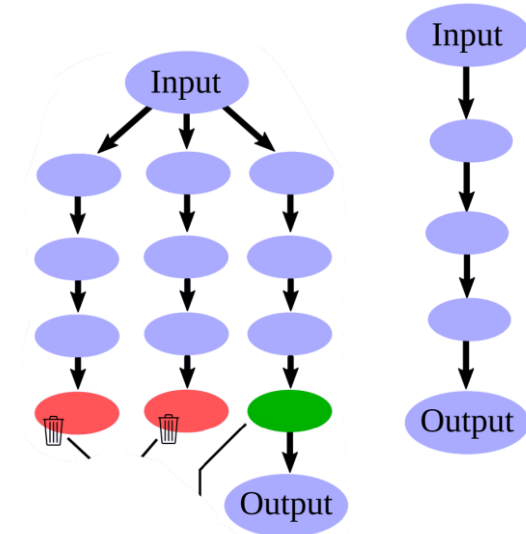
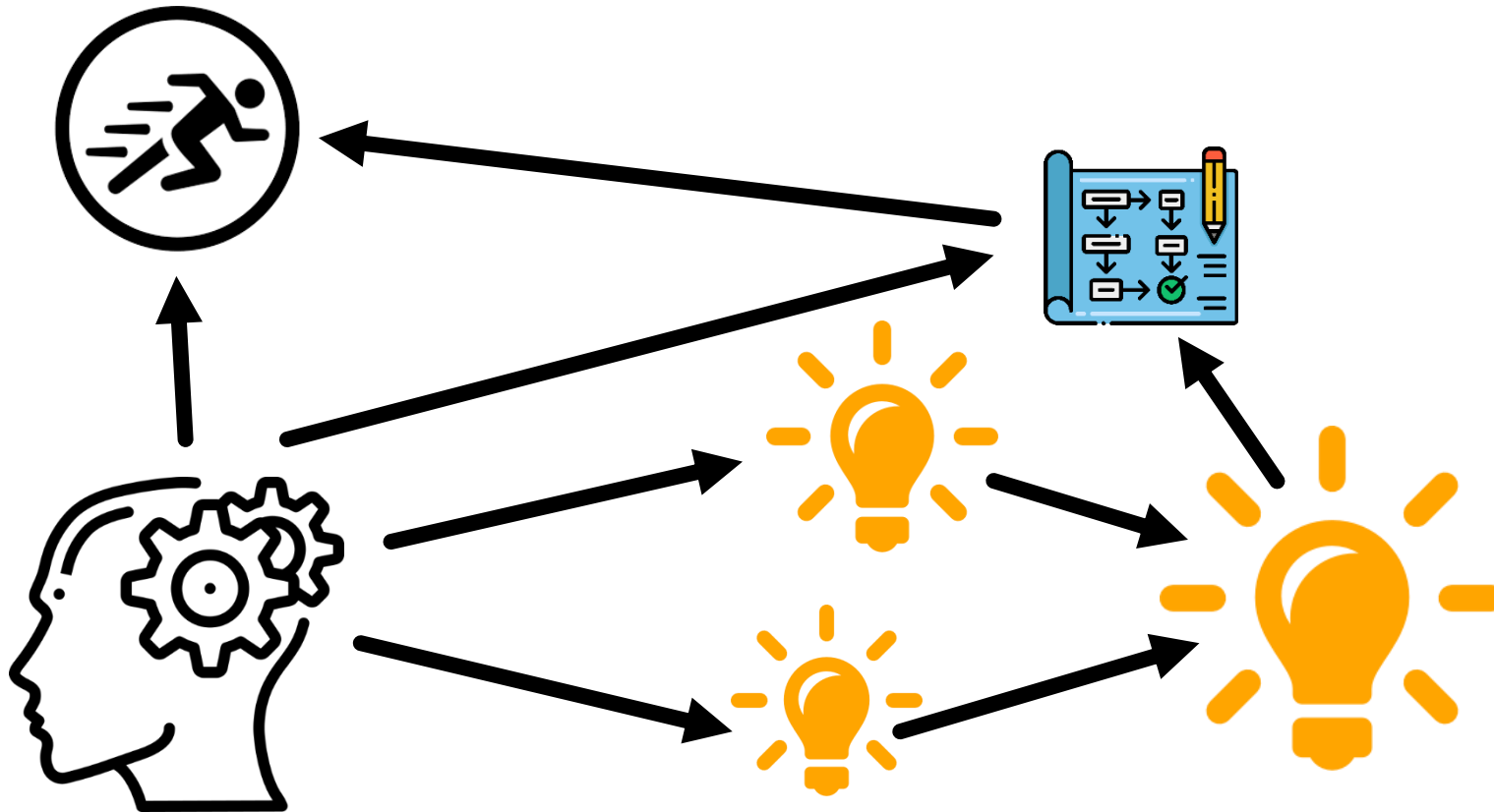
Inspiration for Next Step: Human Thinking



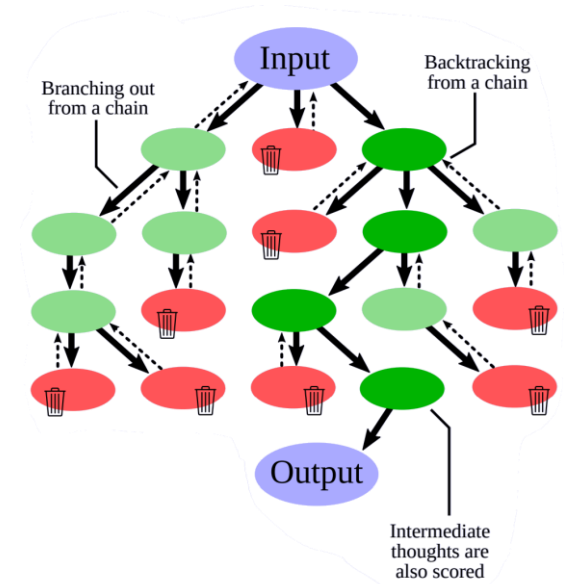
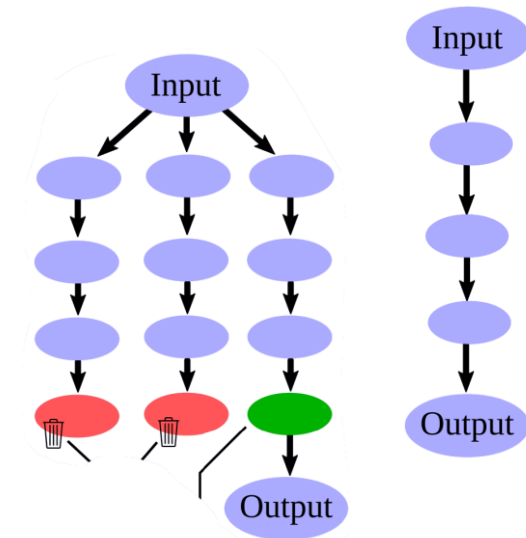
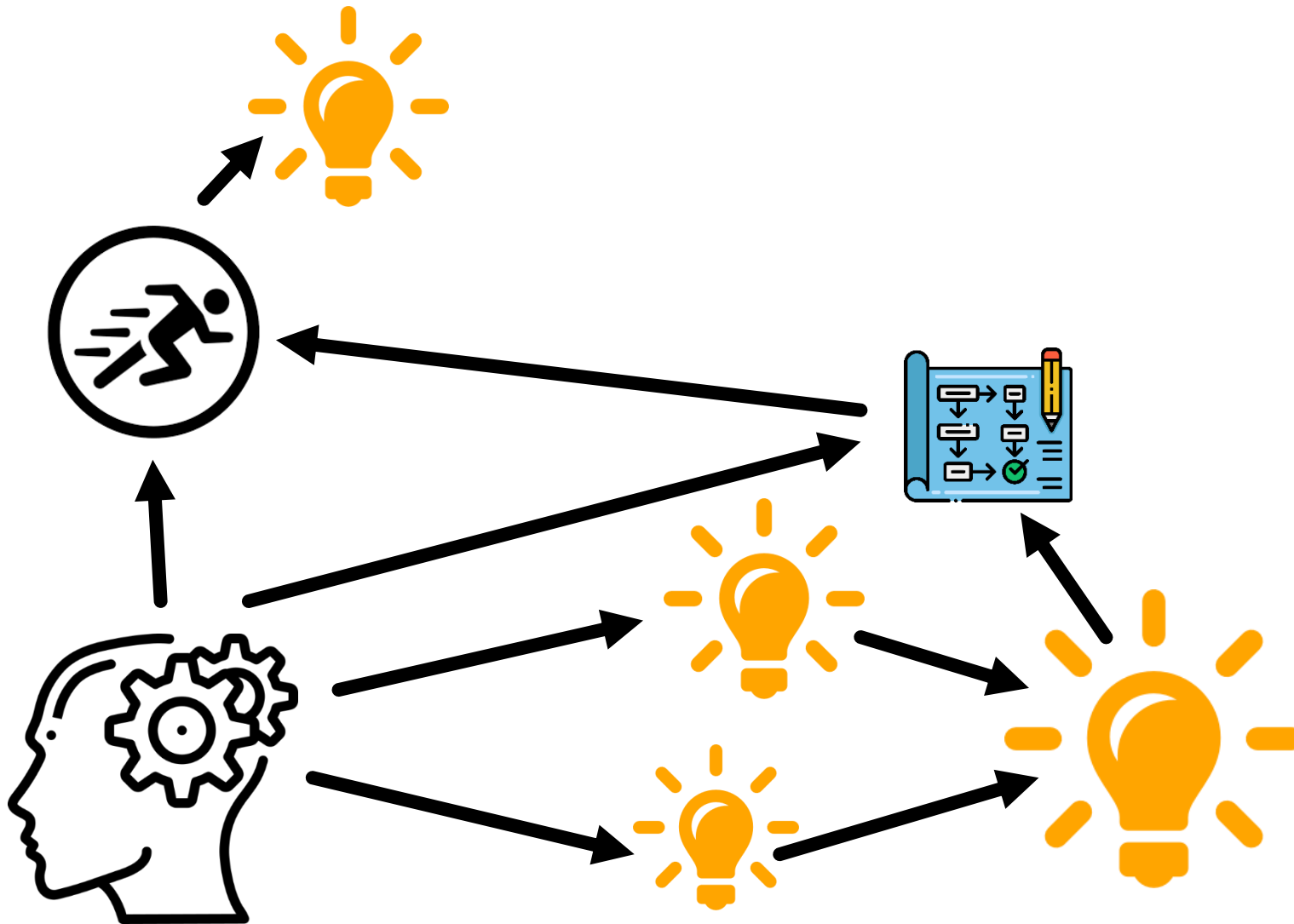
Inspiration for Next Step: Human Thinking



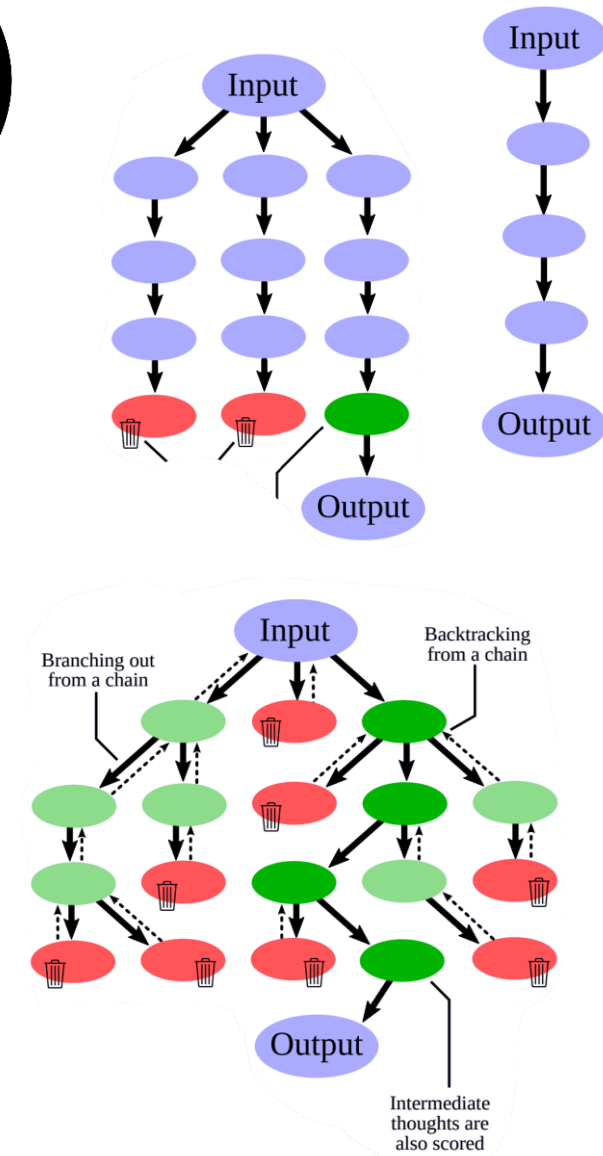
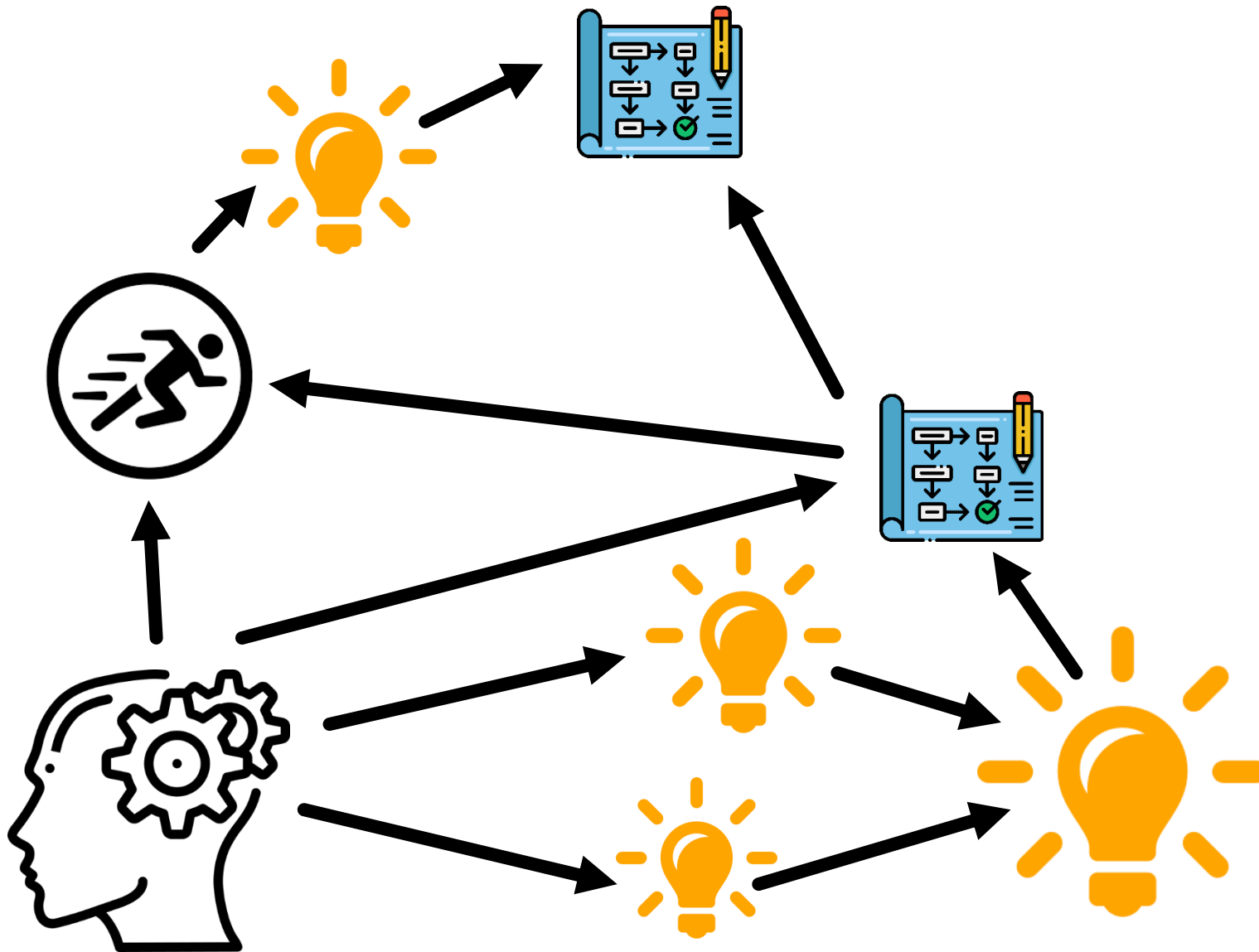
Inspiration for Next Step: Human Thinking



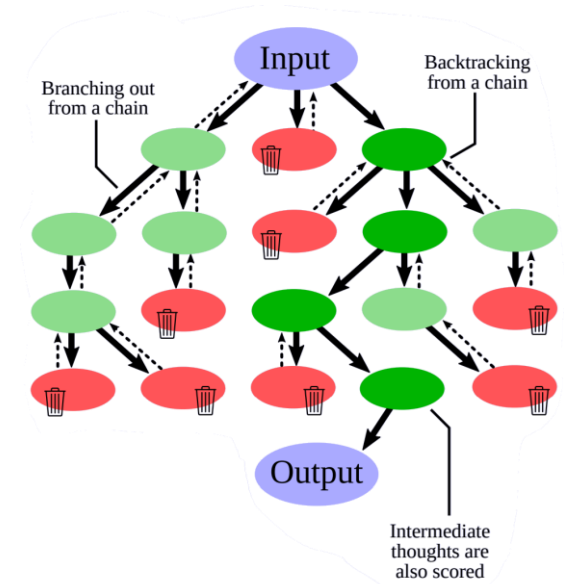
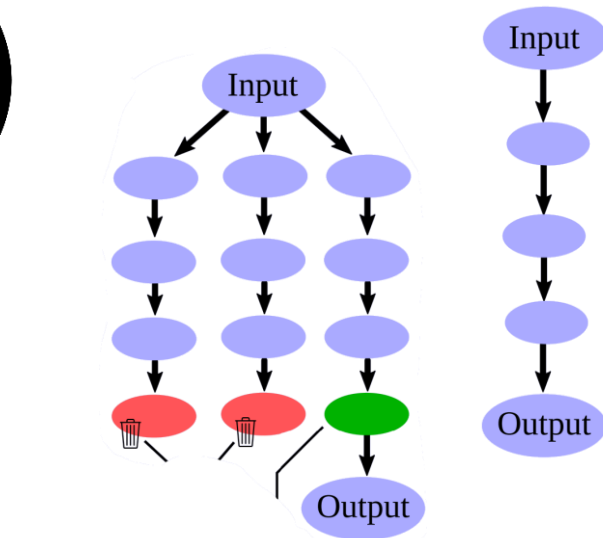
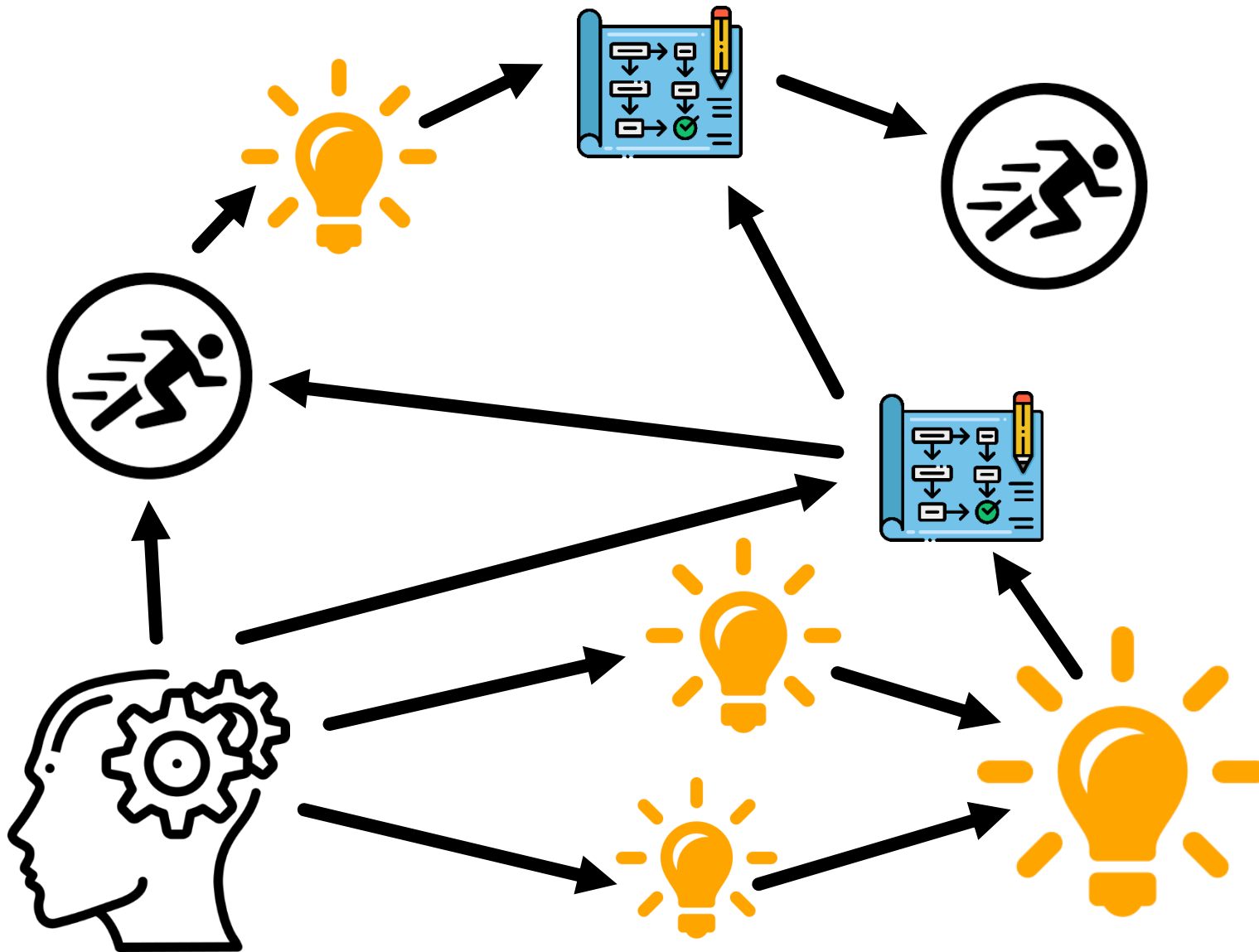
Inspiration for Next Step: Human Thinking



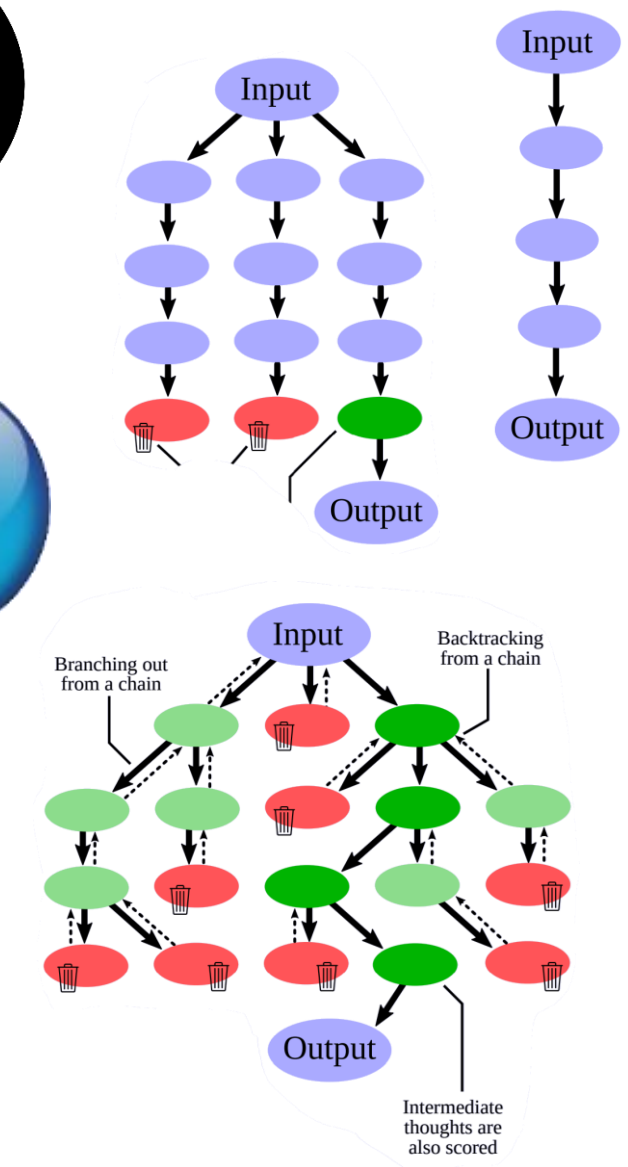
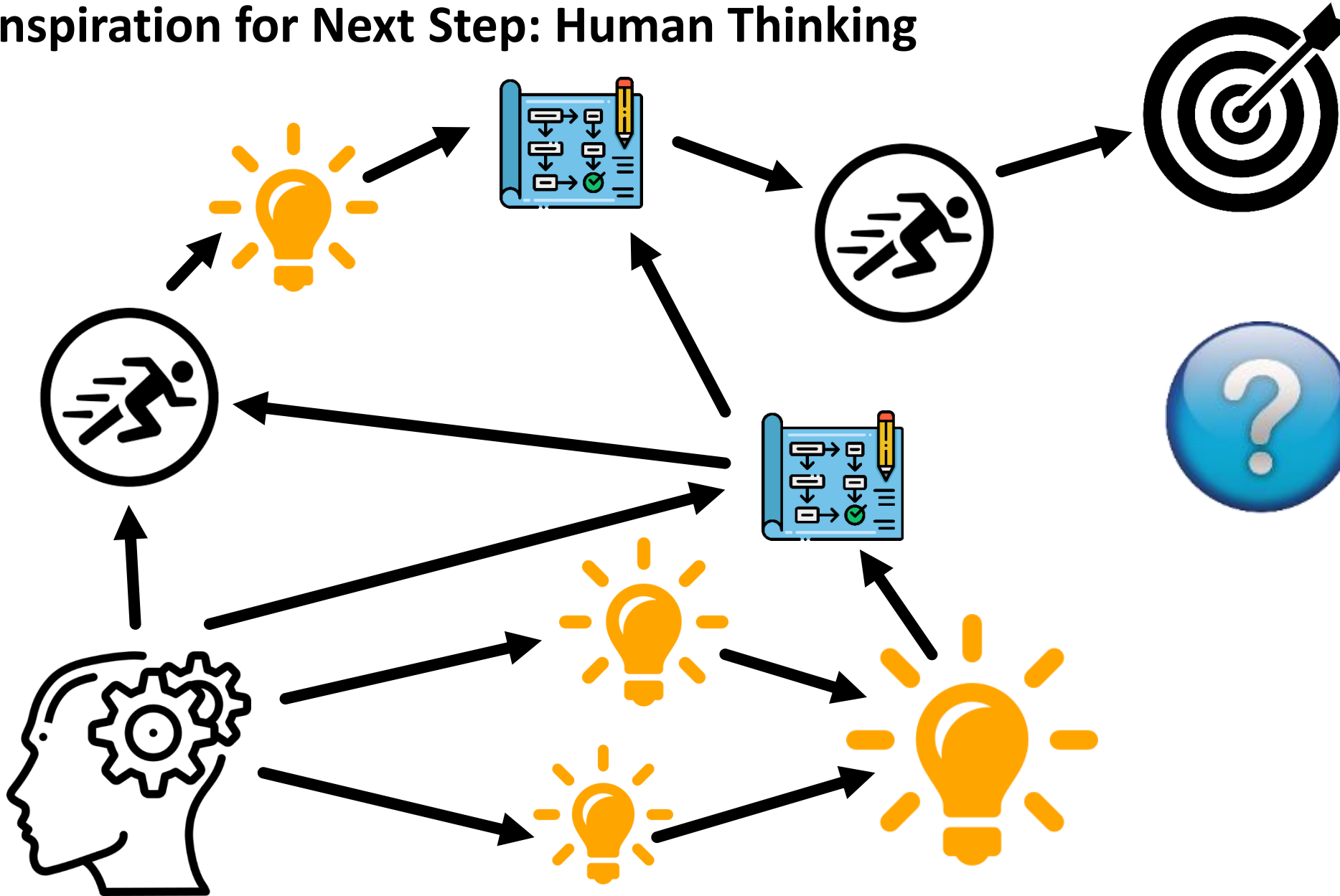
Inspiration for Next Step: Human Thinking



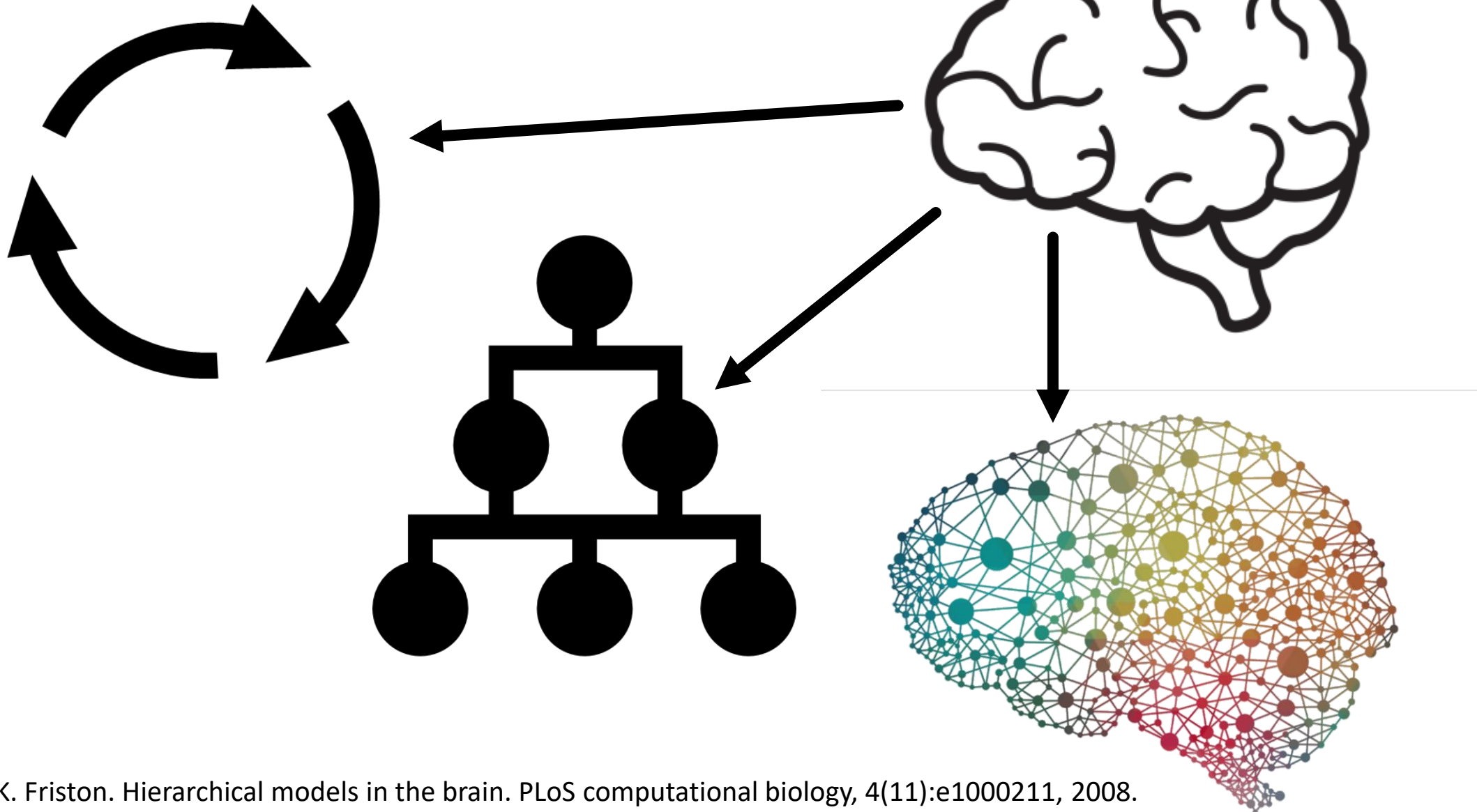
Inspiration for Next Step: Human Thinking



Inspiration for Next Step: Human Thinking



Inspiration for Next Step: Brain Structure [1]



[1] K. Friston. Hierarchical models in the brain. PLoS computational biology, 4(11):e1000211, 2008.

Prompting Paradigms: Graph of Thoughts

[Wei et al.,
Jan'22]

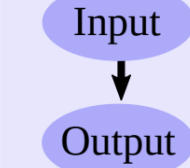
[Wang et al.,
March'22]

[Long et al.,
May'23]

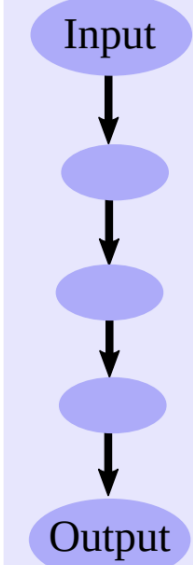
[Yao et al.,
May'23]



Basic Input-Output (IO)

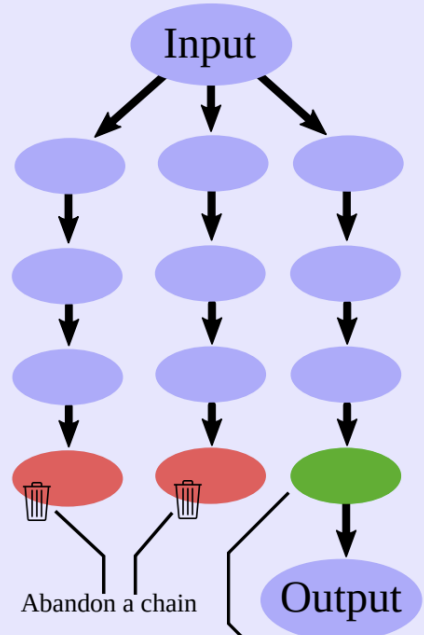


Chain-of-Thought (CoT)



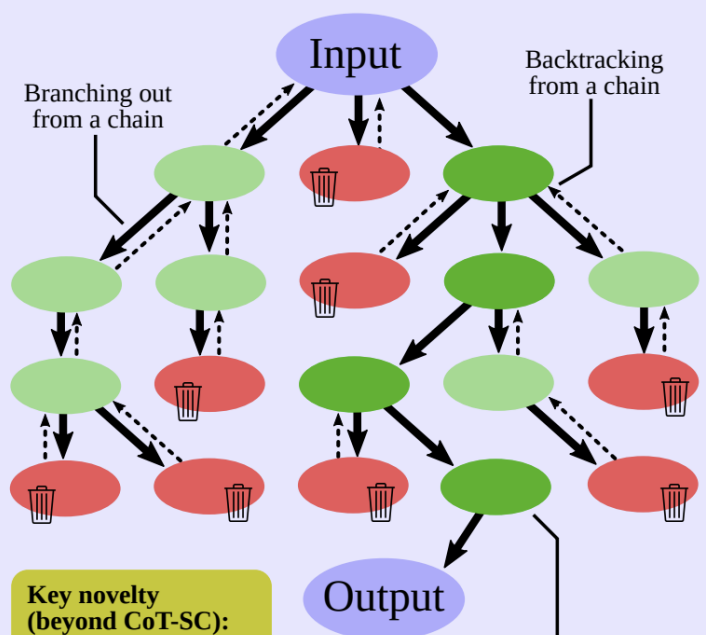
Key novelty:
Intermediate LLM thoughts within a chain

Multiple CoTs (CoT-SC)



Key novelty (beyond CoT):
Harnessing multiple independent chains of thoughts

Tree of Thoughts (ToT)



Key novelty (beyond CoT-SC):
Generating several new thoughts based on a given arbitrary thought, exploring it further, and possibly backtracking from it

Intermediate thoughts are also scored

Legend

- Thoughts:
 - Unscored
 - Positive score
 - Negative score
- ↓ Dependencies between thoughts
- 🗑️ Abandon thought
- ↔️ Backtrack

Prompting Paradigms: Graph of Thoughts

[Wei et al., Jan'22]

[Wang et al., March'22]

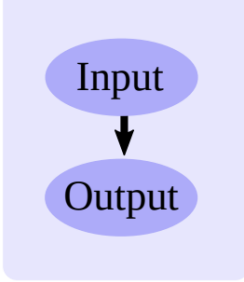
[Long et al., May'23]

[Yao et al., May'23]

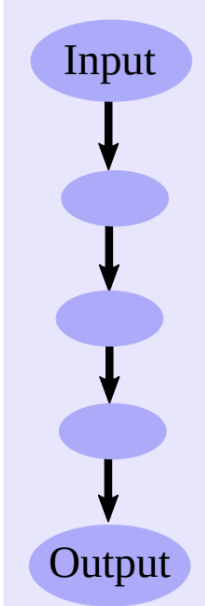
[Our work, August'23]



Basic Input-Output (IO)

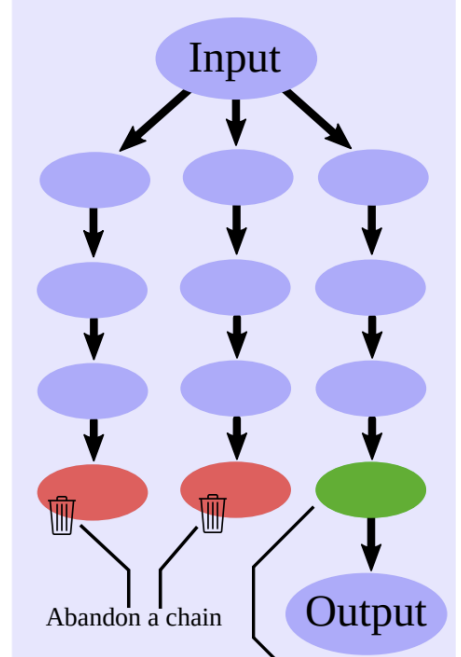


Chain-of-Thought (CoT)



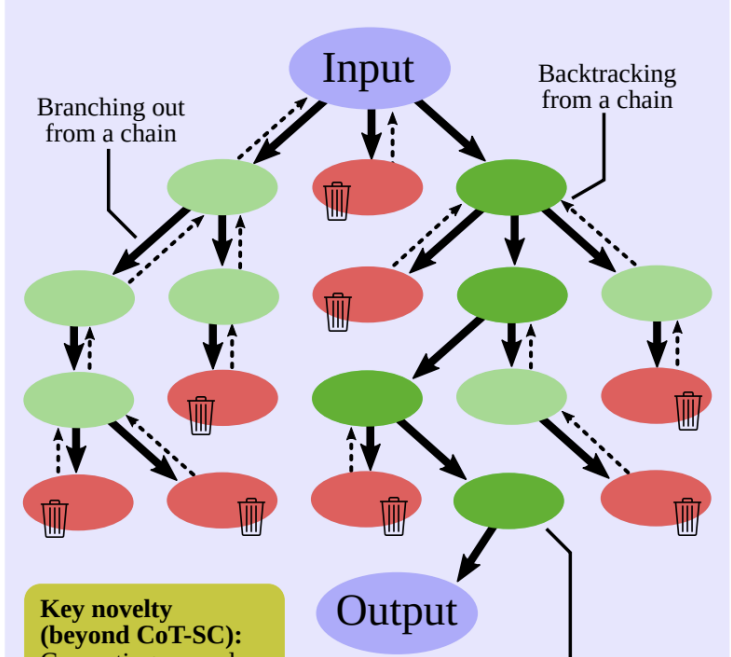
Key novelty: Intermediate LLM thoughts within a chain

Multiple CoTs (CoT-SC)



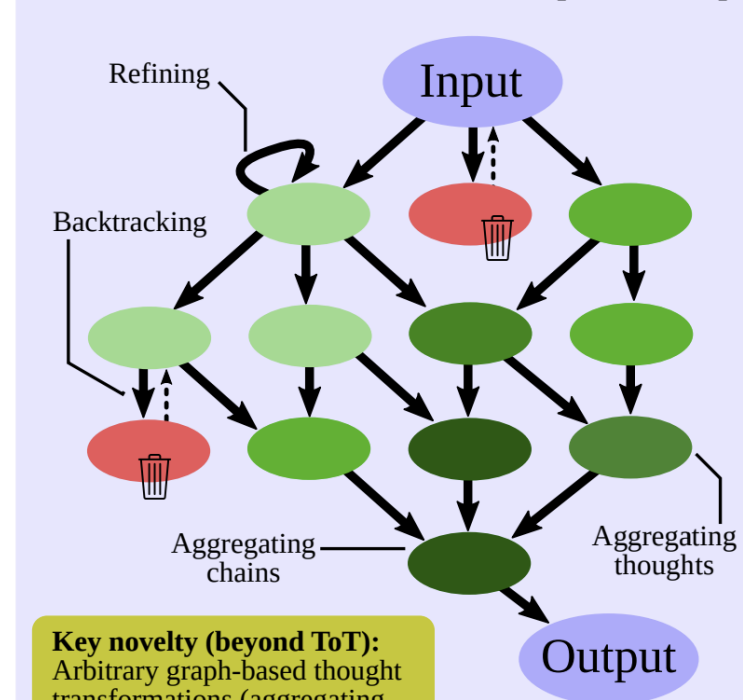
Key novelty (beyond CoT): Harnessing multiple independent chains of thoughts

Tree of Thoughts (ToT)



Key novelty (beyond CoT-SC): Generating several new thoughts based on a given arbitrary thought, exploring it further, and possibly backtracking from it

Graph of Thoughts (GoT)



Key novelty (beyond ToT): Arbitrary graph-based thought transformations (aggregating thoughts into a new one, looping over a thought to refine it)

[This work]

Legend

- Thoughts:
- Unscored
- Positive score
- Negative score
- ↓ Dependencies between thoughts
- 🗑️ Abandon thought
- ↔️ Backtrack

Prompting Paradigms: Graph of Thoughts

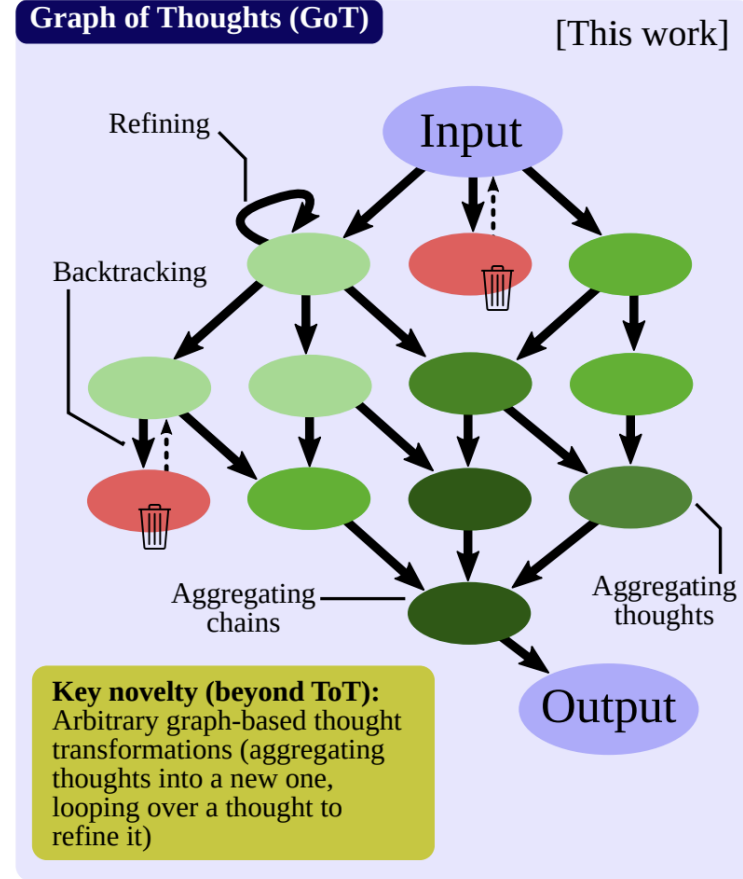
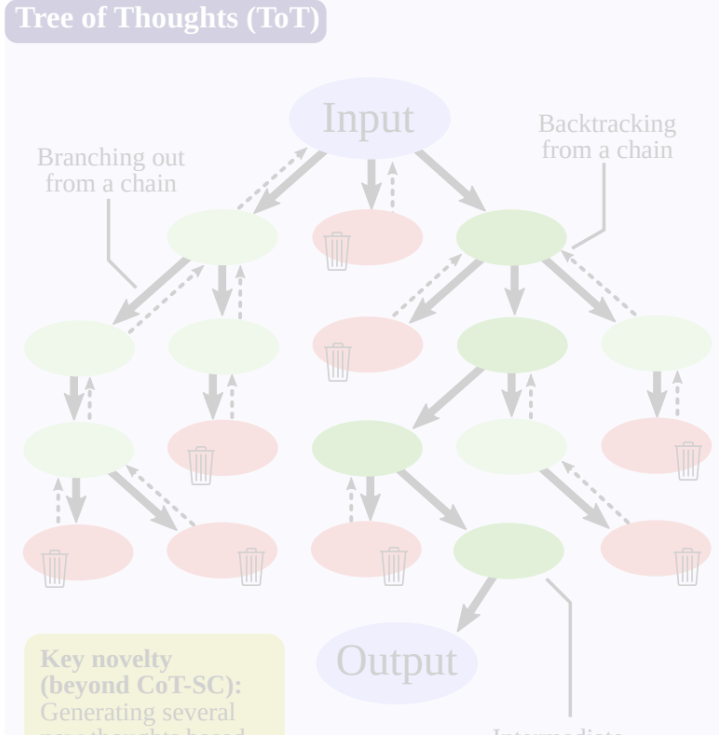
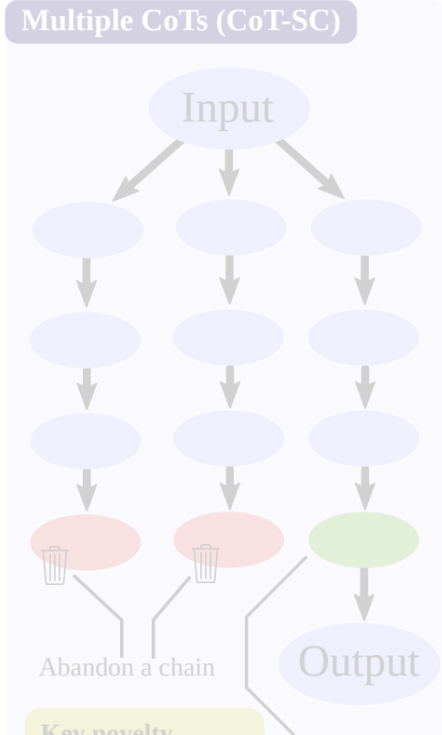
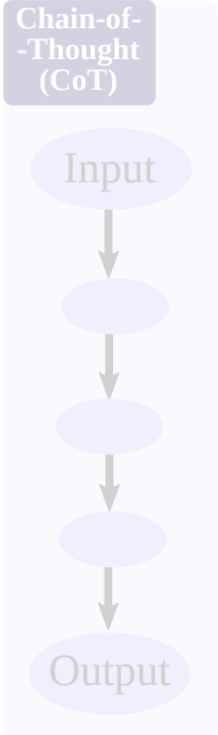
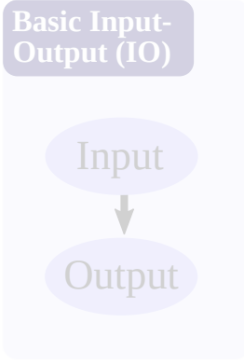
„Thought transformations“

[Wei et al., Jan'22]

[Wang et al., March'22]

[Long et al., May'23]

[Yao et al., May'23]



Legend

Thoughts:

- Unscored
- Positive score
- Negative score

↓ Dependencies between thoughts

🗑️ Abandon thought

↔️ Backtrack

Key novelty: Intermediate LLM thoughts within a chain

Key novelty (beyond CoT): Harnessing multiple independent chains of thoughts

Selecting a chain with the best score

Key novelty (beyond CoT-SC): Generating several new thoughts based on a given arbitrary thought, exploring it further, and possibly backtracking from it

Intermediate thoughts are also scored

Key novelty (beyond ToT): Arbitrary graph-based thought transformations (aggregating thoughts into a new one, looping over a thought to refine it)

Prompting Paradigms: Graph of Thoughts

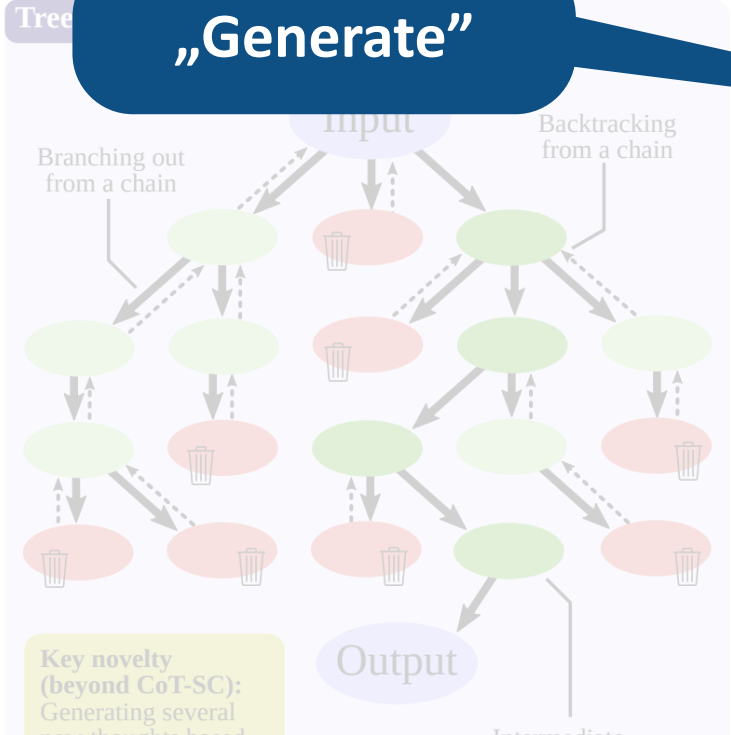
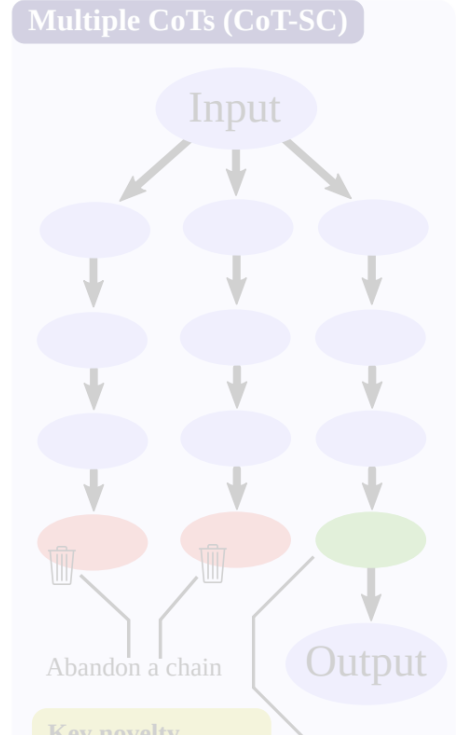
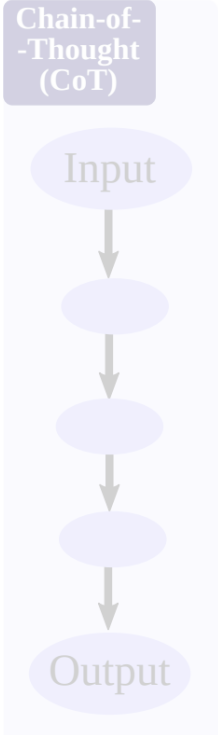
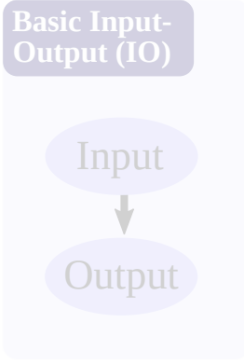
„Thought transformations“

[Wei et al., Jan'22]

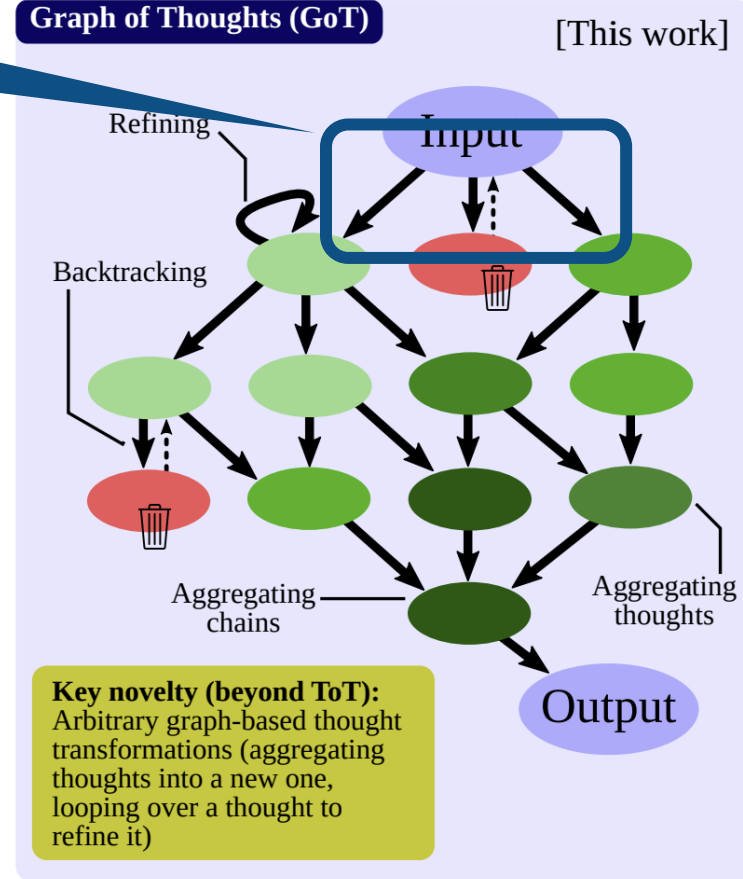
[Wang et al., March'22]

[Long et al., May'23]

[Yao et al., May'23]



„Generate“



Legend

Thoughts:

- Unscored (light blue oval)
- Positive score (green oval)
- Negative score (red oval)

↓ Dependencies between thoughts

🗑️ Abandon thought

↔️ Backtrack

Key novelty: Intermediate LLM thoughts within a chain

Key novelty (beyond CoT): Harnessing multiple independent chains of thoughts

Selecting a chain with the best score

Key novelty (beyond CoT-SC): Generating several new thoughts based on a given arbitrary thought, exploring it further, and possibly backtracking from it

Intermediate thoughts are also scored

Key novelty (beyond ToT): Arbitrary graph-based thought transformations (aggregating thoughts into a new one, looping over a thought to refine it)

Prompting Paradigms: Graph of Thoughts

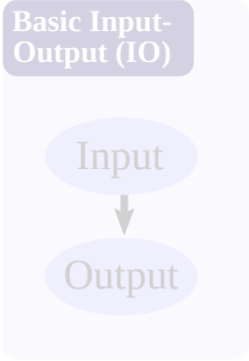
„Thought transformations”

[Wei et al., Jan'22]

[Wang et al., March'22]

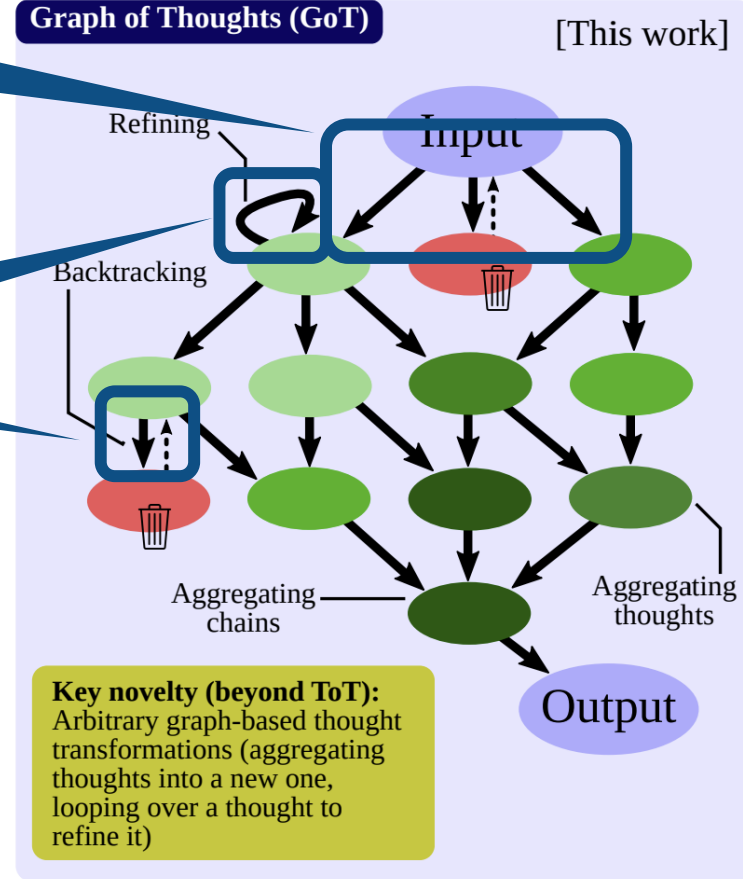
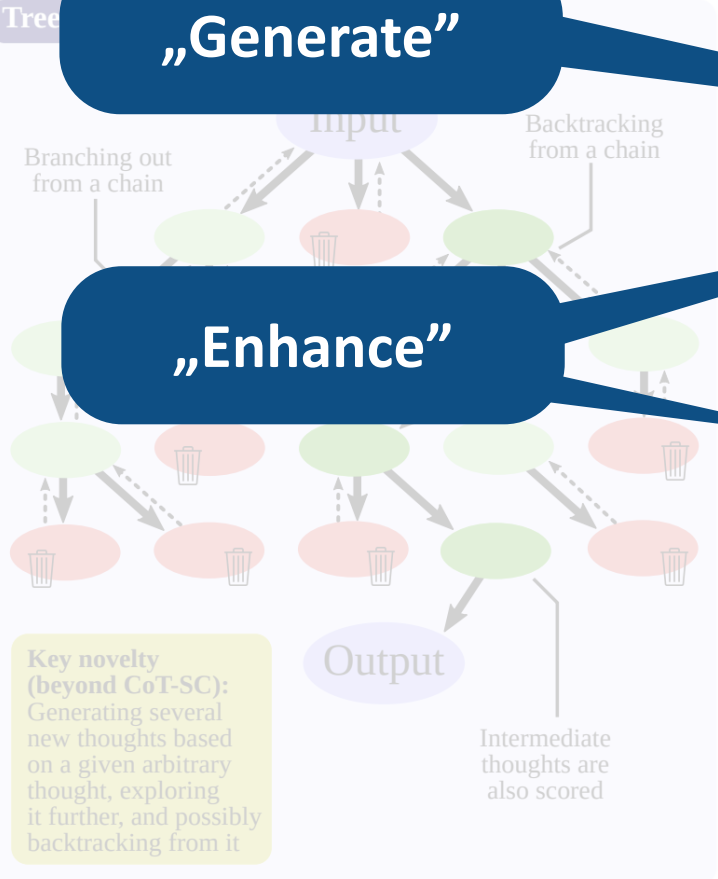
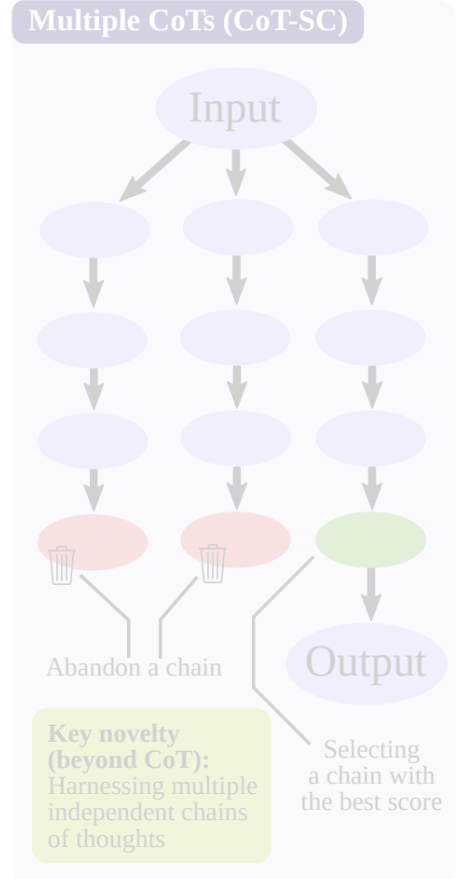
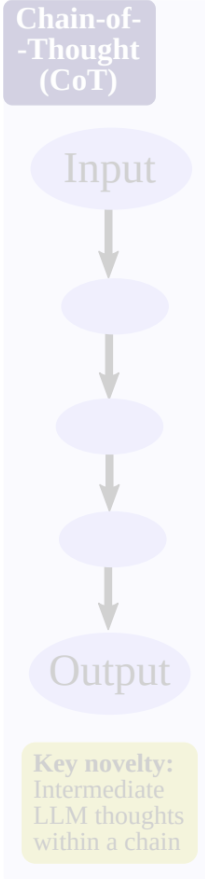
[Long et al., May'23]

[Yao et al., May'23]



Legend

- Thoughts:
 - Unscored (light blue oval)
 - Positive score (green oval)
 - Negative score (red oval)
- ↓ Dependencies between thoughts
- 🗑️ Abandon thought
- ↔️ Backtrack



Prompting Paradigms: Graph of Thoughts

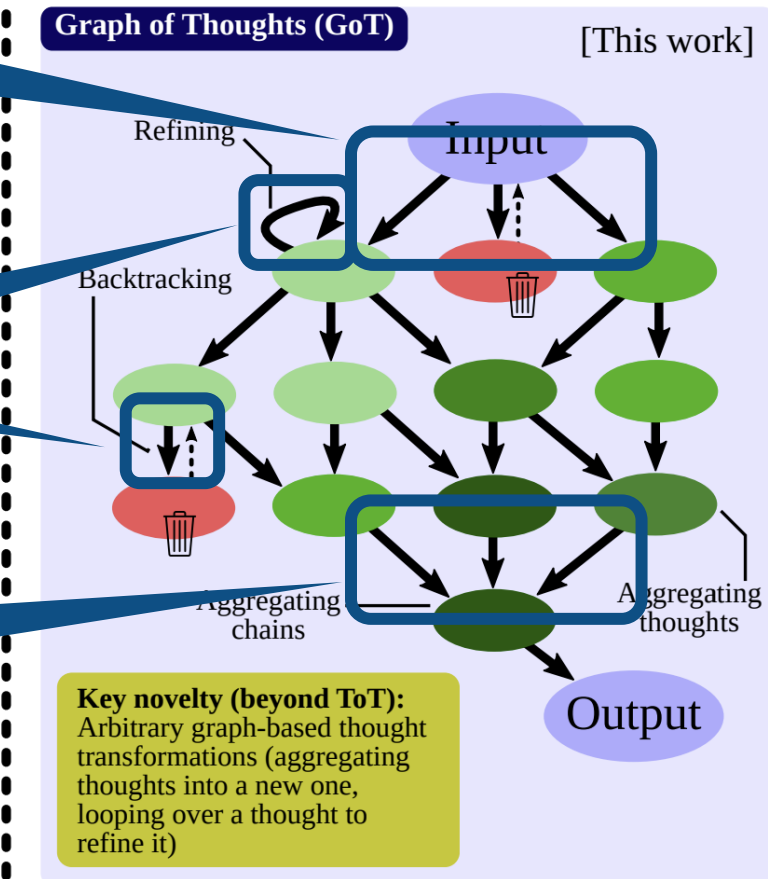
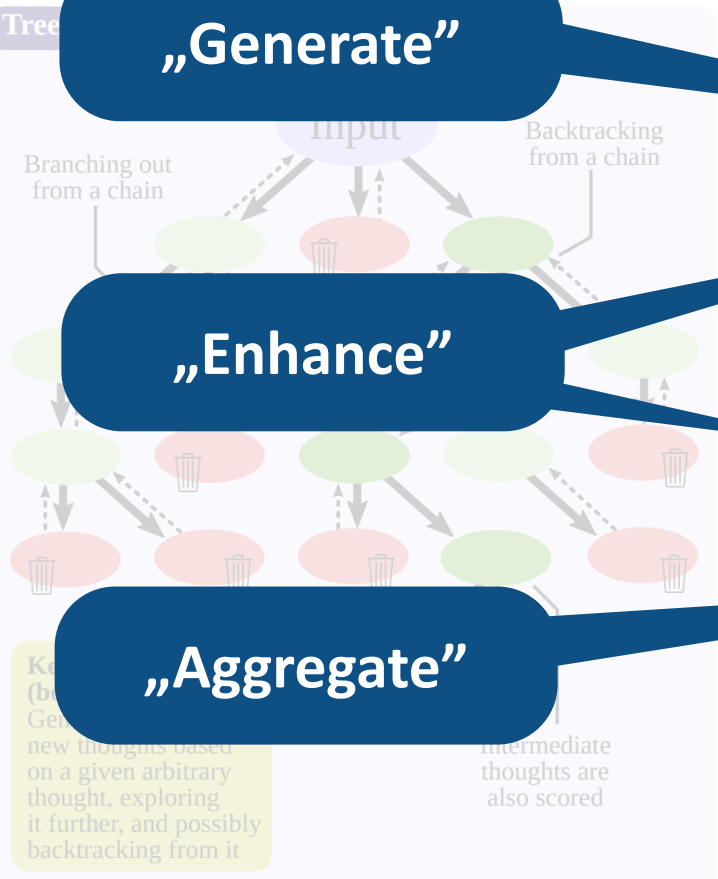
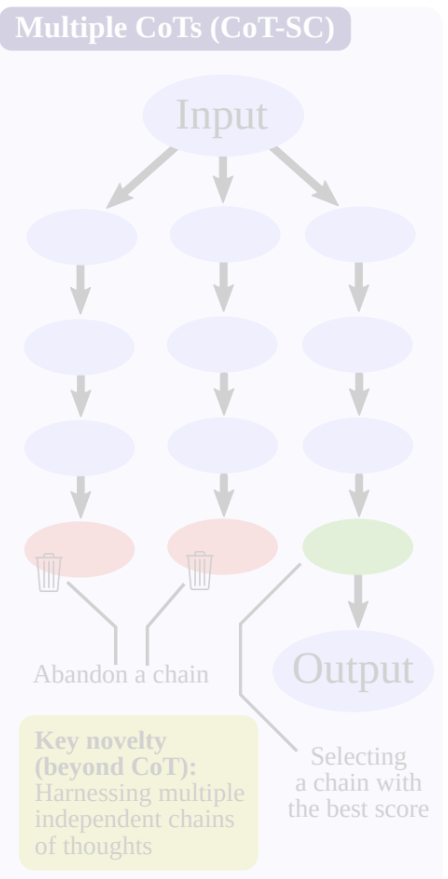
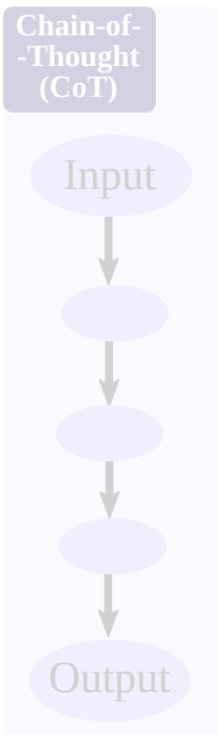
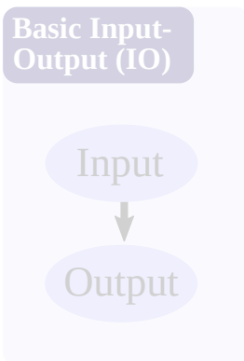
„Thought transformations“

[Wei et al., Jan'22]

[Wang et al., March'22]

[Long et al., May'23]

[Yao et al., May'23]



Legend

- Thoughts:
 - Unscored (light blue)
 - Positive score (green)
 - Negative score (red)
- ↓ Dependencies between thoughts
- 🗑️ Abandon thought
- ↔️ Backtrack

Key novelty (beyond CoT): Intermediate LLM thoughts within a chain

Key novelty (beyond CoT): Harnessing multiple independent chains of thoughts

Key novelty (beyond ToT): Generating new thoughts based on a given arbitrary thought, exploring it further, and possibly backtracking from it

Key novelty (beyond ToT): Arbitrary graph-based thought transformations (aggregating thoughts into a new one, looping over a thought to refine it)

„Generate“

„Enhance“

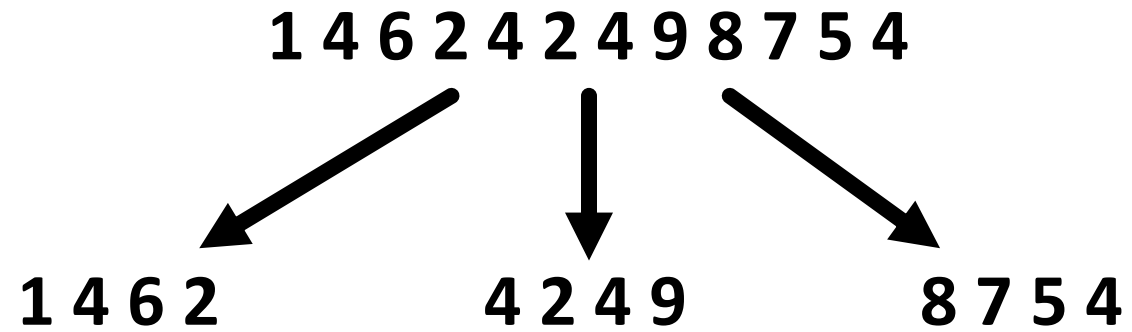
„Aggregate“

Thought Transformations for Sorting

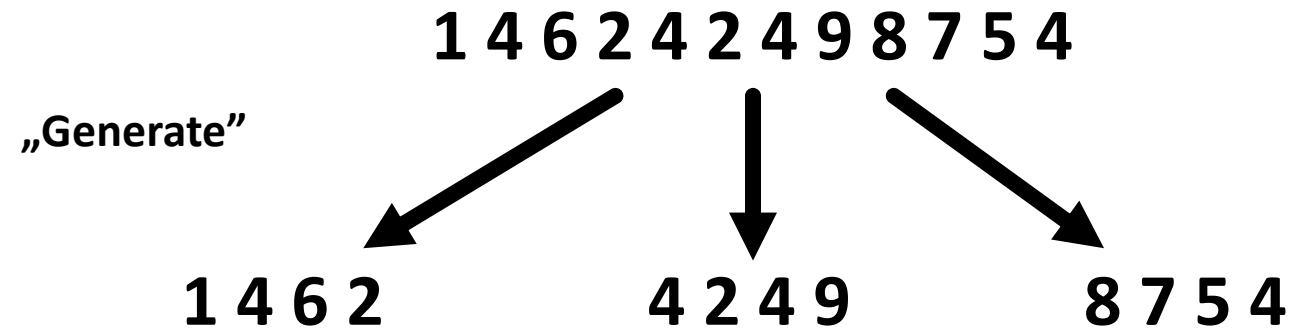
Thought Transformations for Sorting

1 4 6 2 4 2 4 9 8 7 5 4

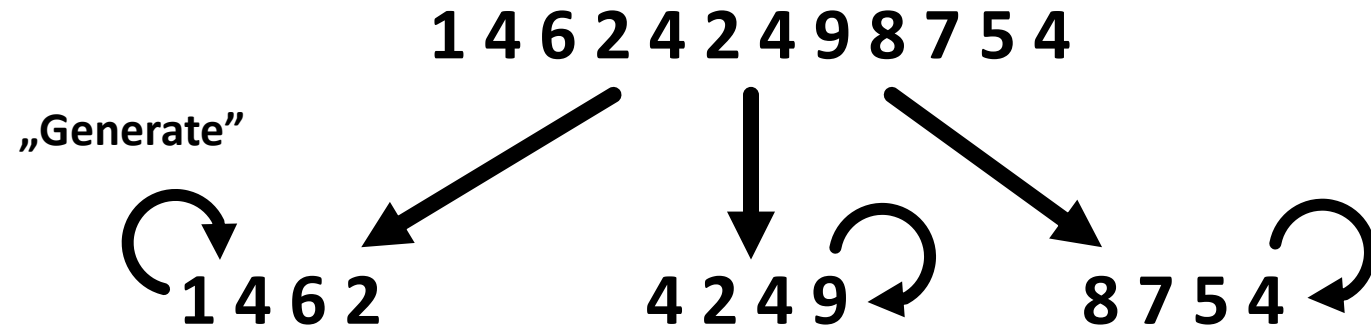
Thought Transformations for Sorting



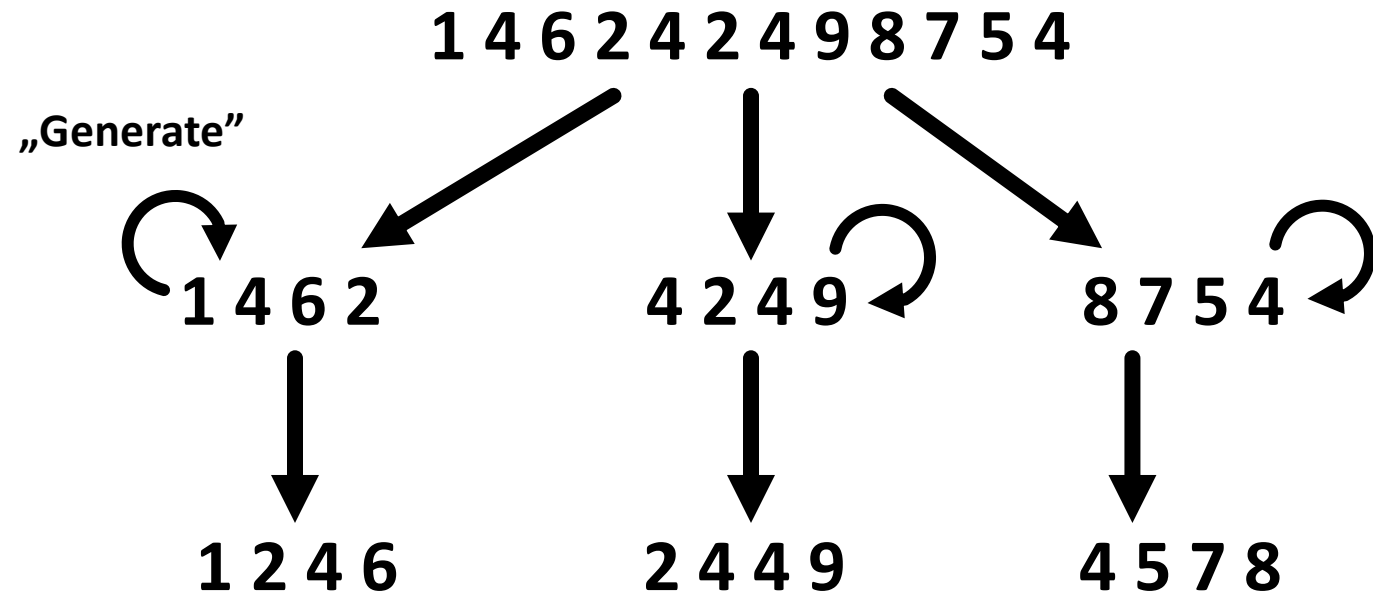
Thought Transformations for Sorting



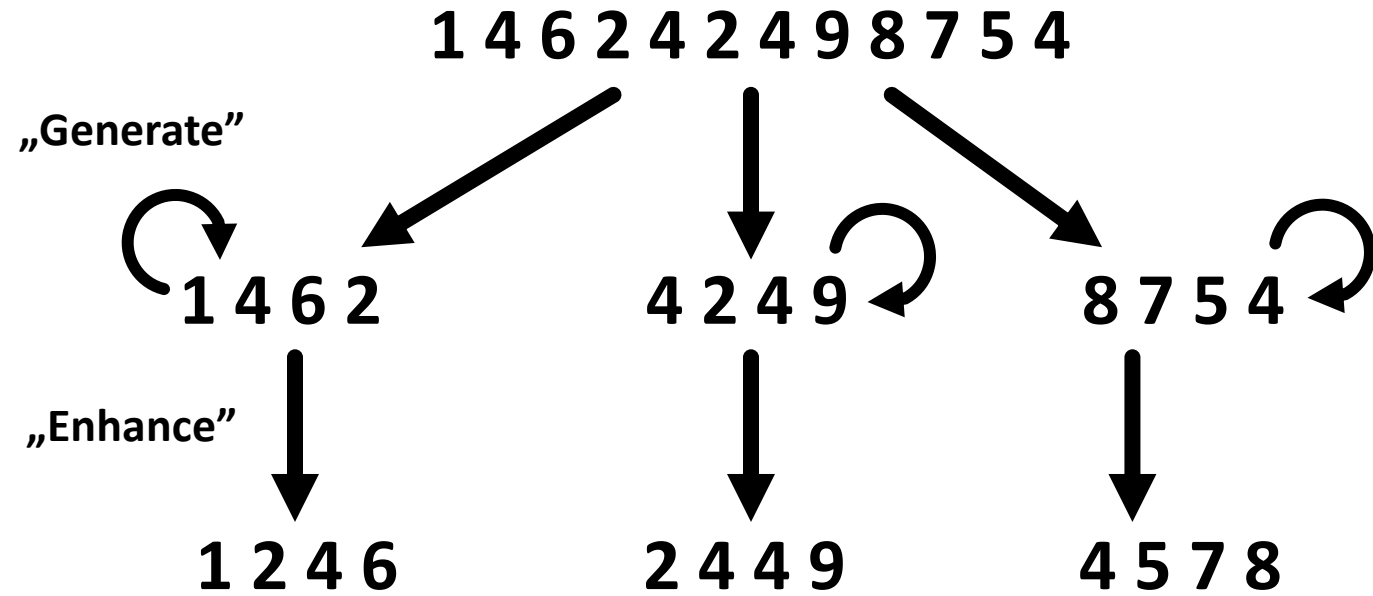
Thought Transformations for Sorting



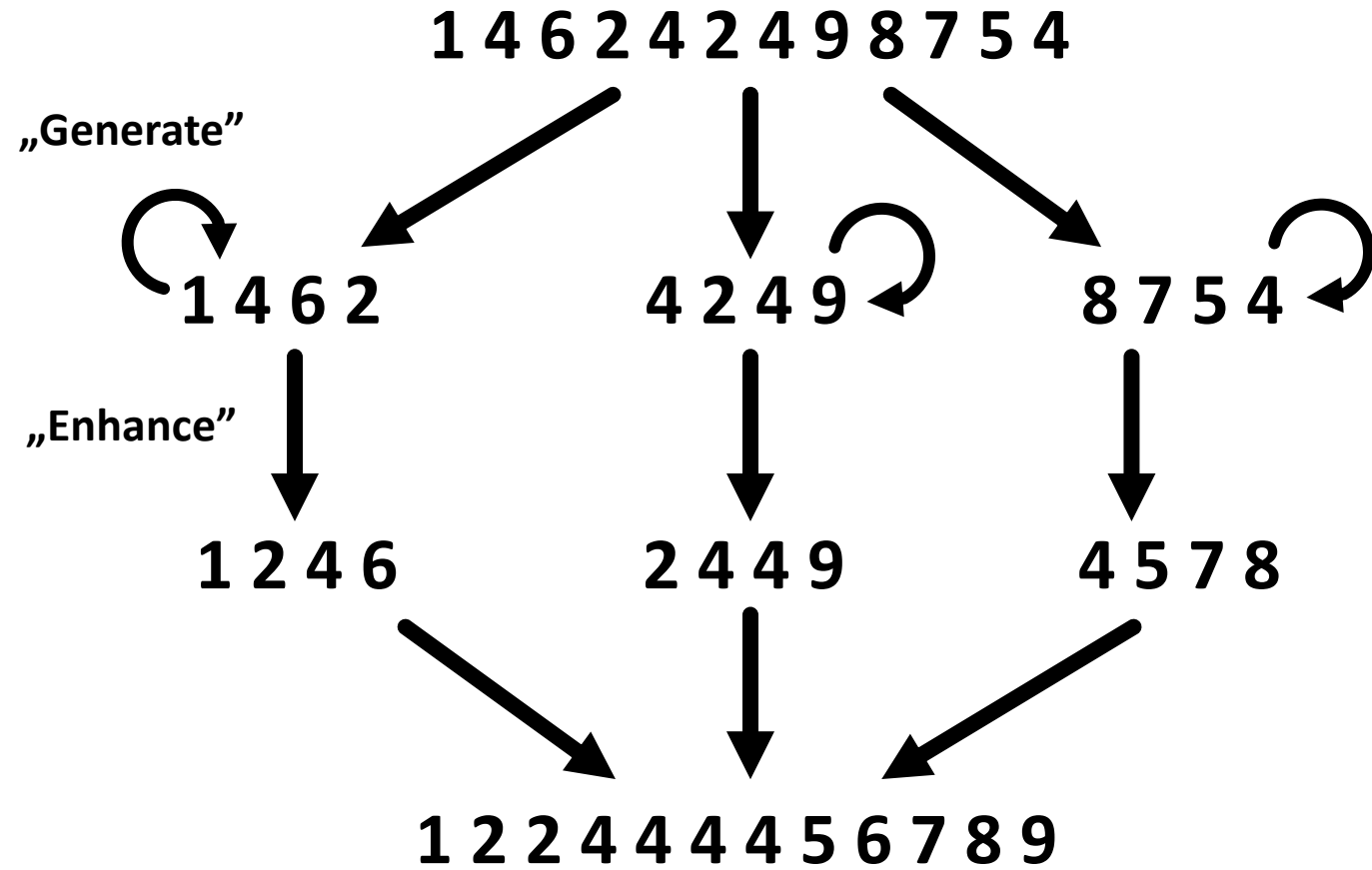
Thought Transformations for Sorting



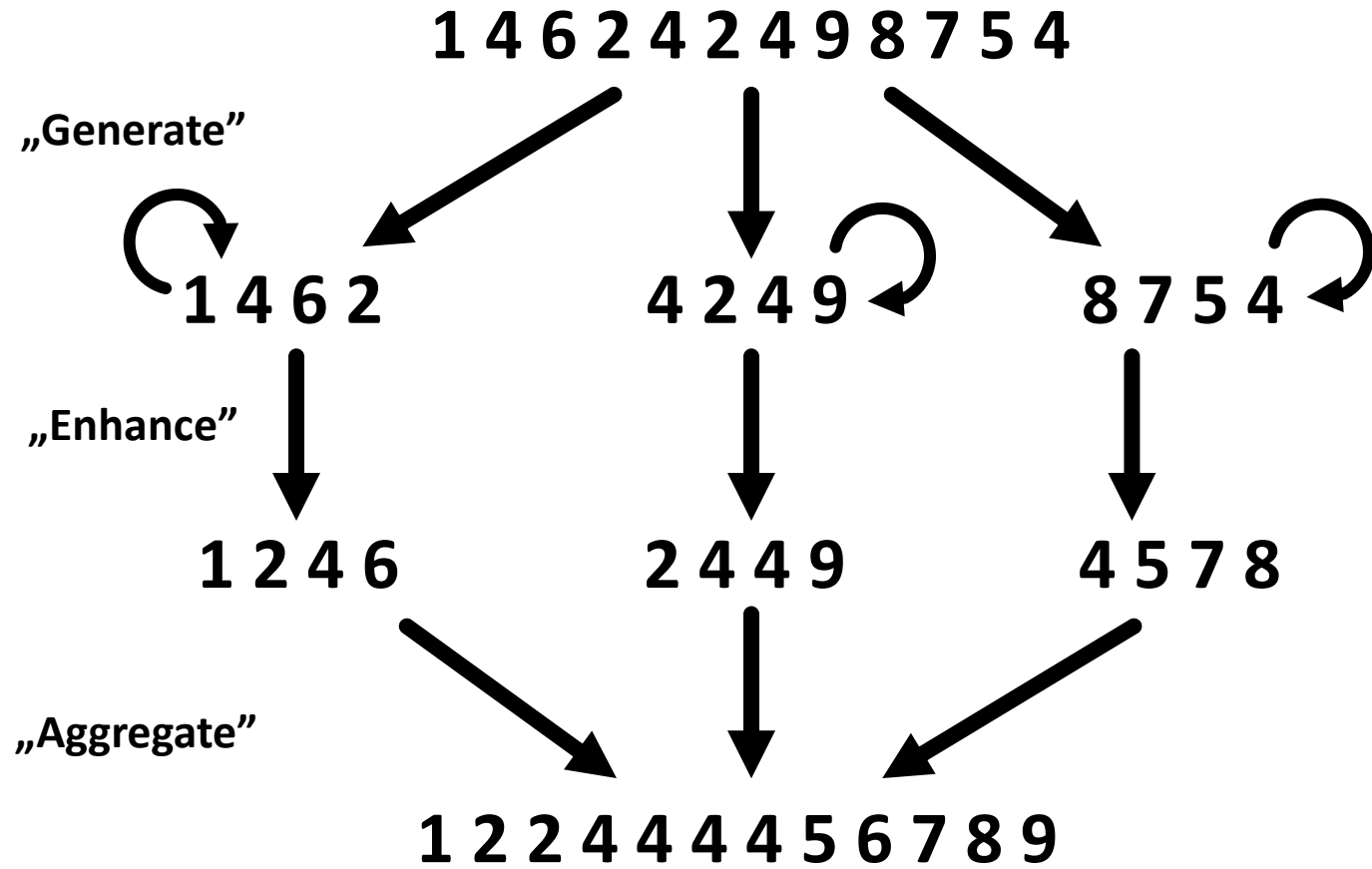
Thought Transformations for Sorting



Thought Transformations for Sorting

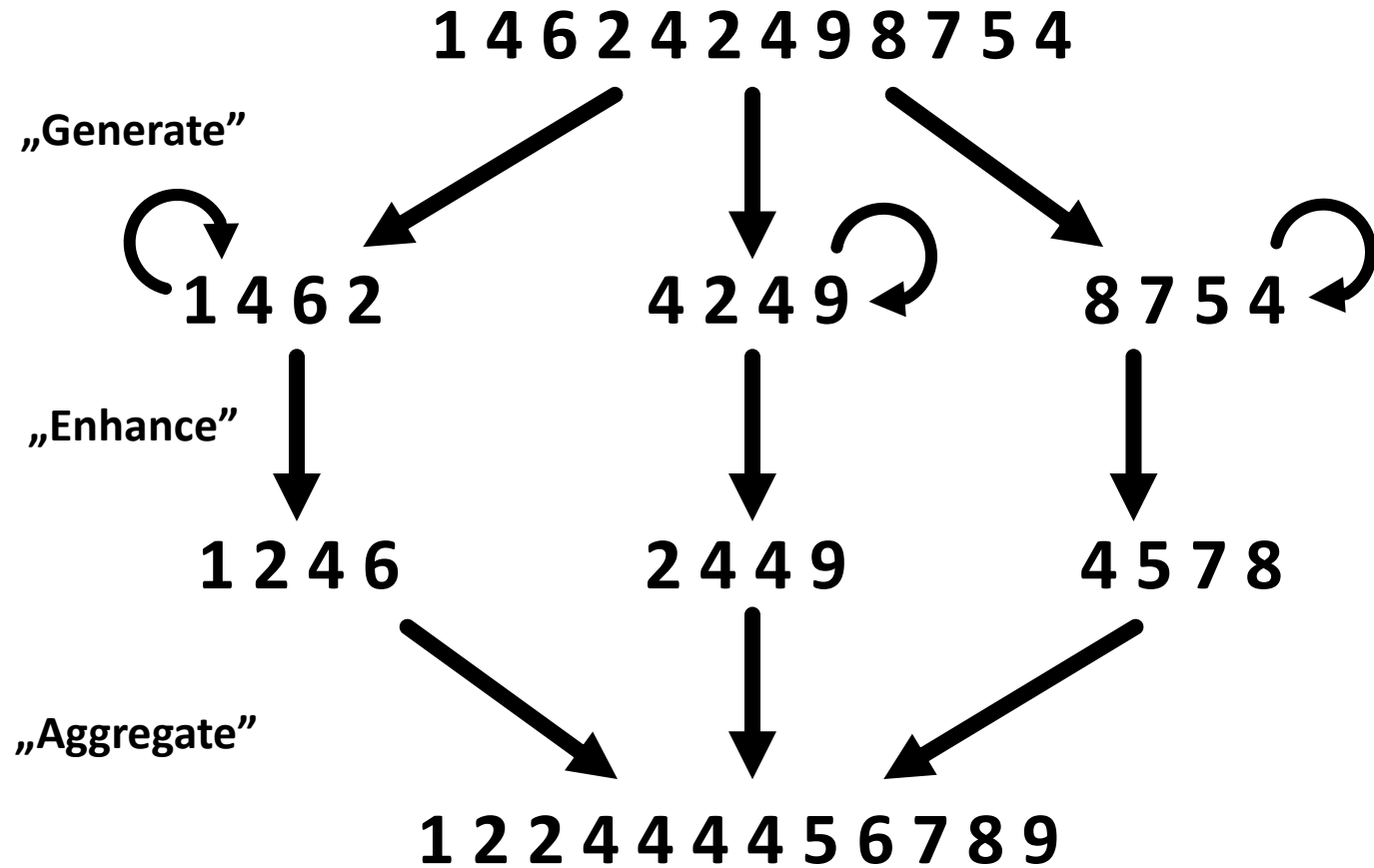


Thought Transformations for Sorting



Thought Transformations for Sorting

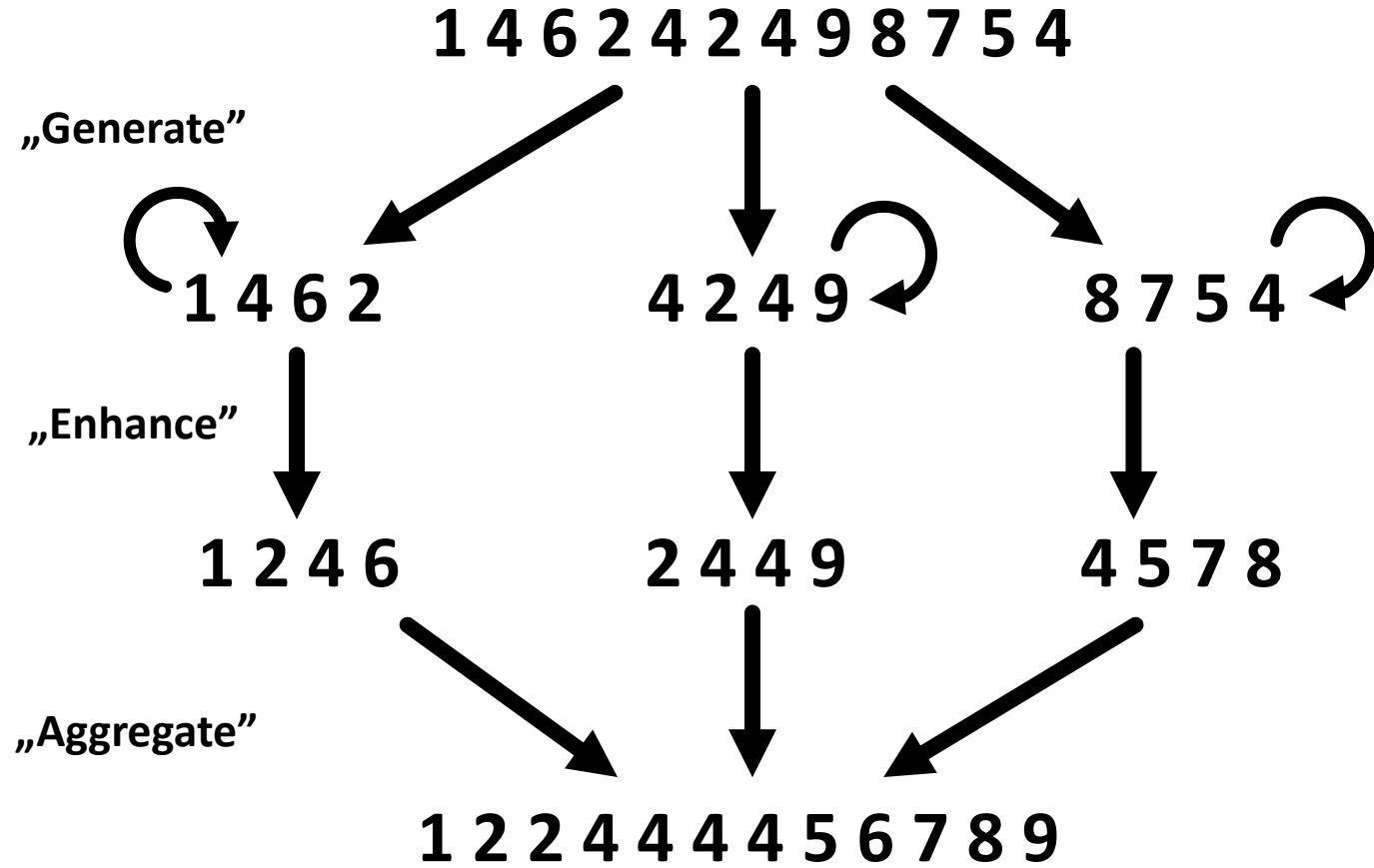
Why sorting? Because it is a fundamental problem in CS, and it still does pose a great challenge for all other baselines



Thought Transformations for Sorting

Why sorting? Because it is a fundamental problem in CS, and it still does pose a great challenge for all other baselines

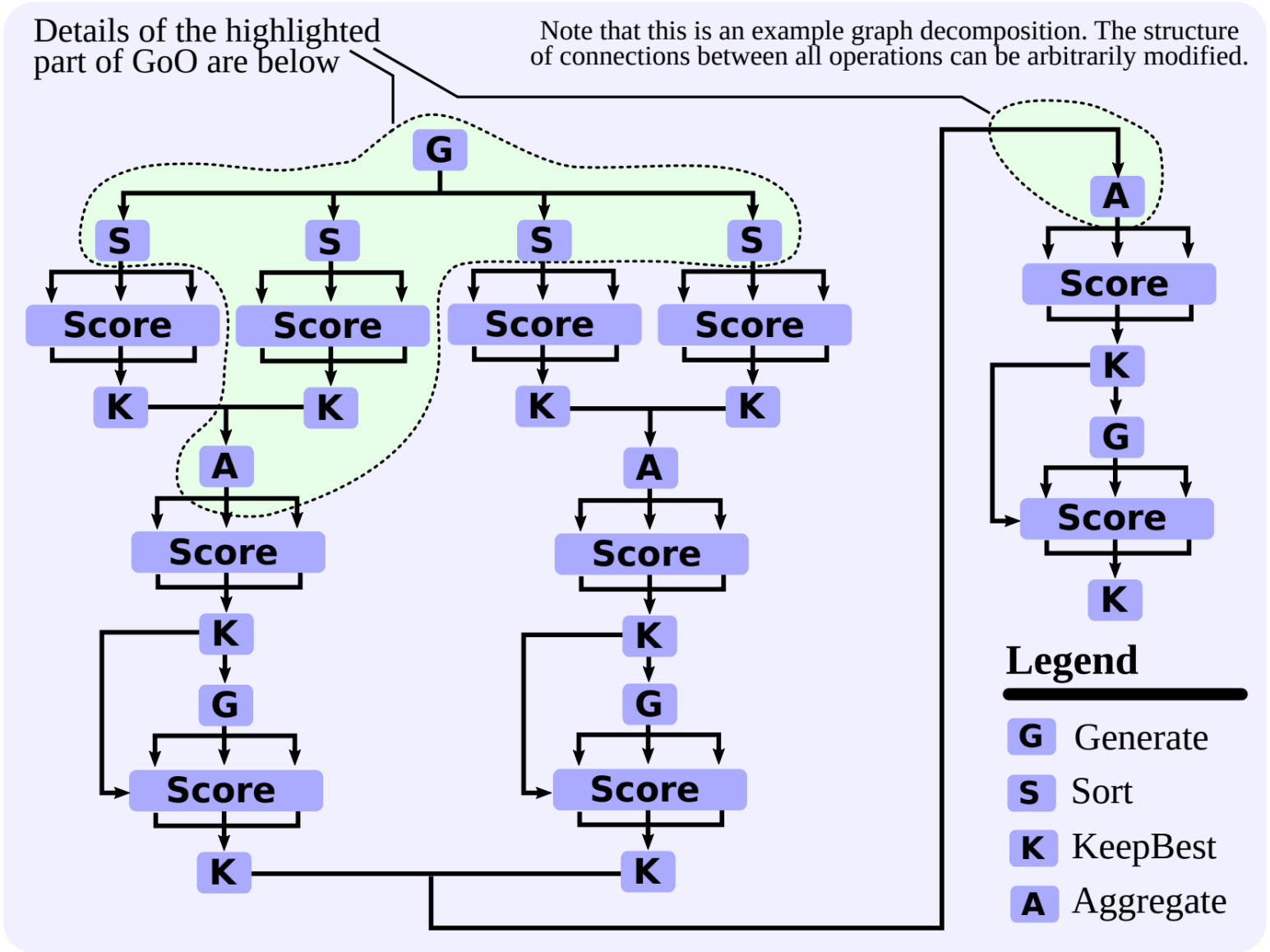
This is a small example; for real use cases, the size is much larger, and the graph gets more complex



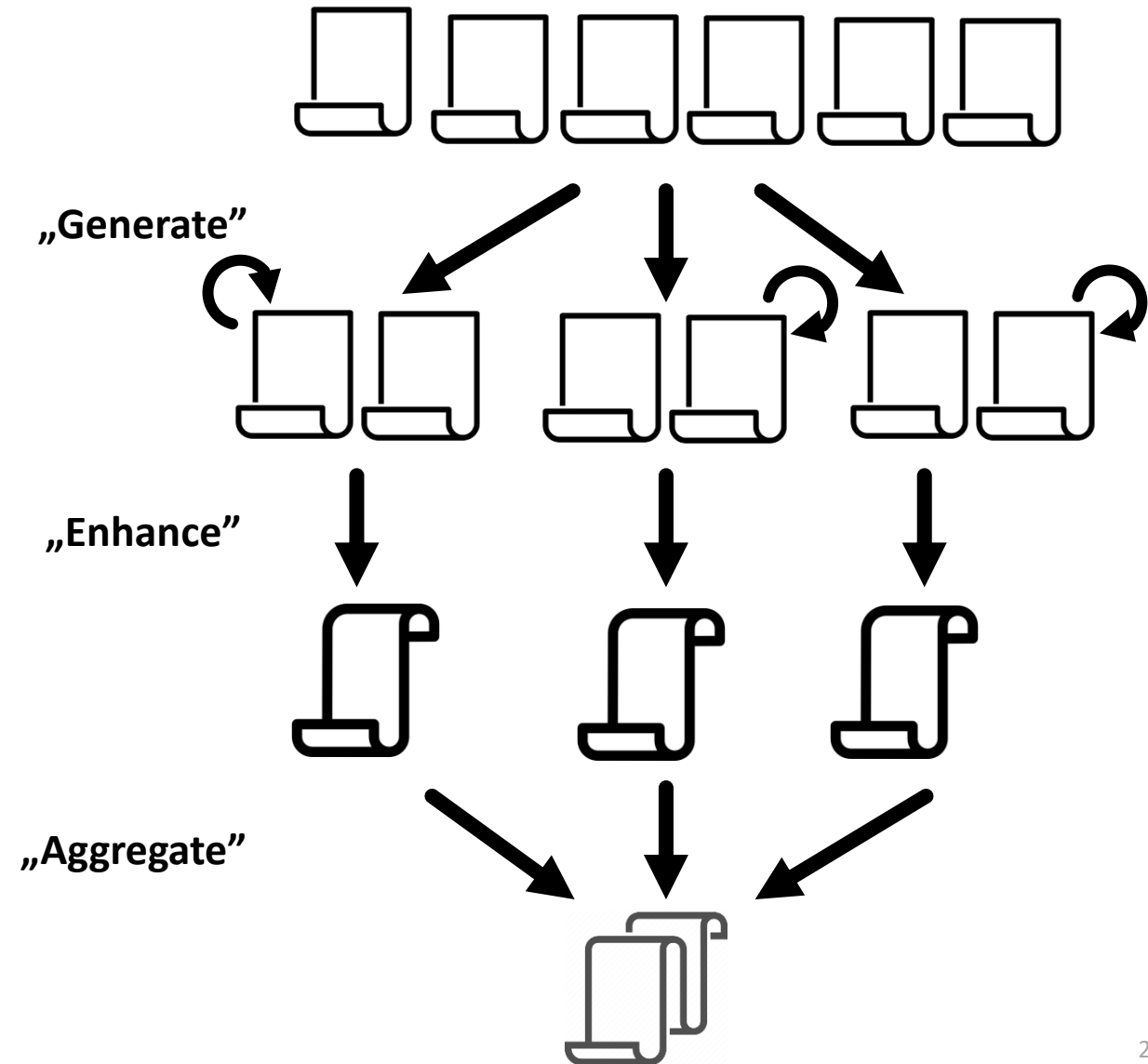
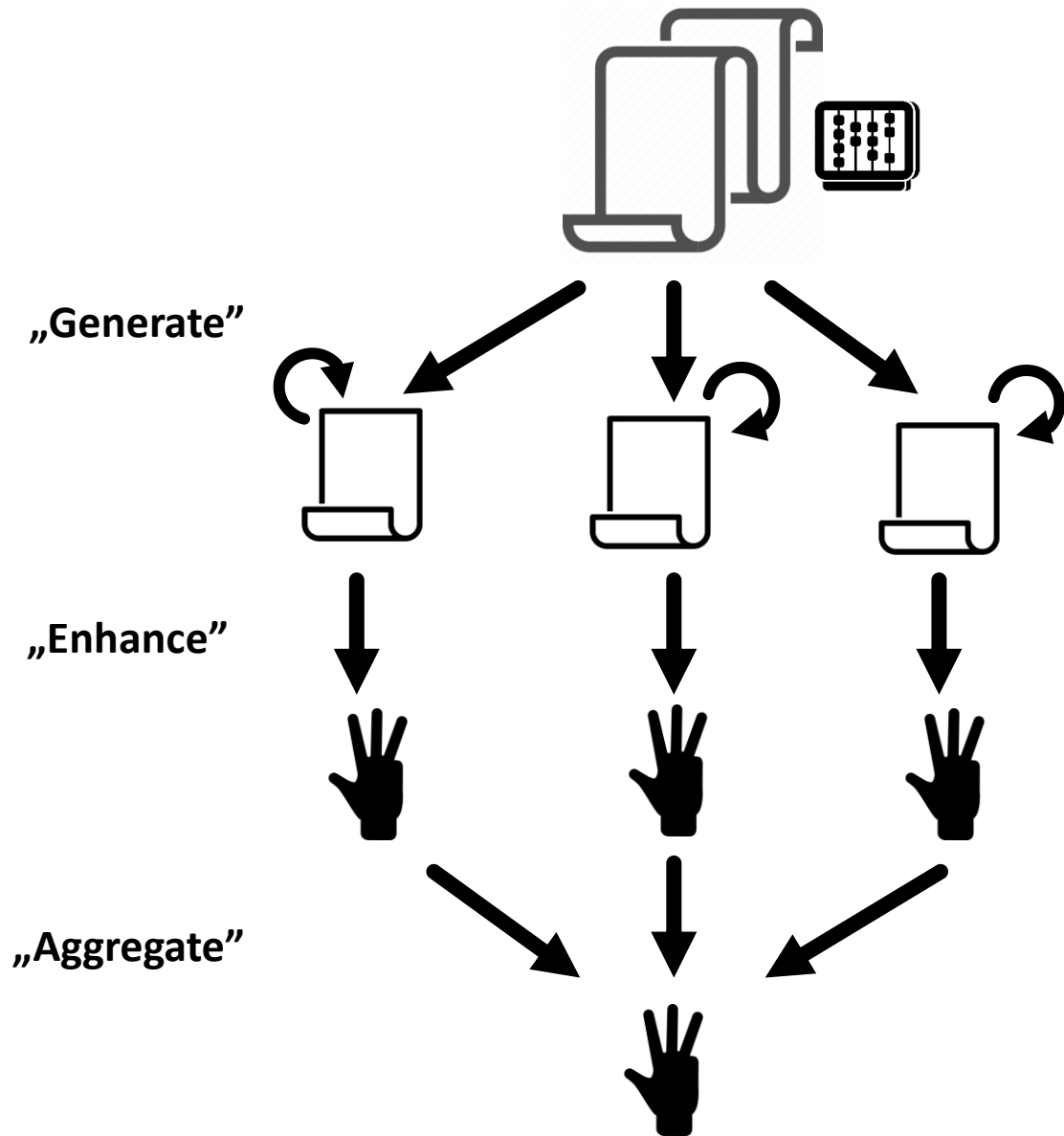
Thought Transformations for Sorting

Why sorting? Because it is a fundamental problem in CS, and it still does pose a great challenge for all other baselines

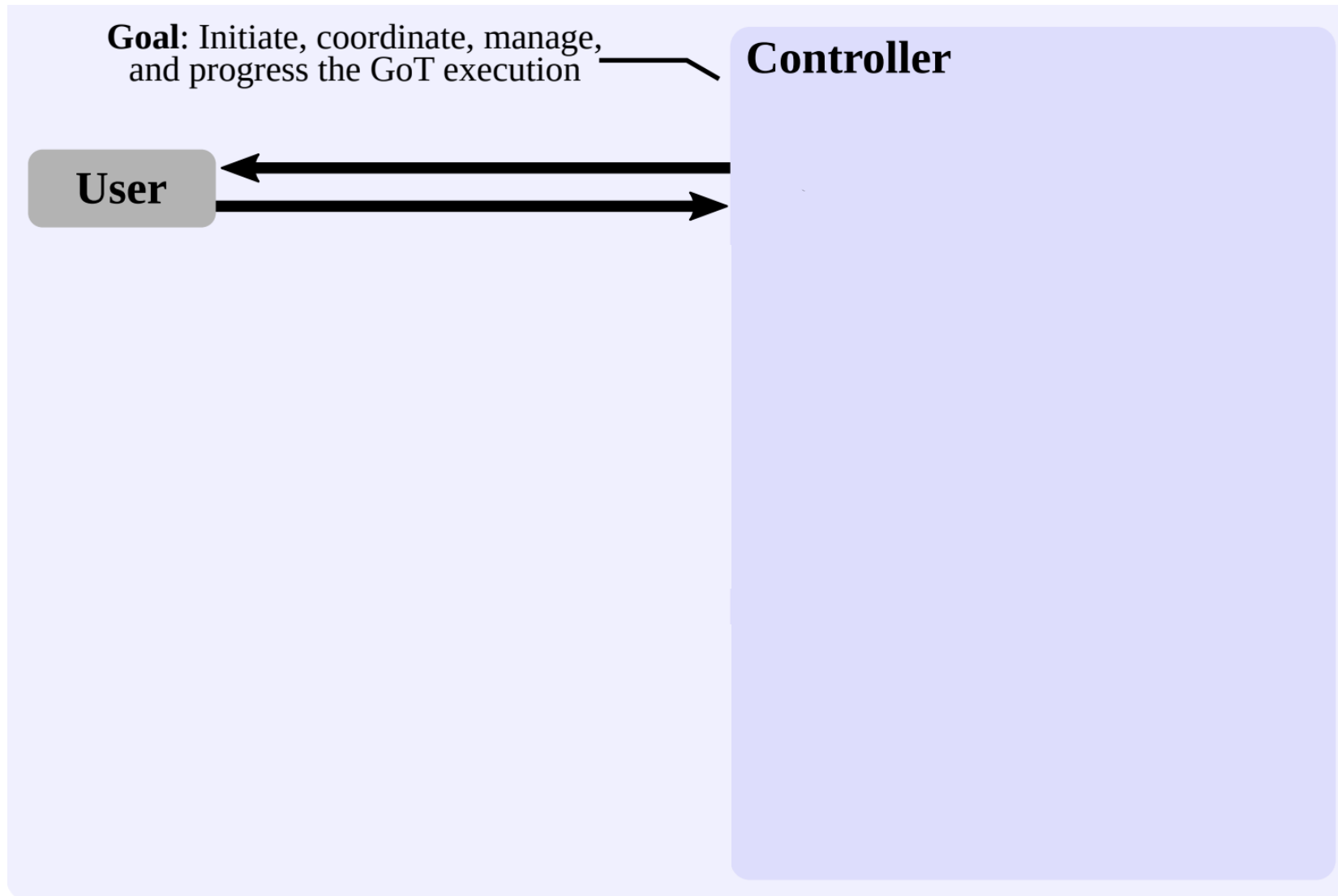
This is a small example; for real use cases, the size is much larger, and the graph gets more complex



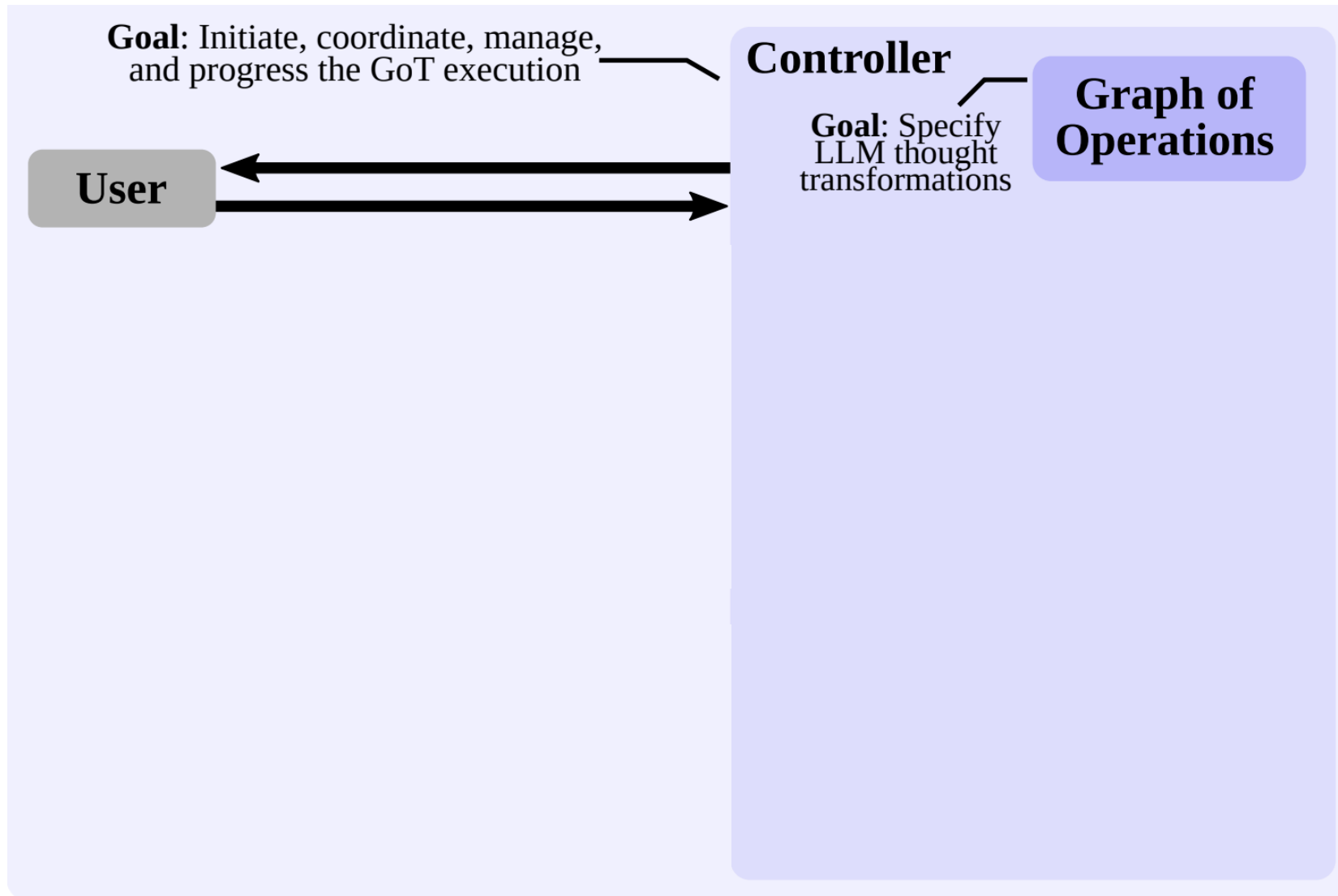
Other Examples: Keyword Counting & NDA merging



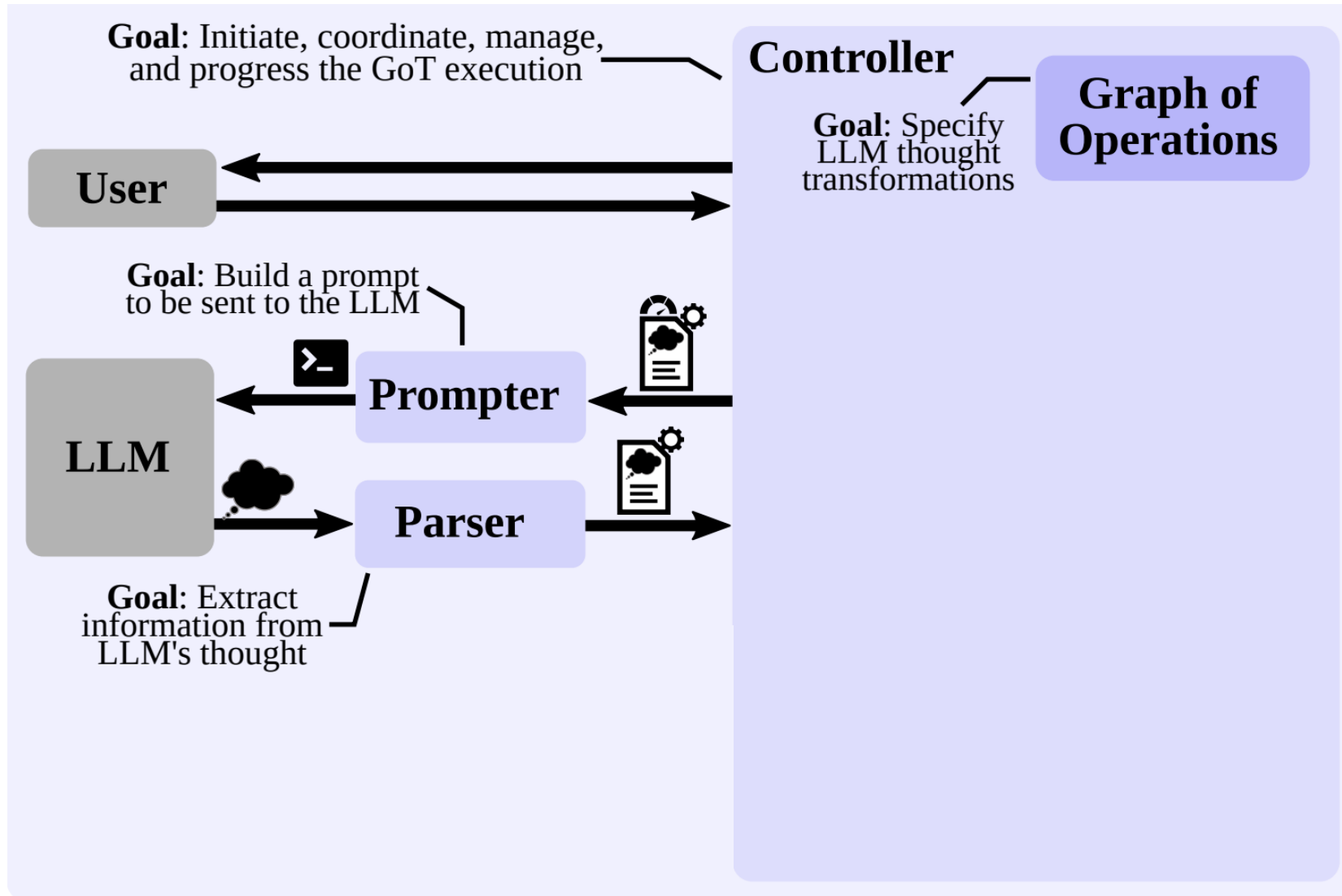
Graph of Thoughts: Architecture & Design



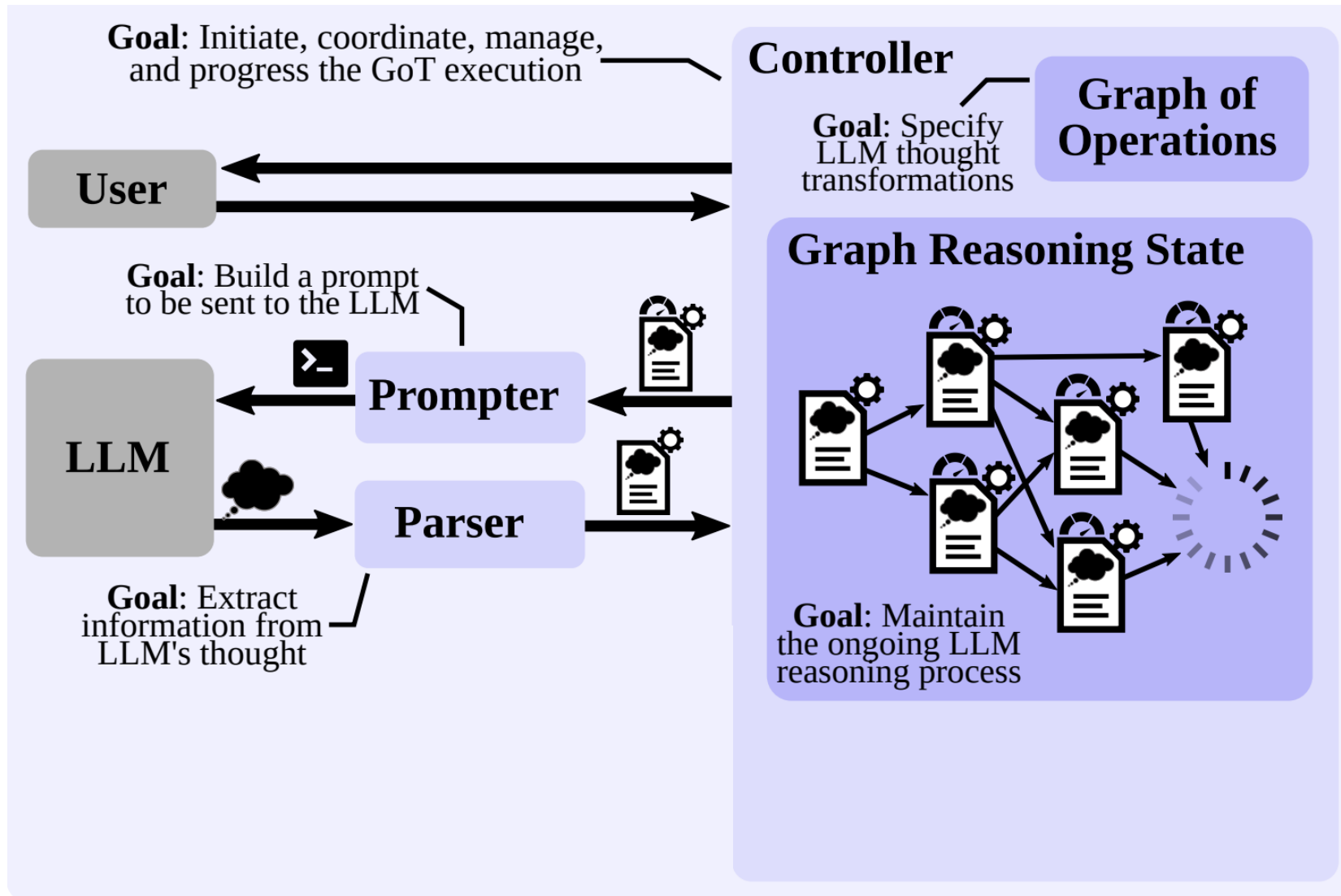
Graph of Thoughts: Architecture & Design



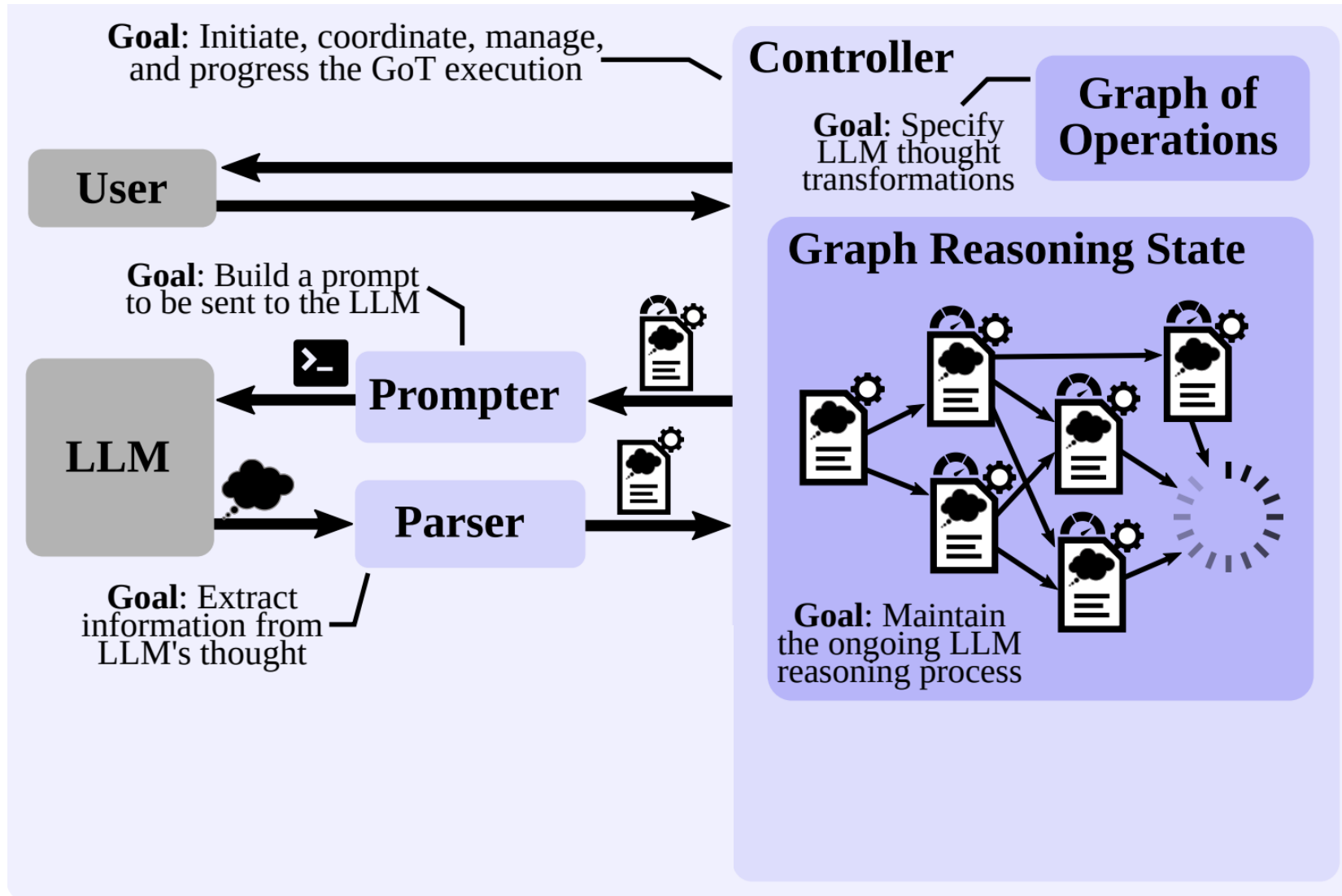
Graph of Thoughts: Architecture & Design



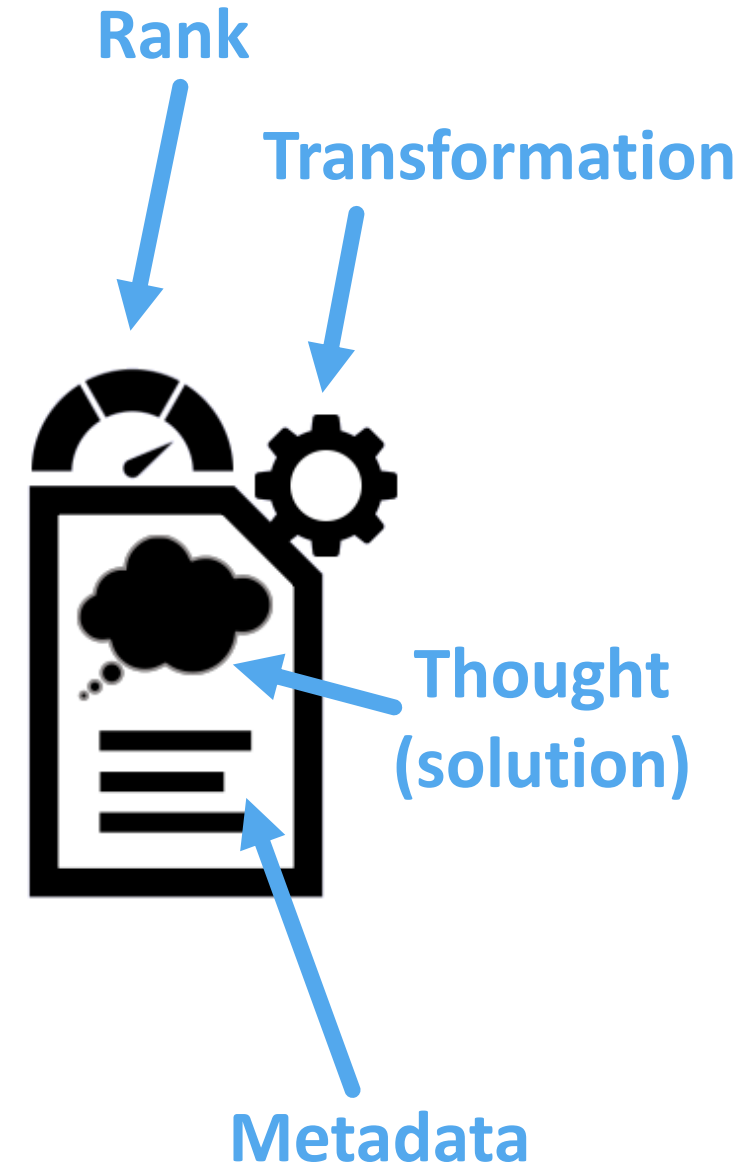
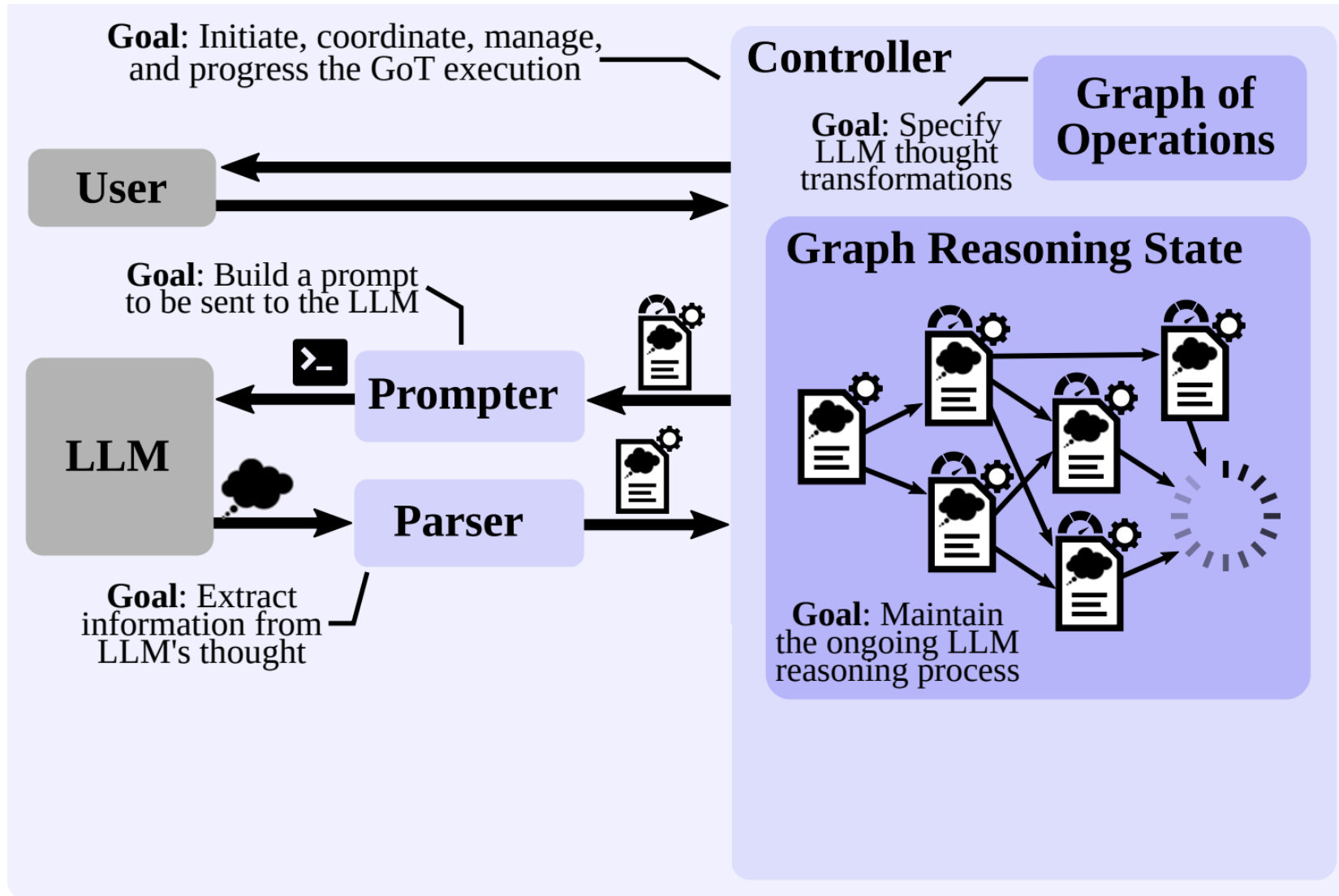
Graph of Thoughts: Architecture & Design



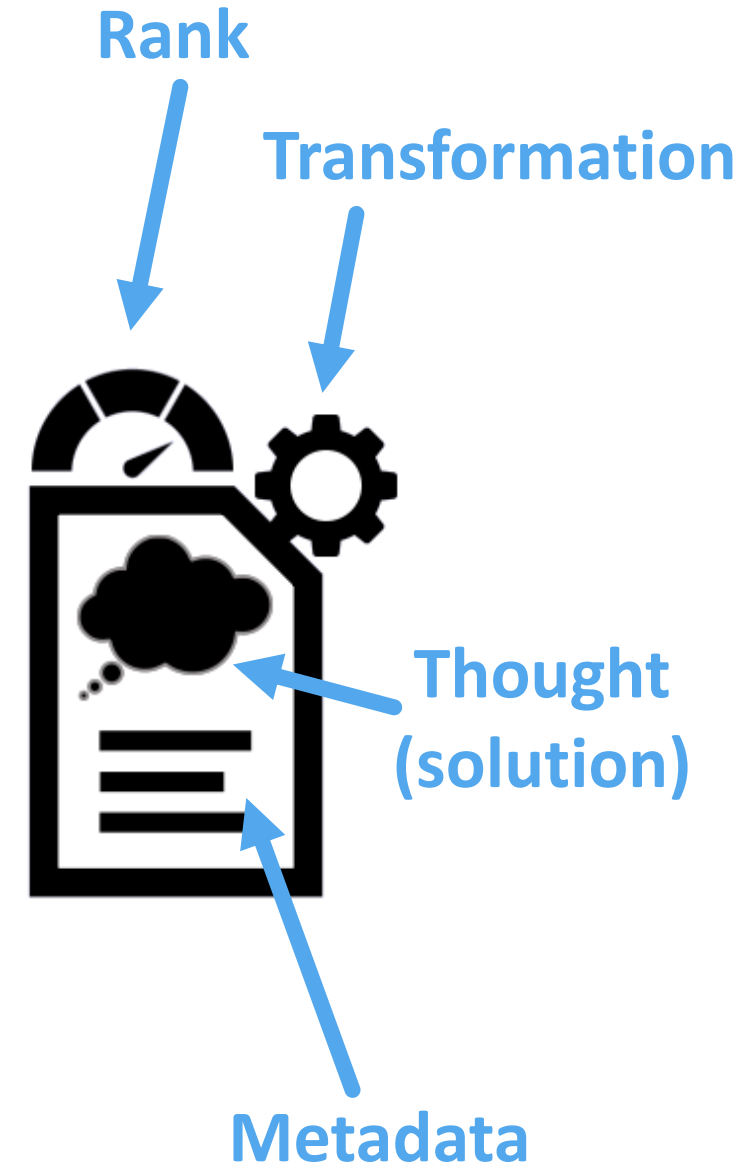
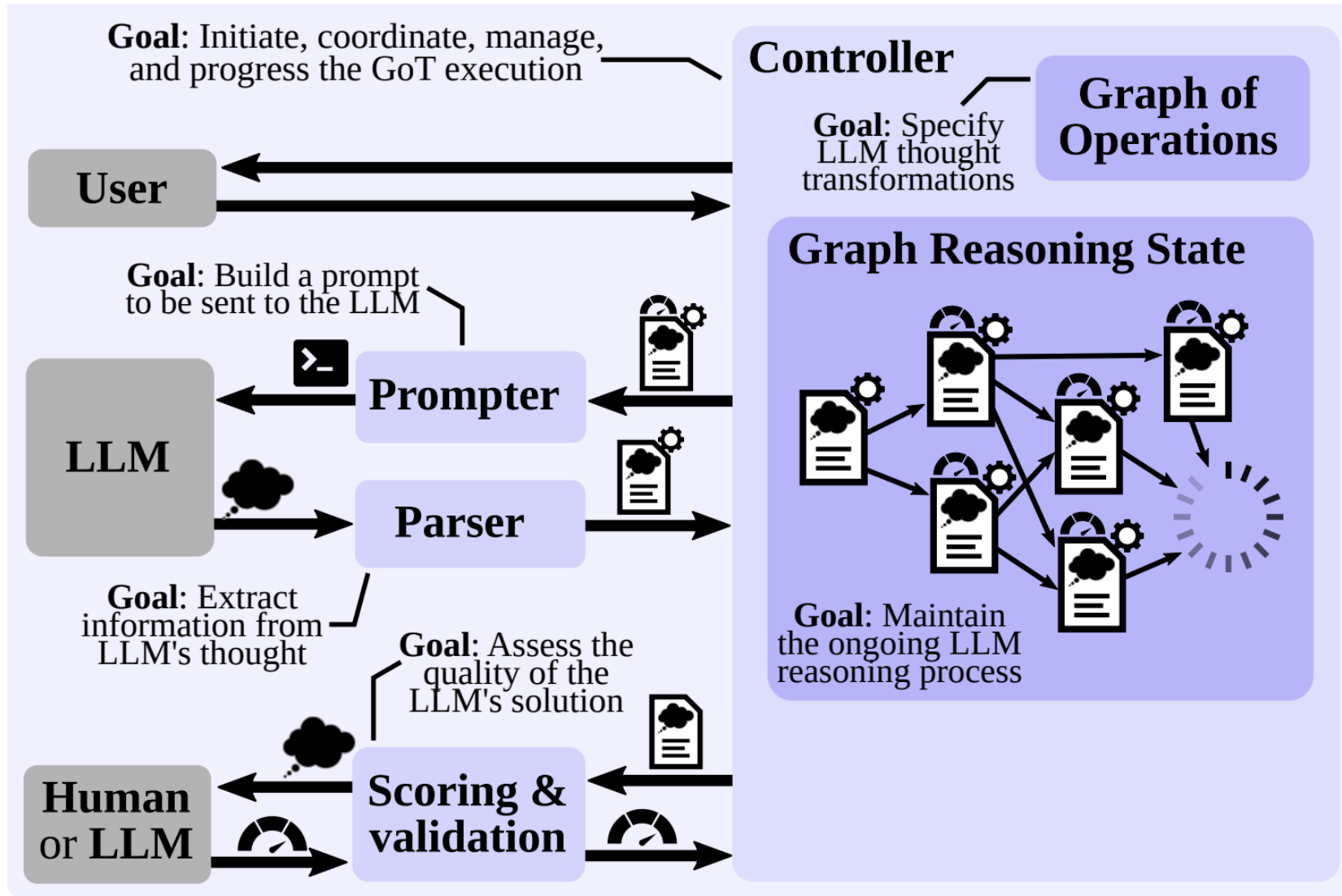
Graph of Thoughts: Architecture & Design



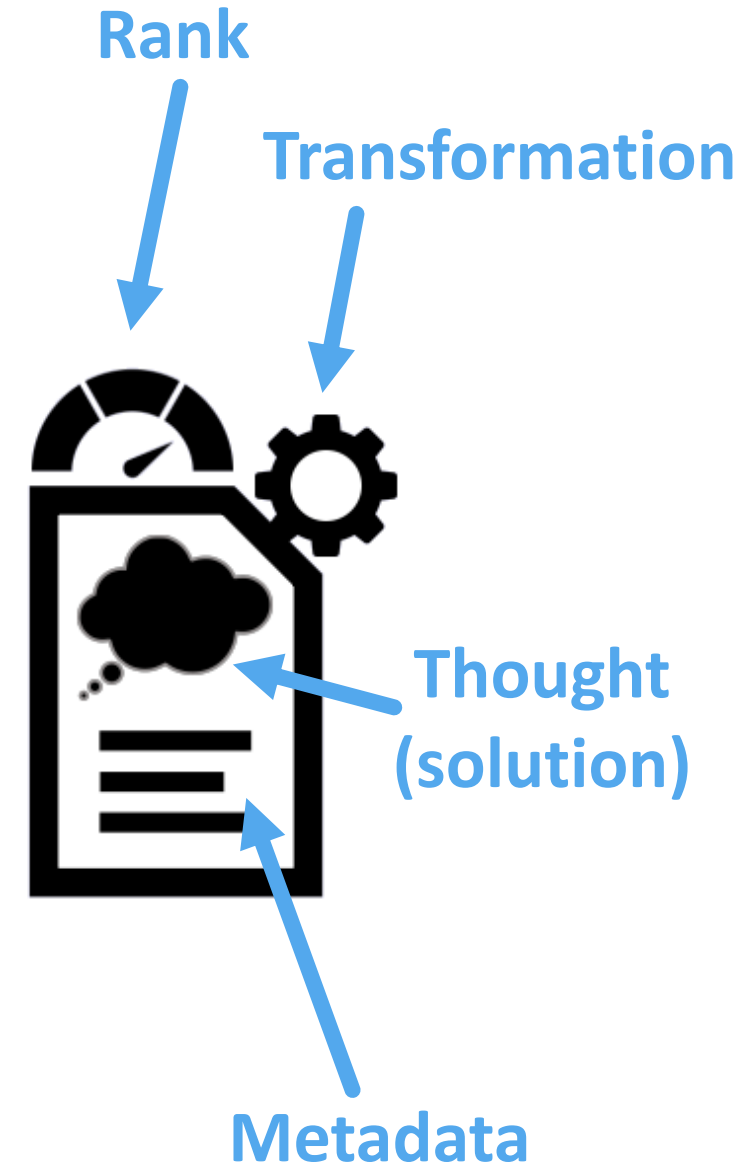
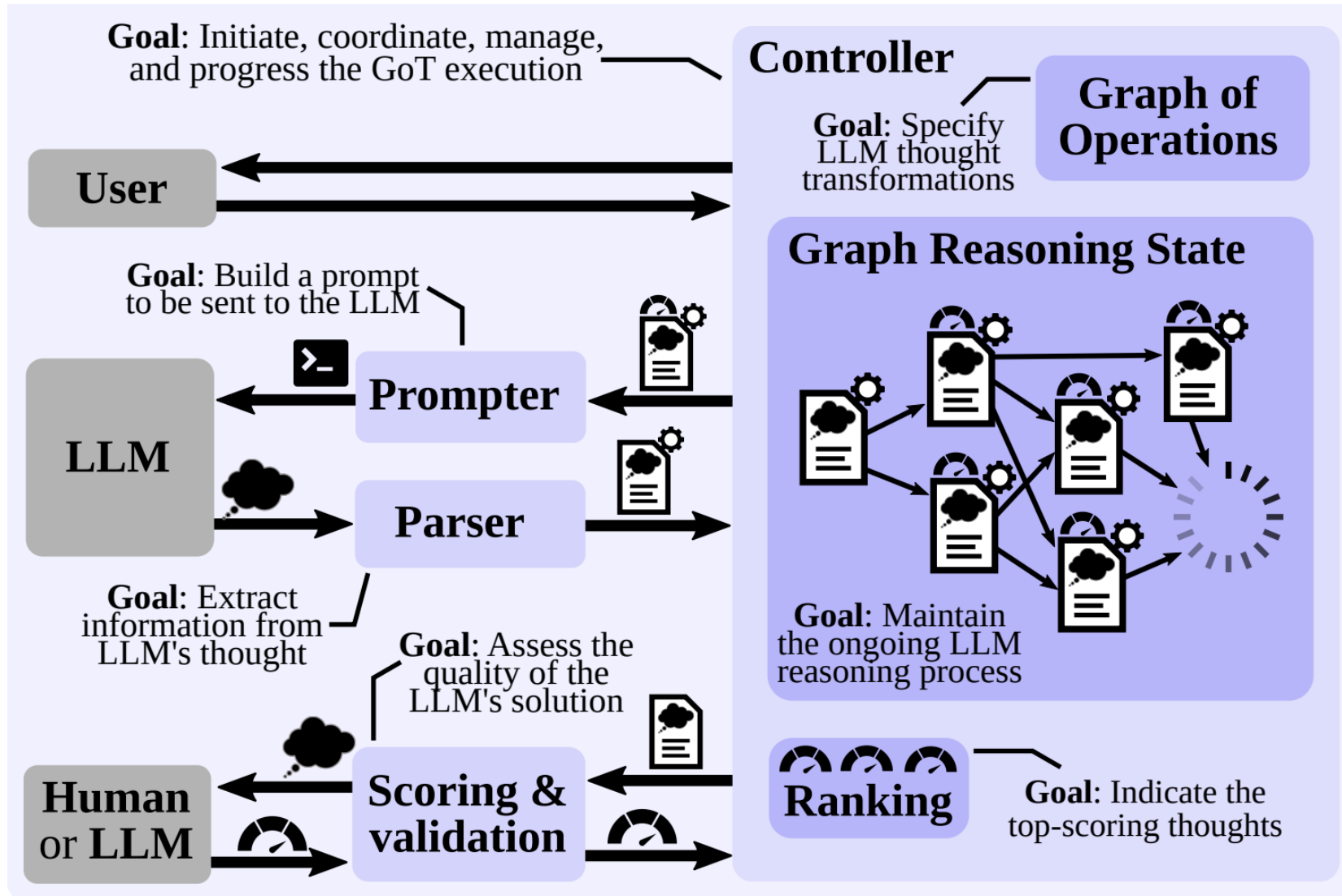
Graph of Thoughts: Architecture & Design



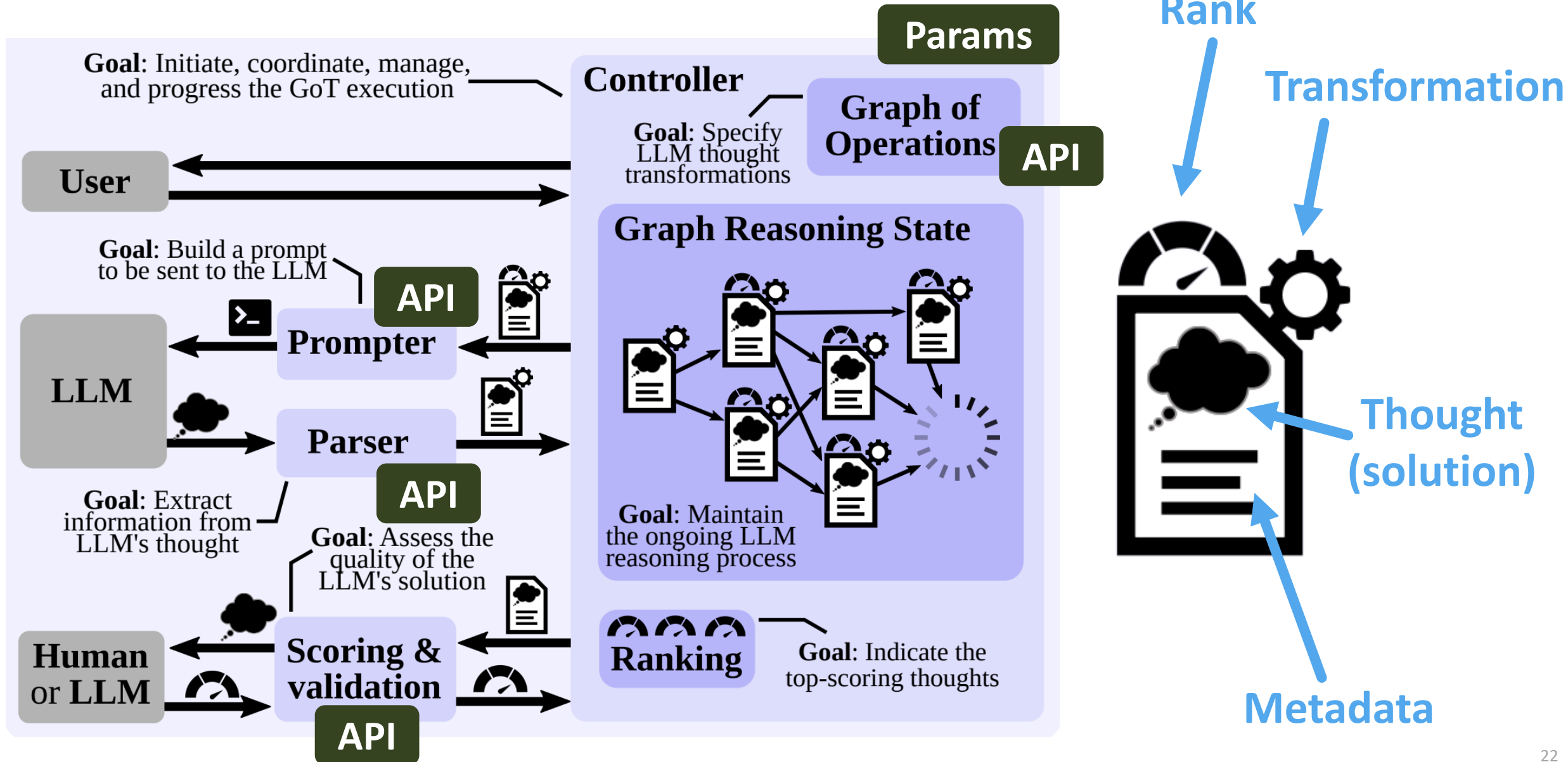
Graph of Thoughts: Architecture & Design



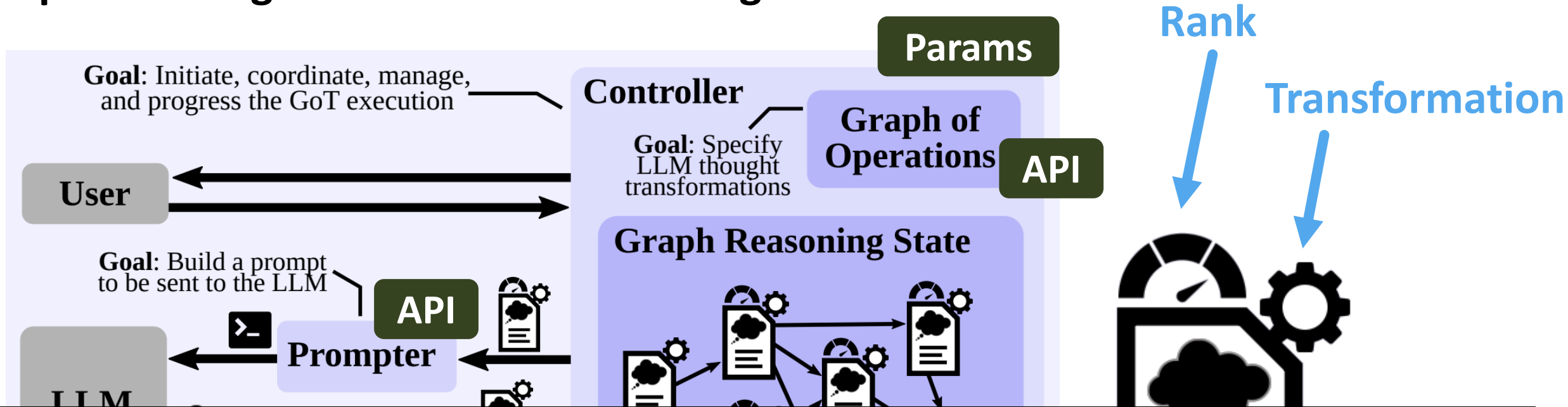
Graph of Thoughts: Architecture & Design



Graph of Thoughts: Architecture & Design



Graph of Thoughts: Architecture & Design

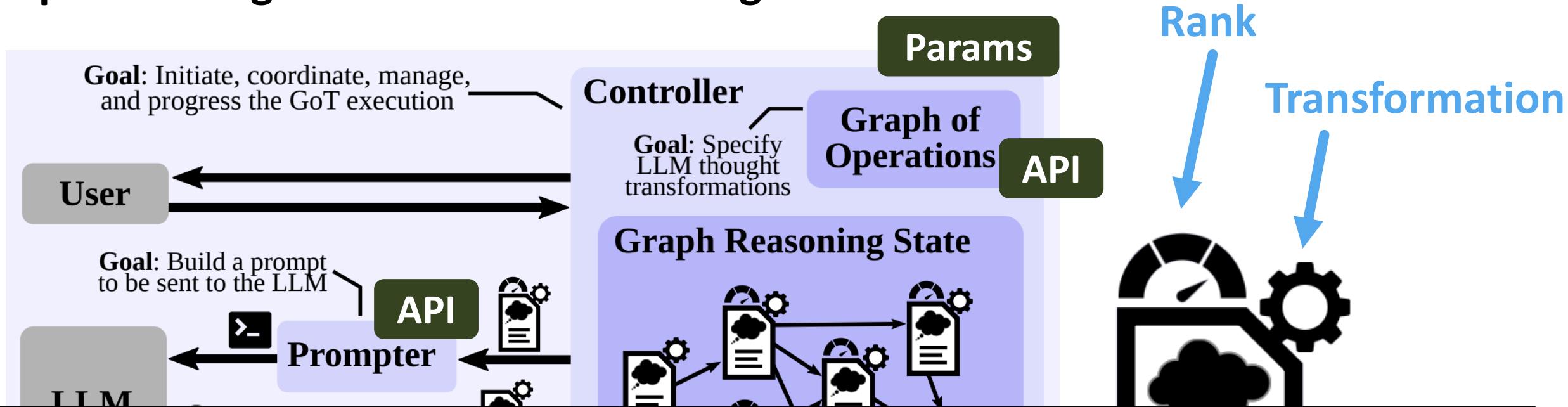


Graph of Thoughts: Solving Elaborate Problems with Large Language Models

Maciej Besta^{1*}, Nils Blach^{1*}, Ales Kubicek¹, Robert Gerstenberger¹,
Lukas Gianinazzi¹, Joanna Gajda², Tomasz Lehmann², Michał Podstawski³,
Hubert Niewiadomski², Piotr Nyczyk², Torsten Hoefler¹

¹ETH Zurich, ²Cledar, ³Warsaw University of Technology
bestam@inf.ethz.ch, nils.blach@inf.ethz.ch, htor@inf.ethz.ch

Graph of Thoughts: Architecture & Design



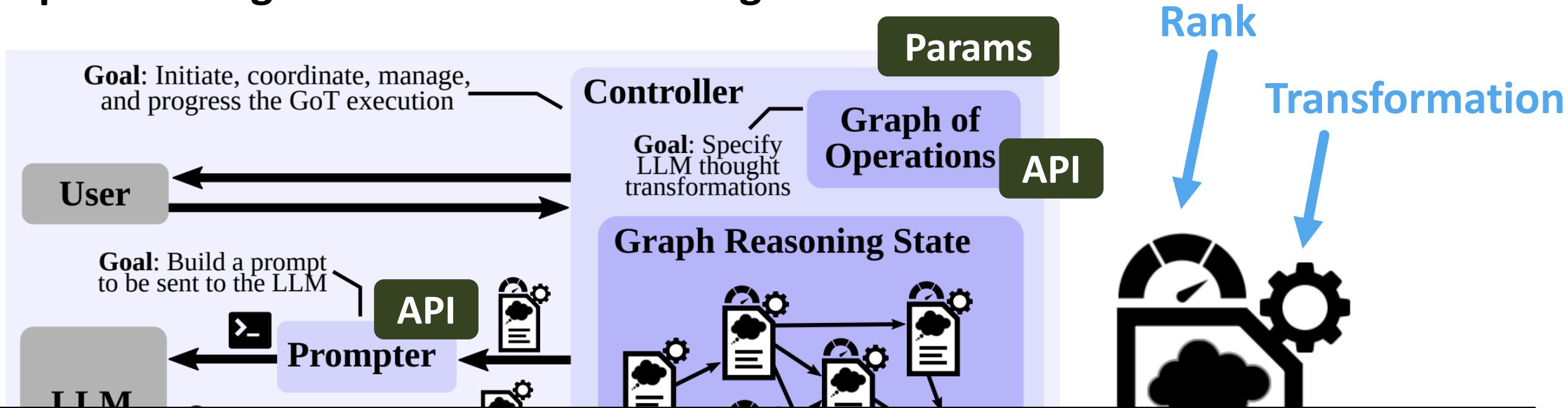
<https://github.com/spcl/graph-of-thoughts>, @AAAI'24

Graph of Thoughts: Solving Elaborate Problems with Large Language Models

Maciej Besta^{1*}, Nils Blach^{1*}, Ales Kubicek¹, Robert Gerstenberger¹,
Lukas Gianinazzi¹, Joanna Gajda², Tomasz Lehmann², Michał Podstawski³,
Hubert Niewiadomski², Piotr Nyczyk², Torsten Hoefler¹

¹ETH Zurich, ²Cledar, ³Warsaw University of Technology
bestam@inf.ethz.ch, nils.blach@inf.ethz.ch, htor@inf.ethz.ch

Graph of Thoughts: Architecture & Design



<https://github.com/spcl/graph-of-thoughts>,

@AAAI'24

☆ 1.6k stars 🔗 91 forks

Graph of Thoughts: Solving Elaborate Problems with Large Language Models

Maciej Besta^{1*}, Nils Blach^{1*}, Ales Kubicek¹, Robert Gerstenberger¹,
Lukas Gianinazzi¹, Joanna Gajda², Tomasz Lehmann², Michał Podstawski³,
Hubert Niewiadomski², Piotr Nyczyk², Torsten Hoefler¹

¹ETH Zurich, ²Cledar, ³Warsaw University of Technology
bestam@inf.ethz.ch, nils.blach@inf.ethz.ch, htor@inf.ethz.ch

Evaluation: Used Machine & Objectives



Evaluation: Used Machine & Objectives

CSCS Cray Piz Daint & Ault
64GB – 2TB memory per server



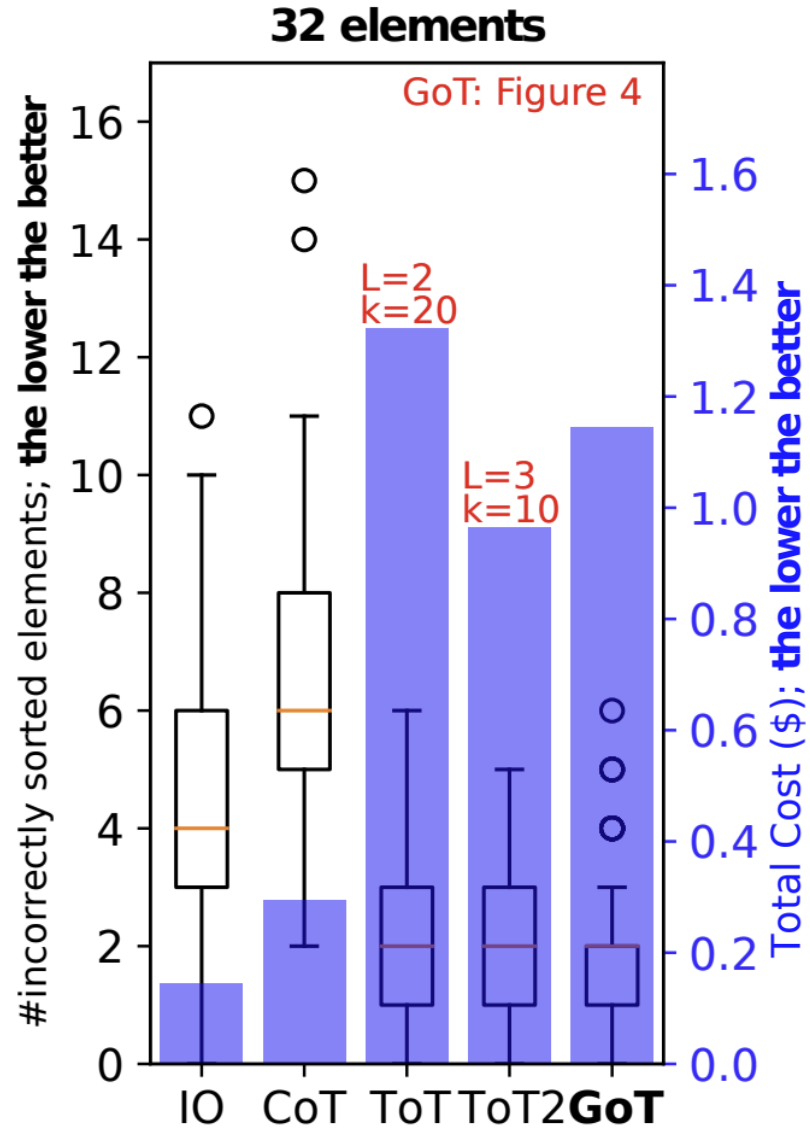
Evaluation: Used Machine & Objectives

CSCS Cray Piz Daint & Ault
64GB – 2TB memory per server

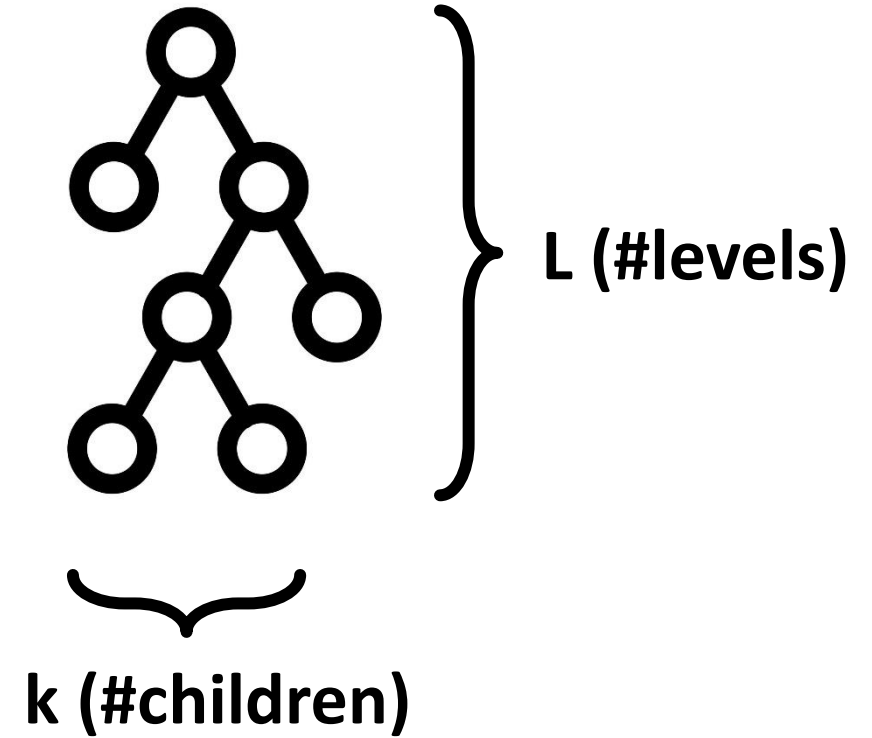
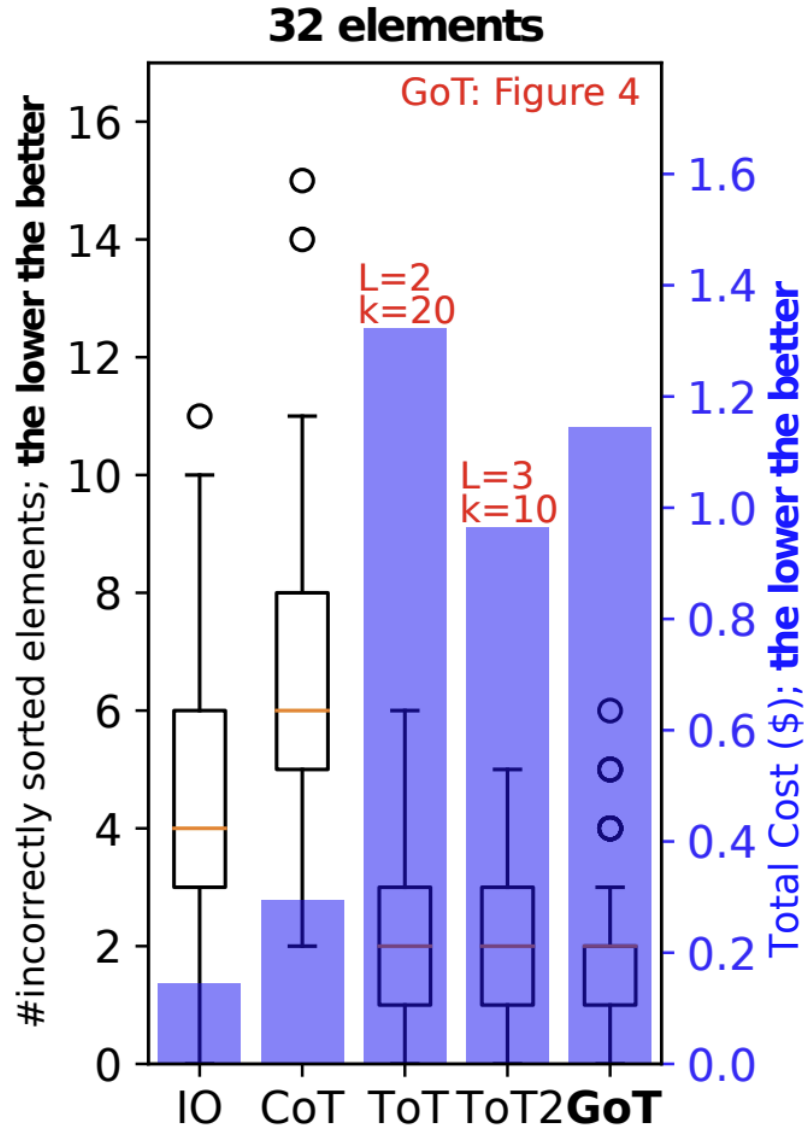
Main goal: show that GoT successfully harnesses the graph abstraction to enable more efficient queries



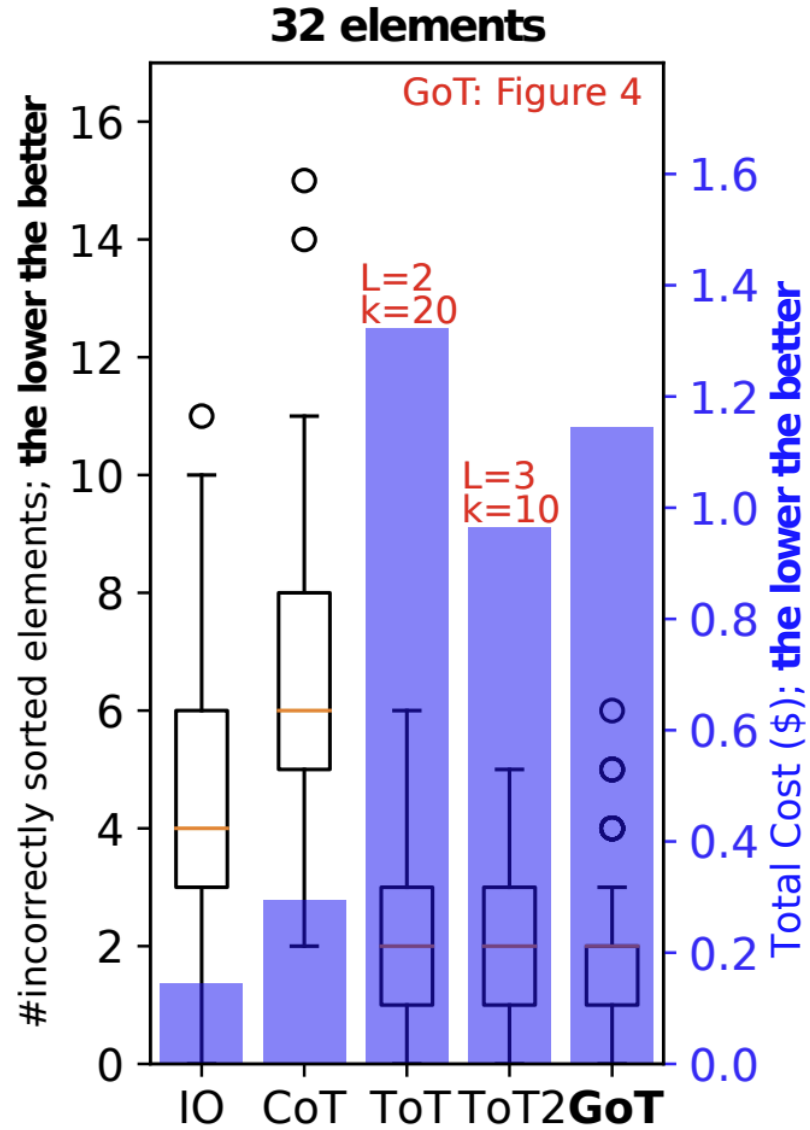
Sorting Numbers



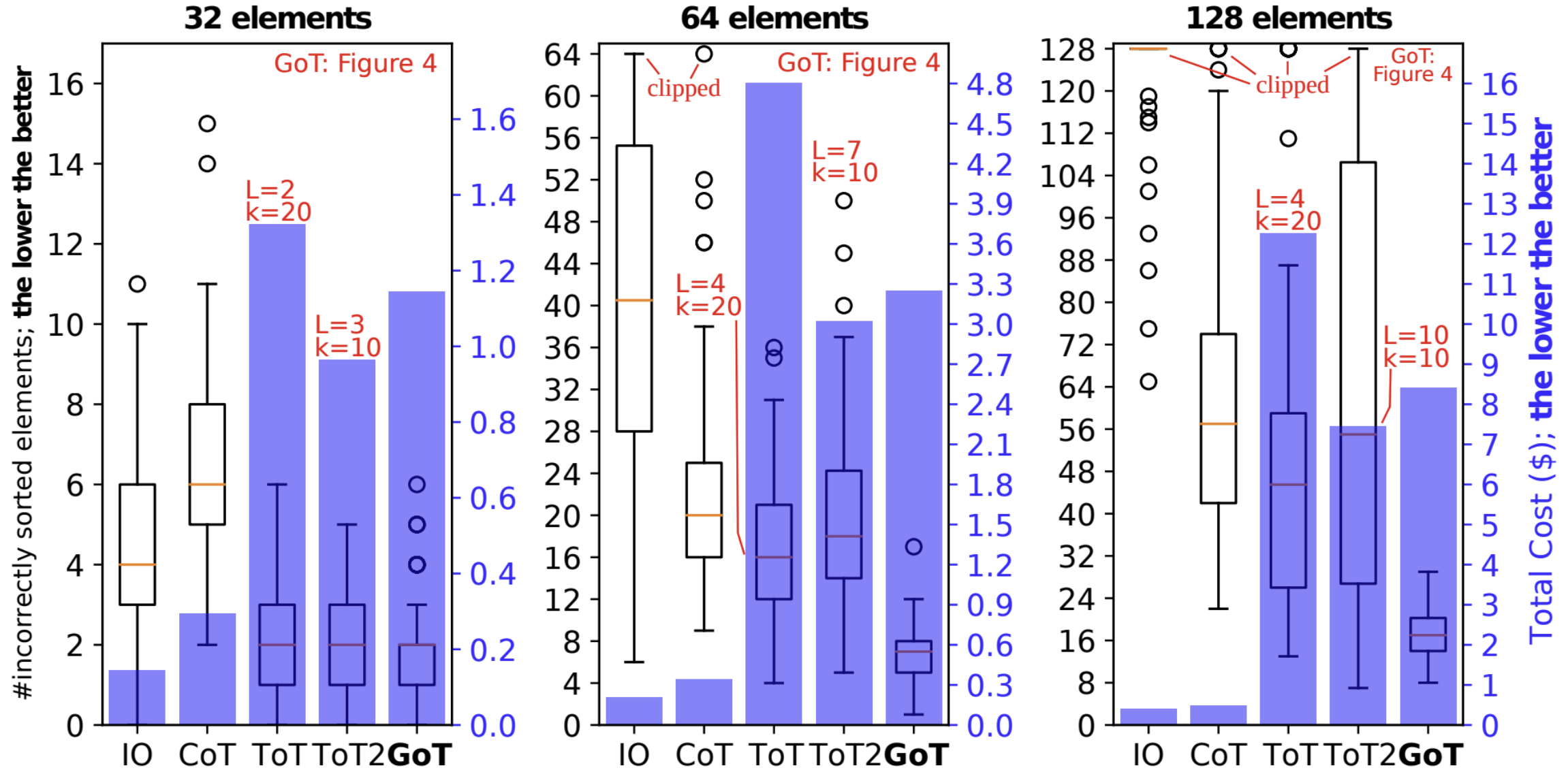
Sorting Numbers



Sorting Numbers

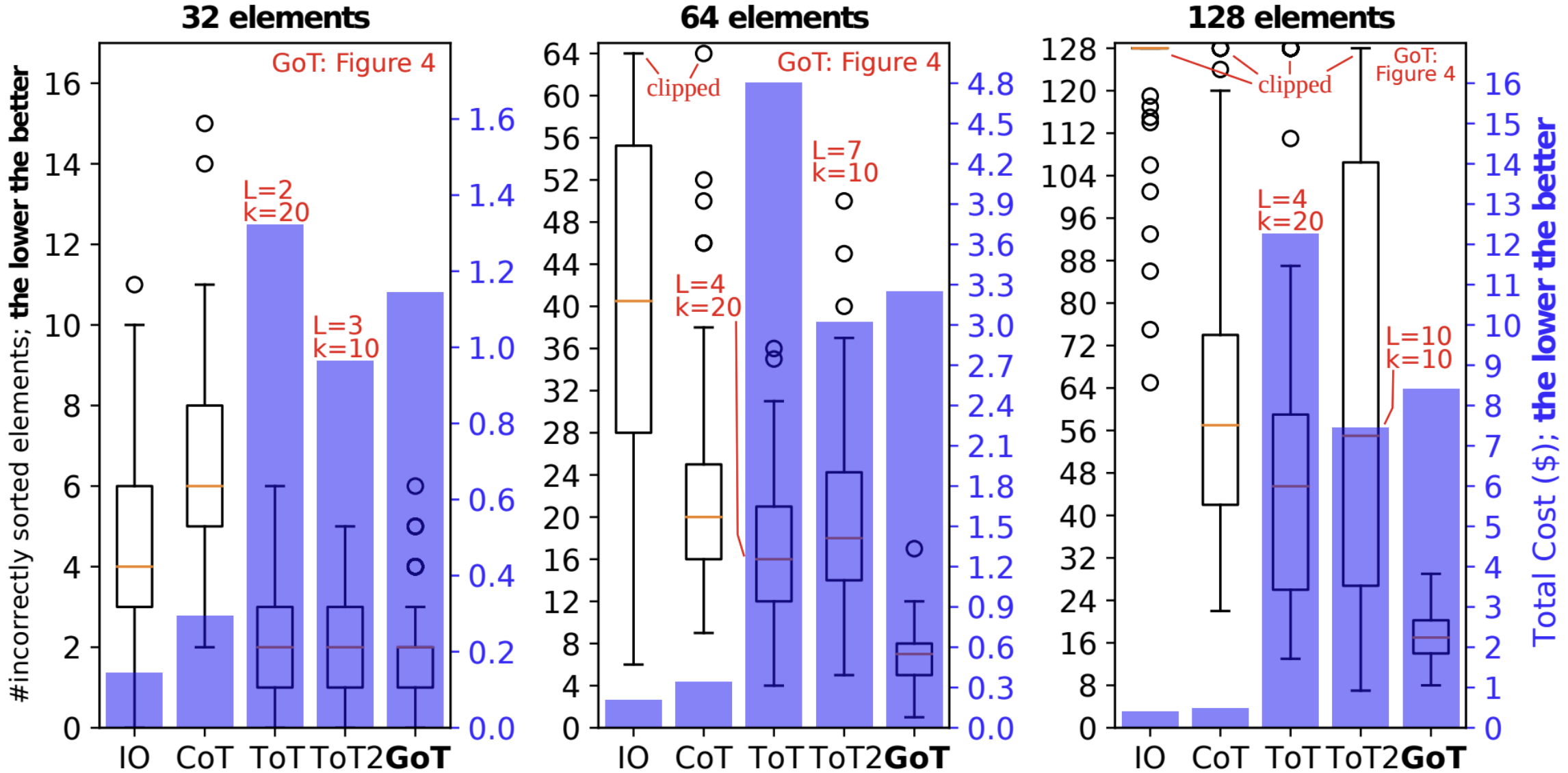


Sorting Numbers

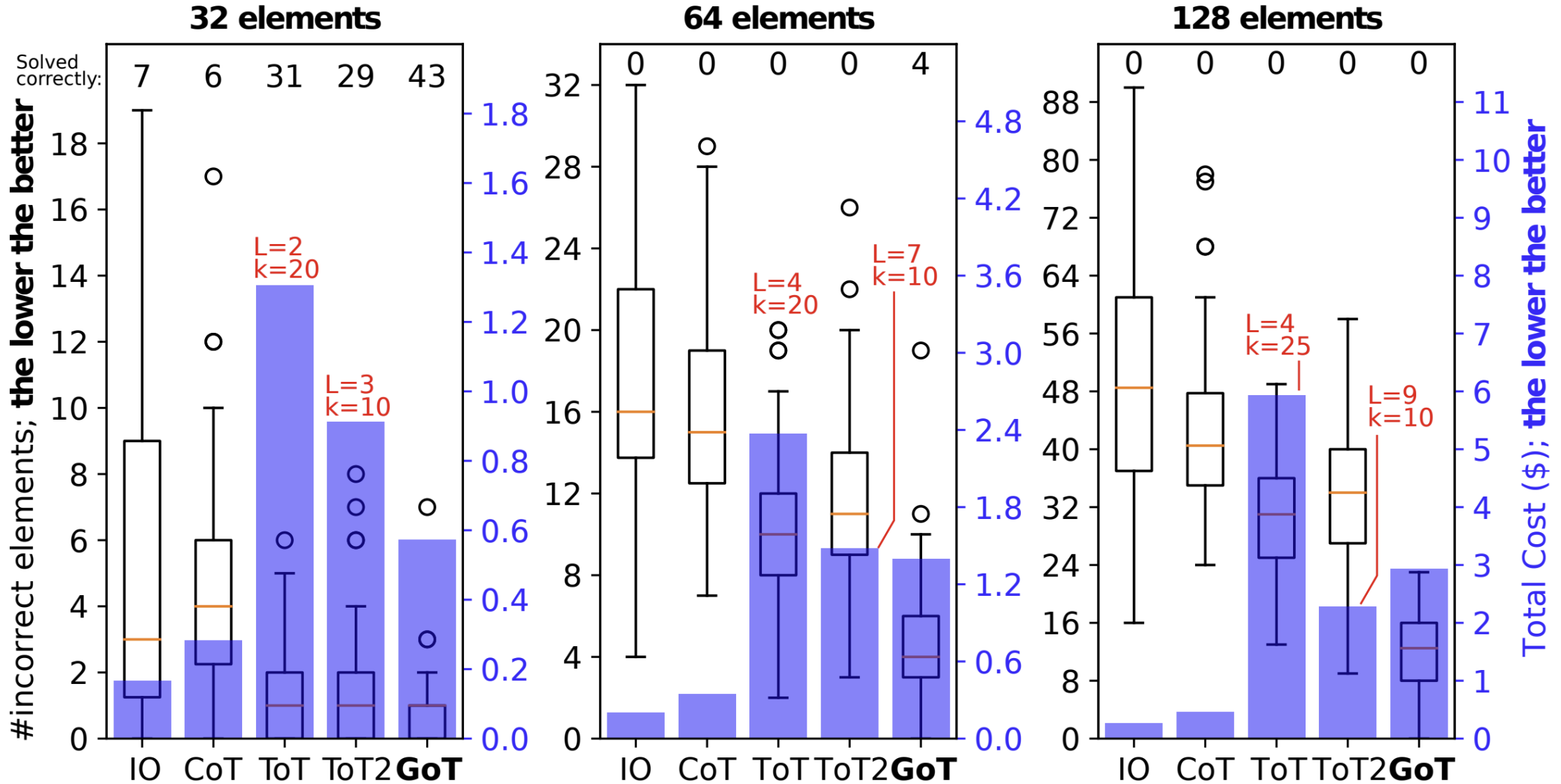


Sorting Numbers

The longer the sequence, the higher the gain

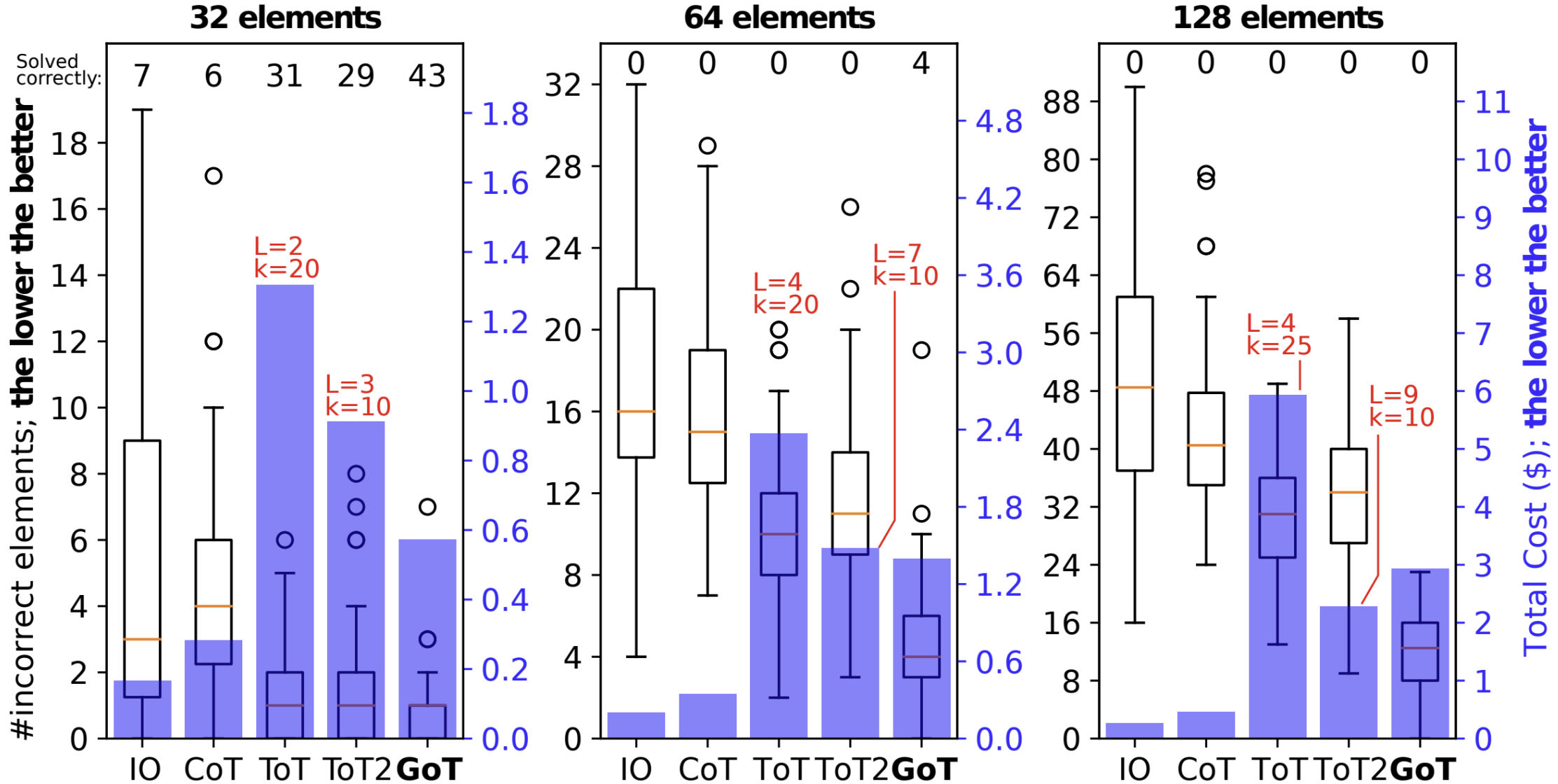


Intersecting Sets of Numbers

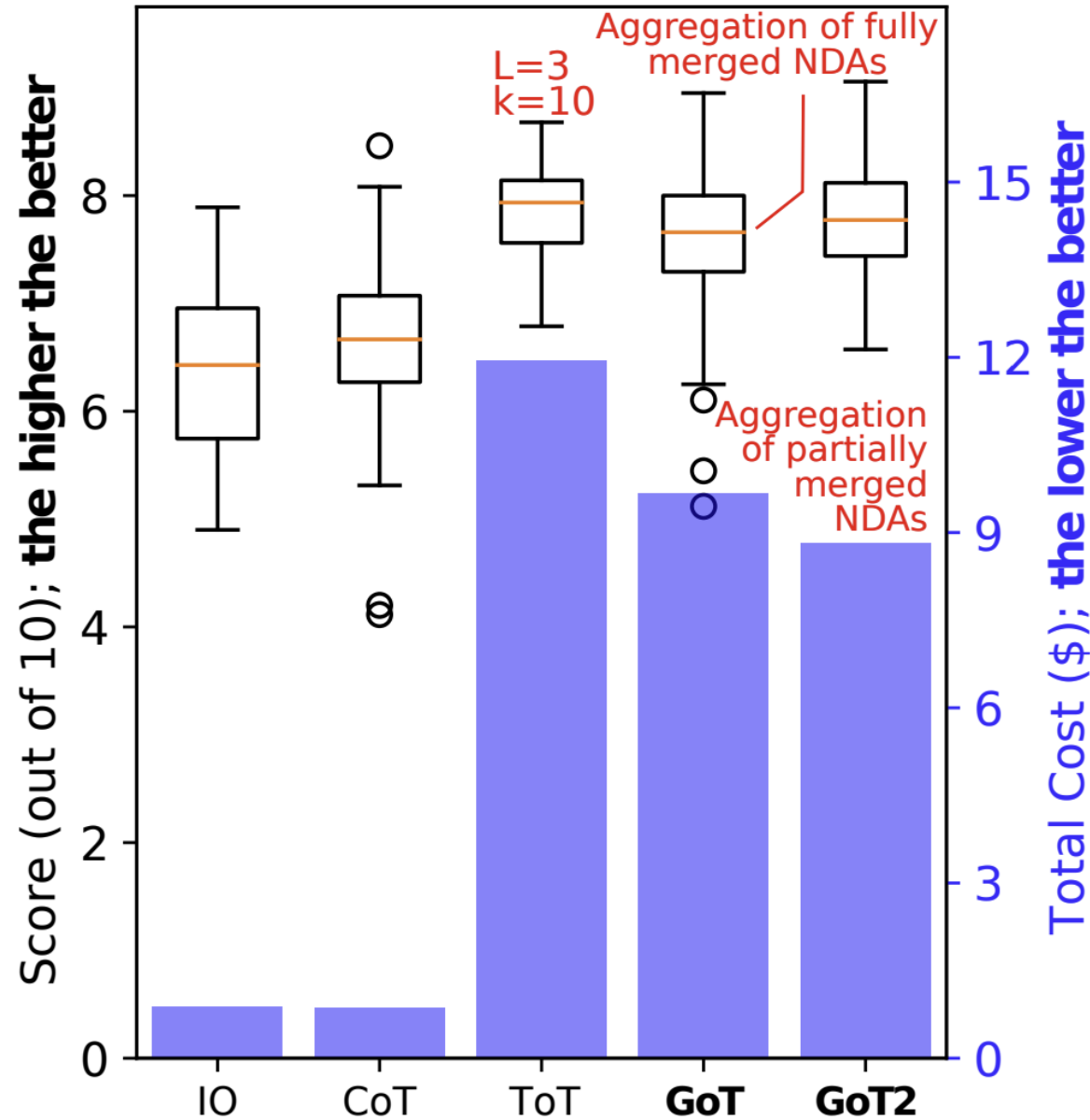


Intersecting Sets of Numbers

The longer the sequence, the higher the gain

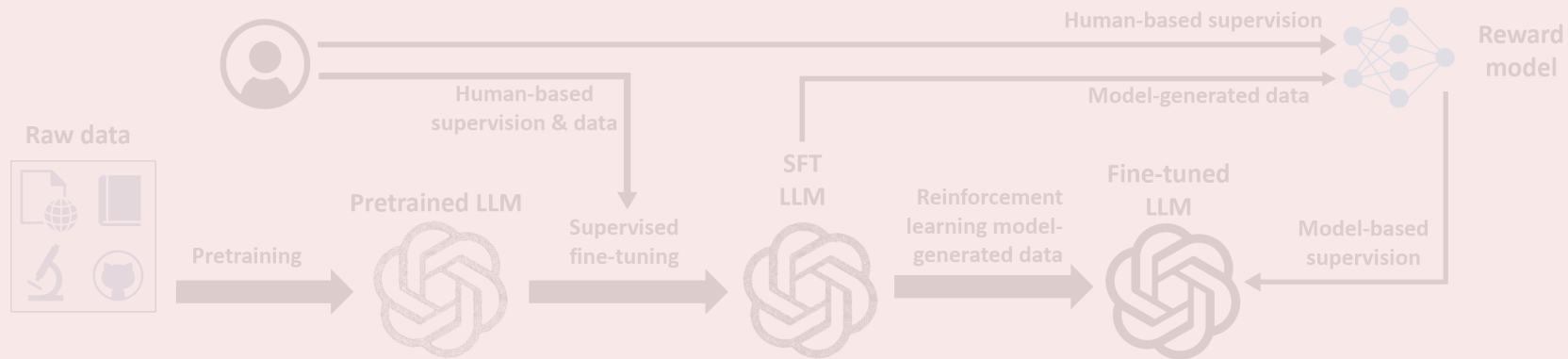


Merging Documents



The Emergence of the „Generative AI Ecosystem”

Training related



Inference related

Prompting Structures

Tools

Psychology

Web

Retrieval

The Emergence of the „Generative AI Ecosystem”

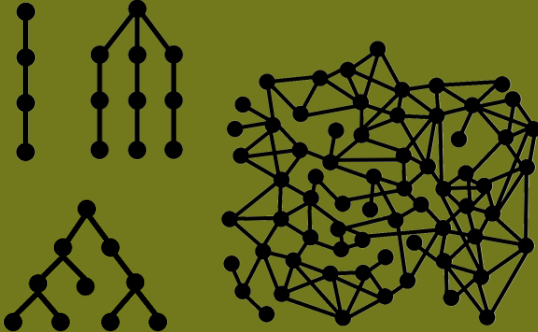
Training related




Why does structured prompting work?

Inference related

Prompting Structures



Tools



Psychology



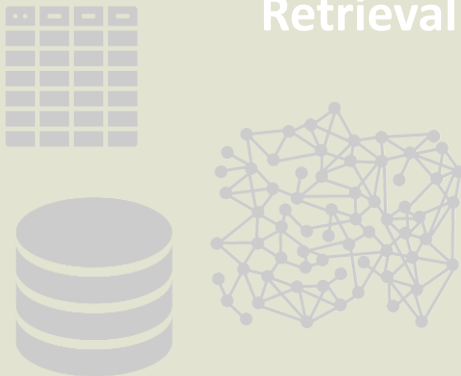
Replies („thoughts”)
Prompts



Web



Retrieval



Why Does Structured Prompting Work?

Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)

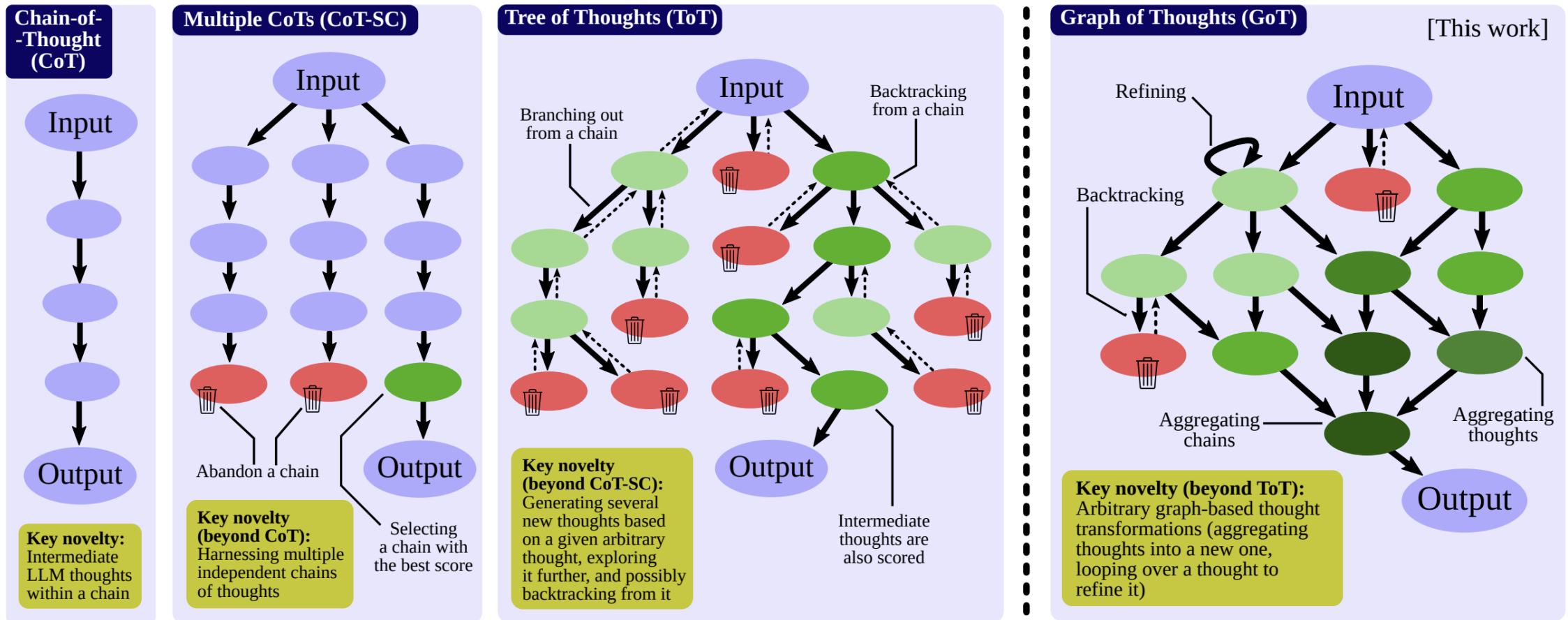
Why Does Structured Prompting Work?

Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)

Volume – for a given thought t – is the number of preceding LLM thoughts that could have impacted t

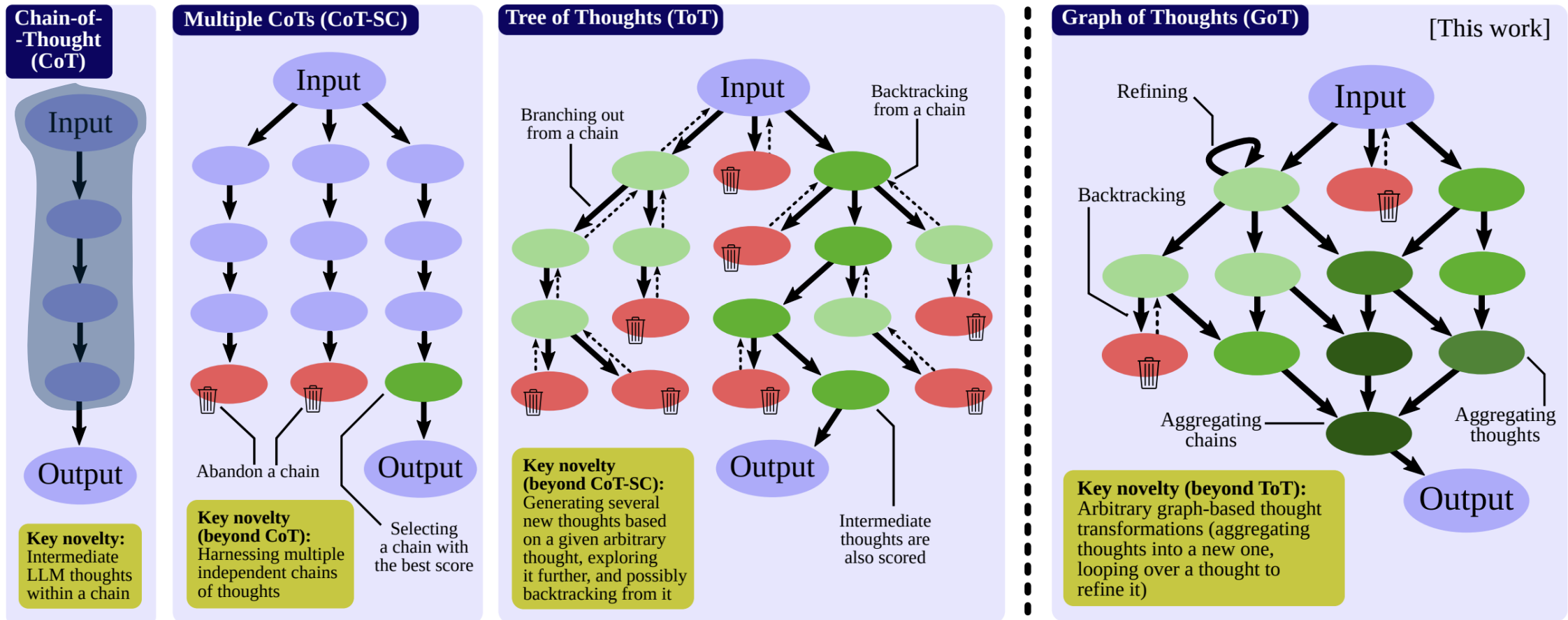
Why Does Structured Prompting Work?

Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)



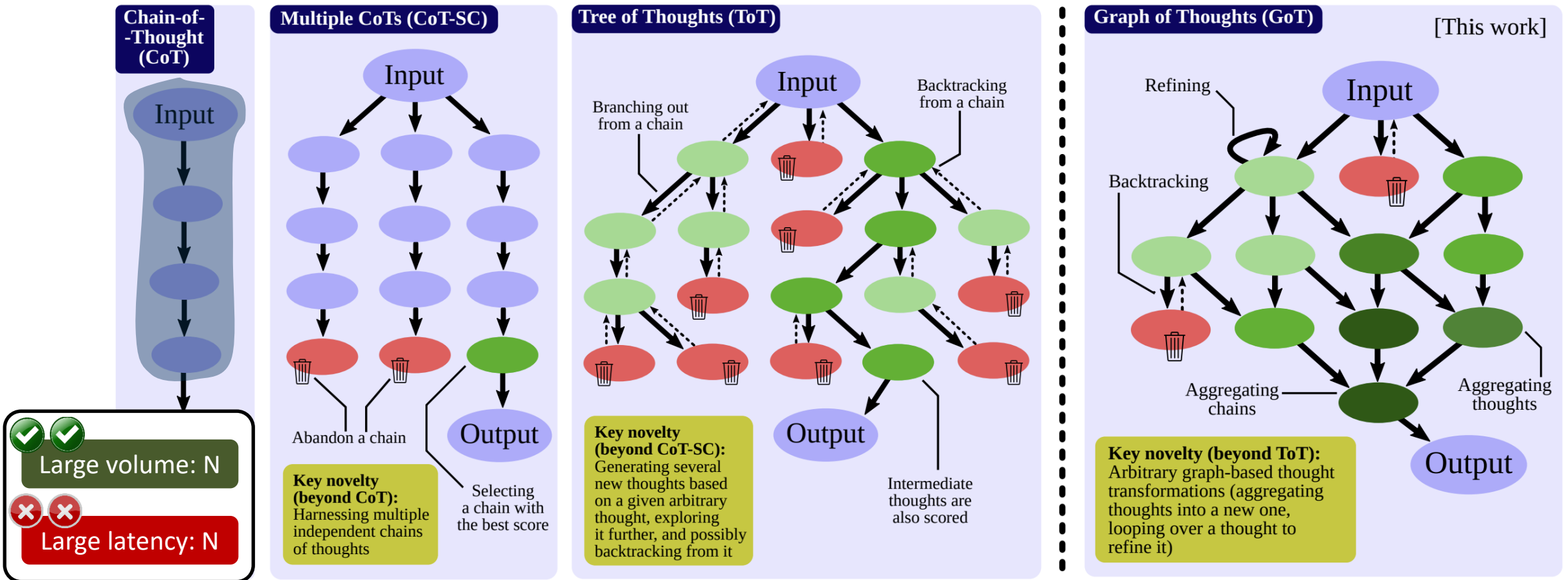
Why Does Structured Prompting Work?

Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)



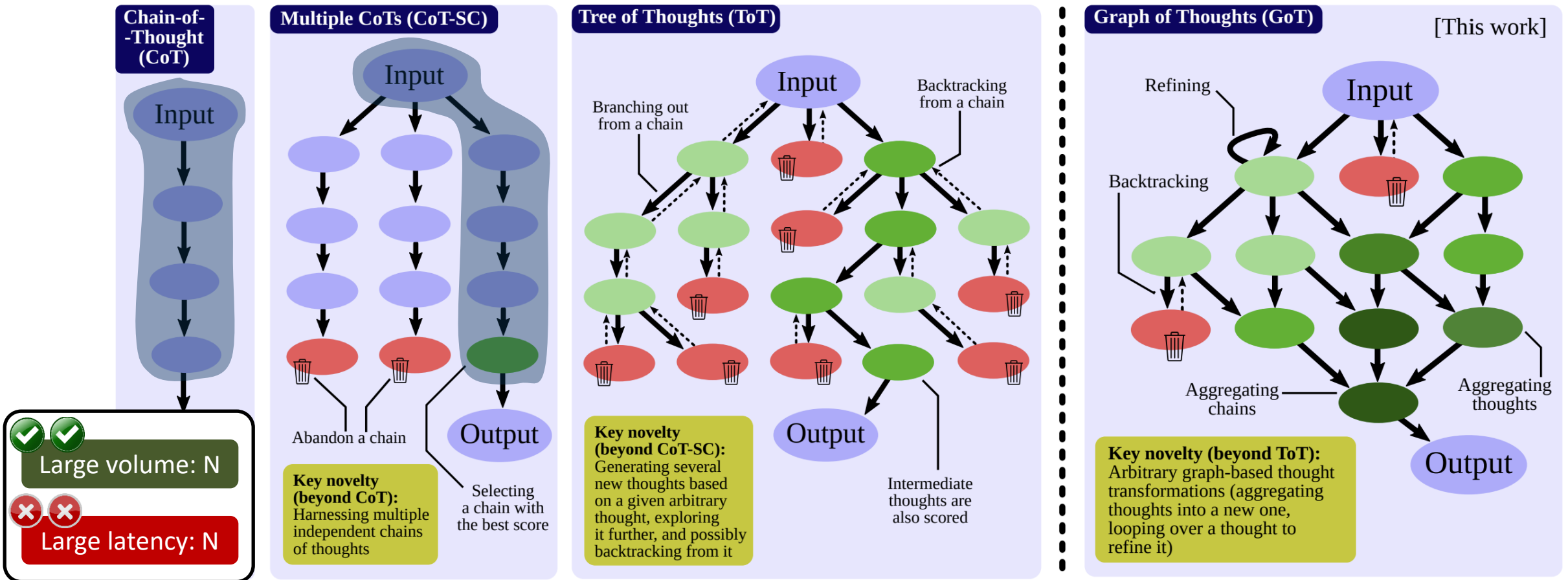
Why Does Structured Prompting Work?

Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)



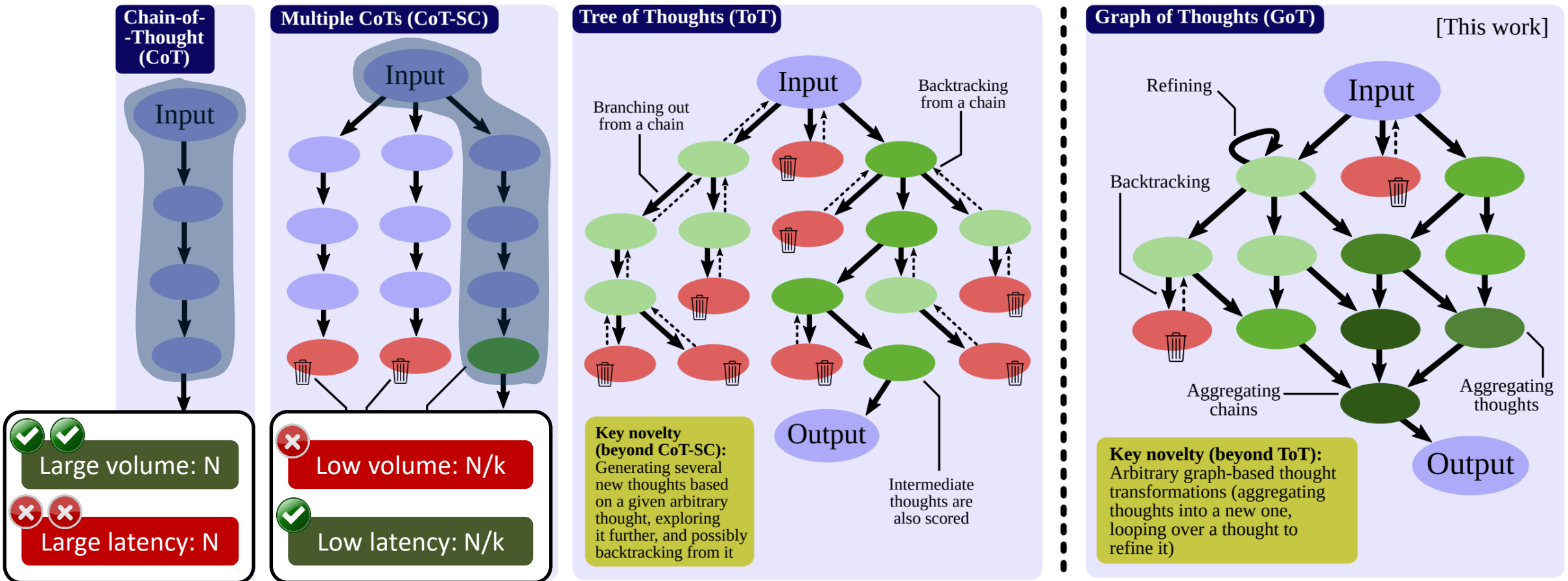
Why Does Structured Prompting Work?

Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)



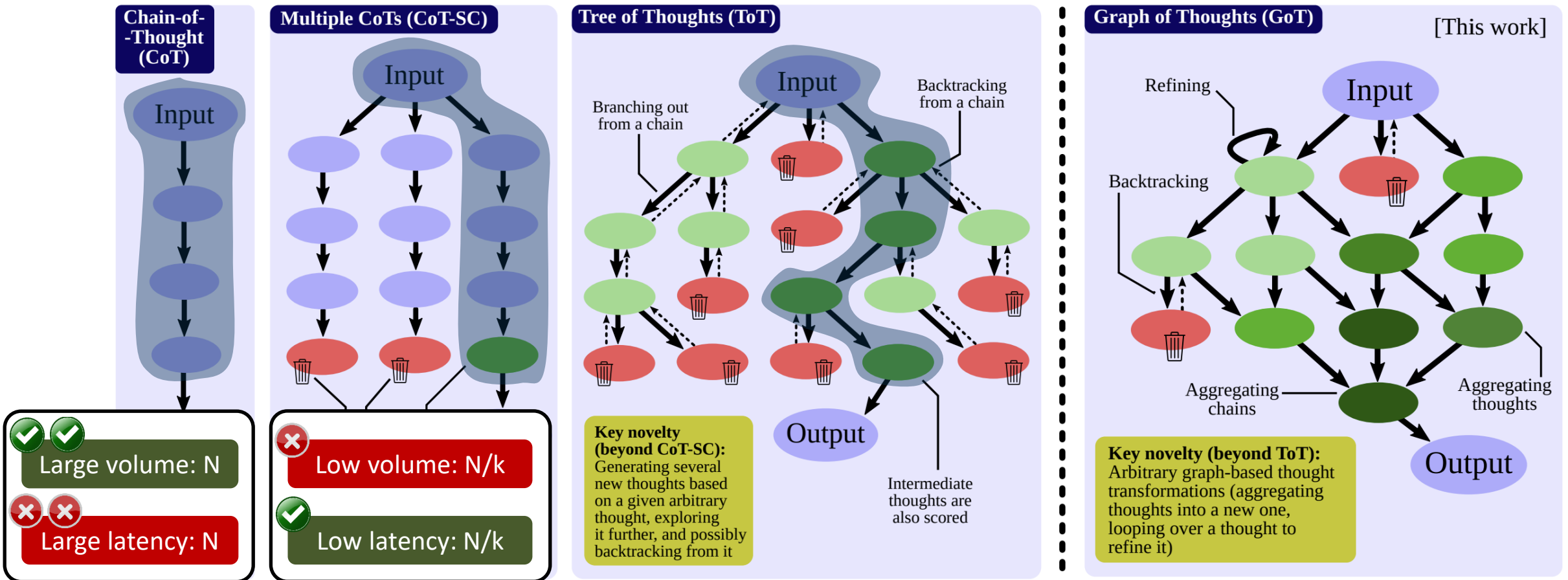
Why Does Structured Prompting Work?

Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)



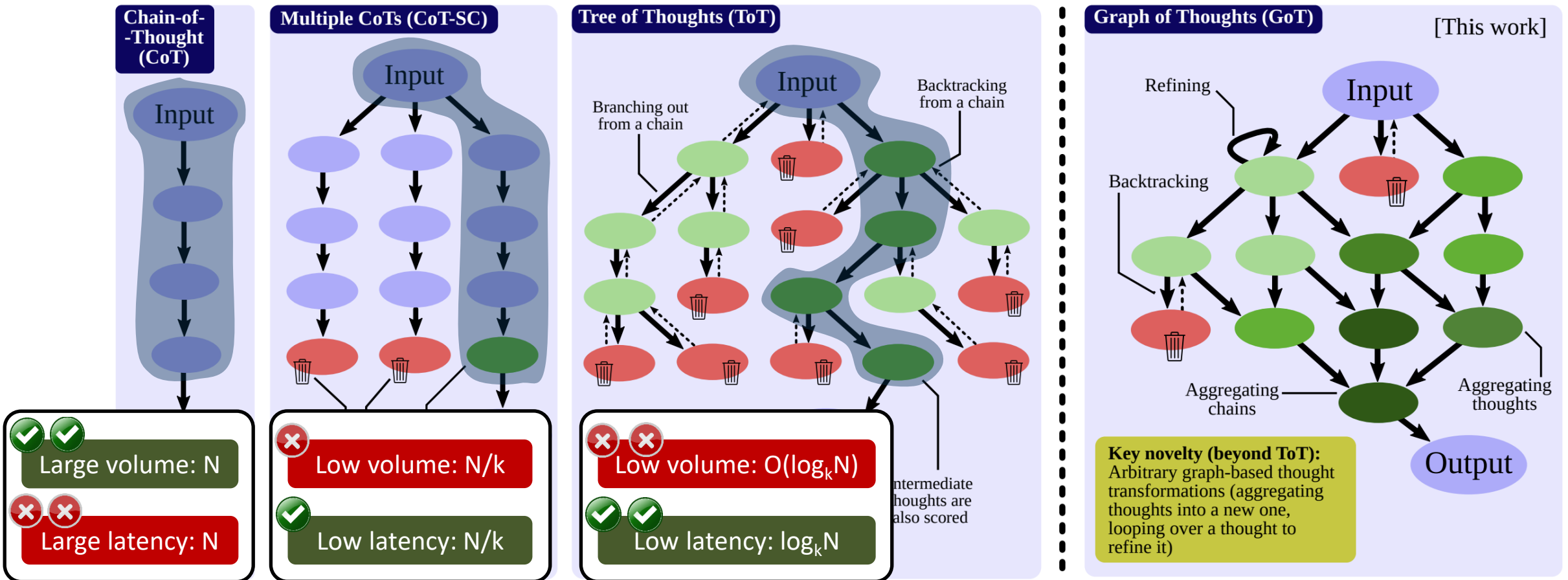
Why Does Structured Prompting Work?

Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)



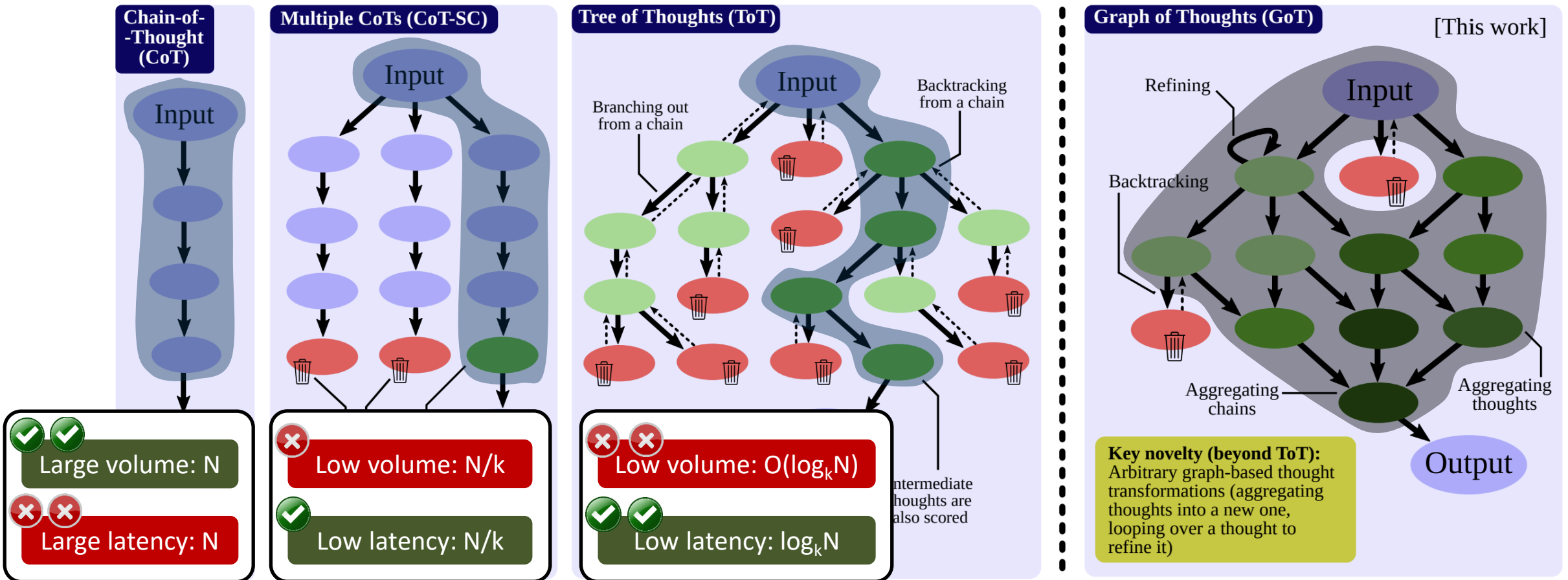
Why Does Structured Prompting Work?

Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)



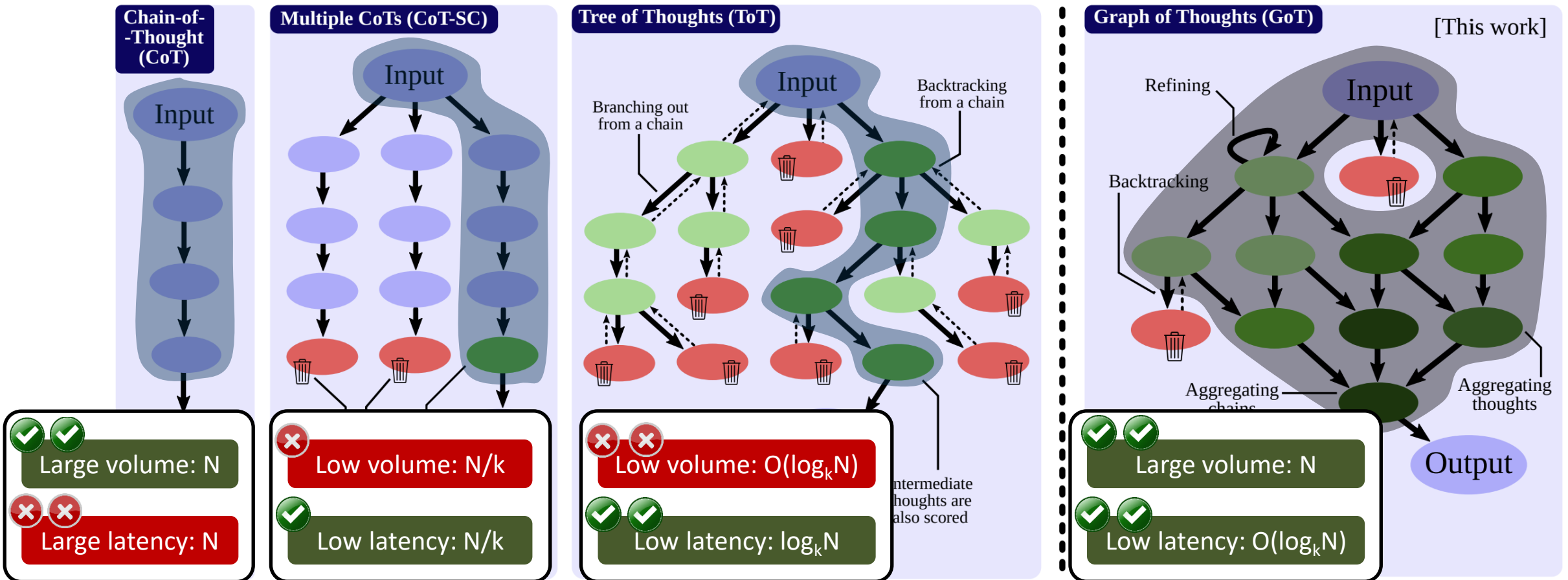
Why Does Structured Prompting Work?

Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)

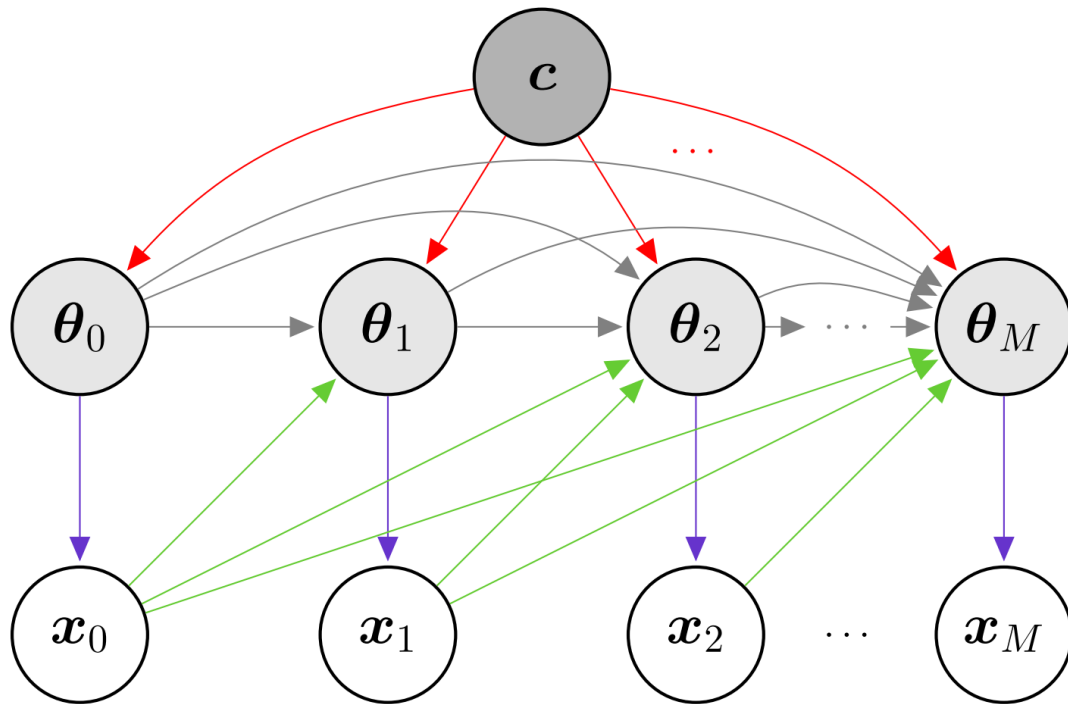


Why Does Structured Prompting Work?

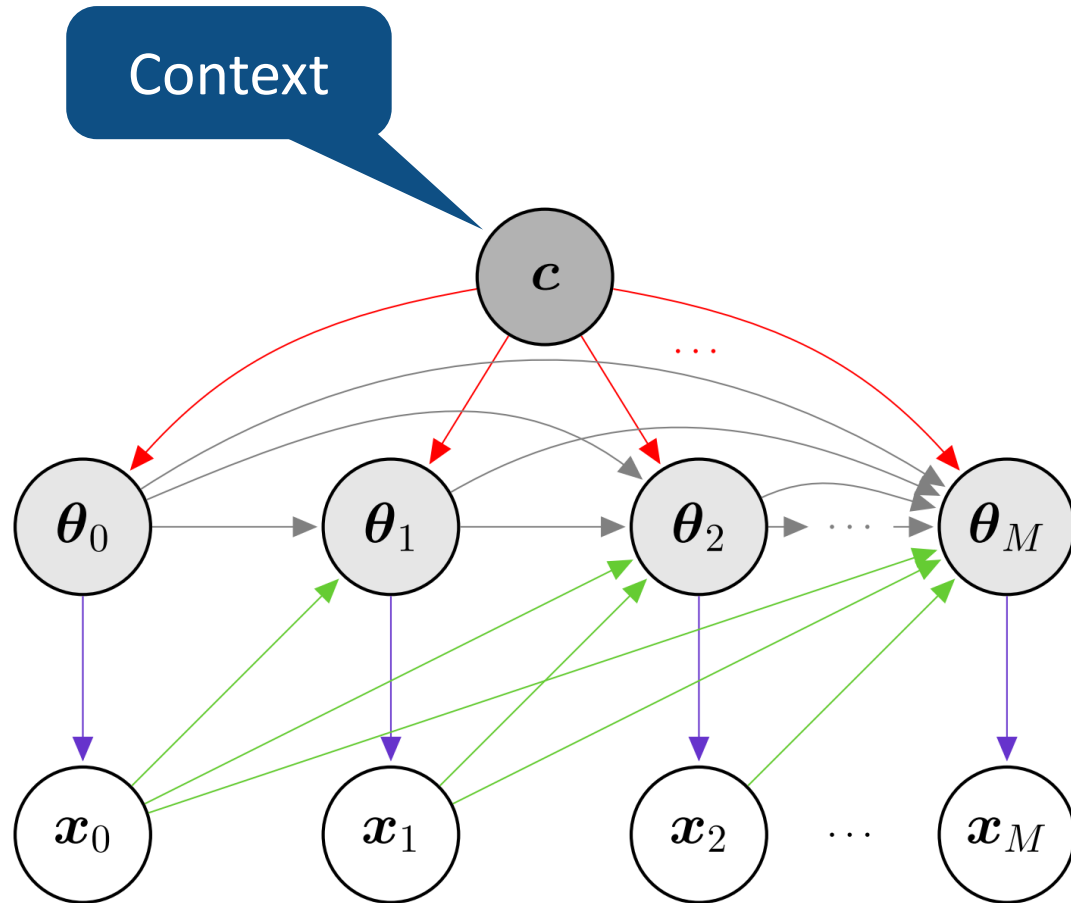
Assume a **fixed thought size** (#tokens) and a **fixed context size** (#thoughts in the LLM context, denoted with N)



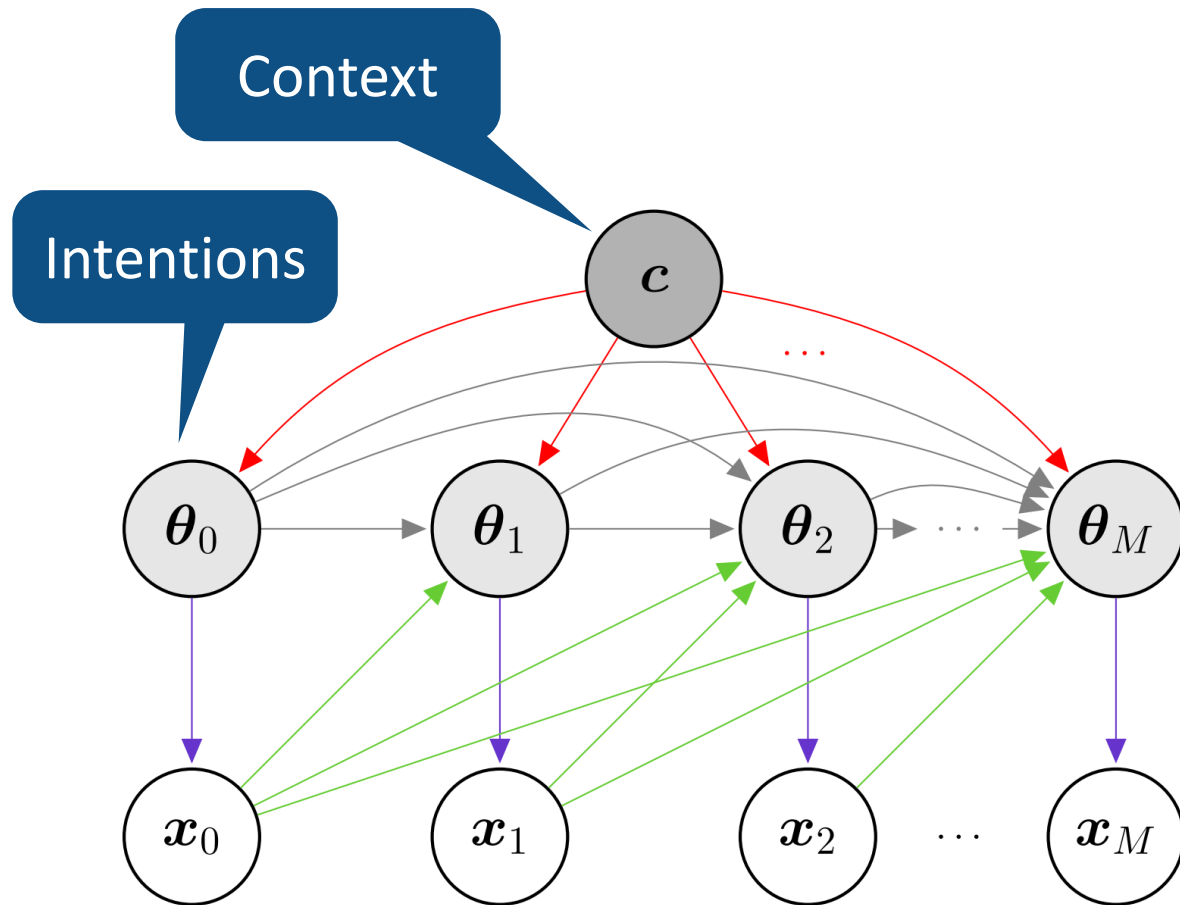
Why Does Structured Prompting Work? Probabilistic Graphical Models [1]



Why Does Structured Prompting Work? Probabilistic Graphical Models [1]

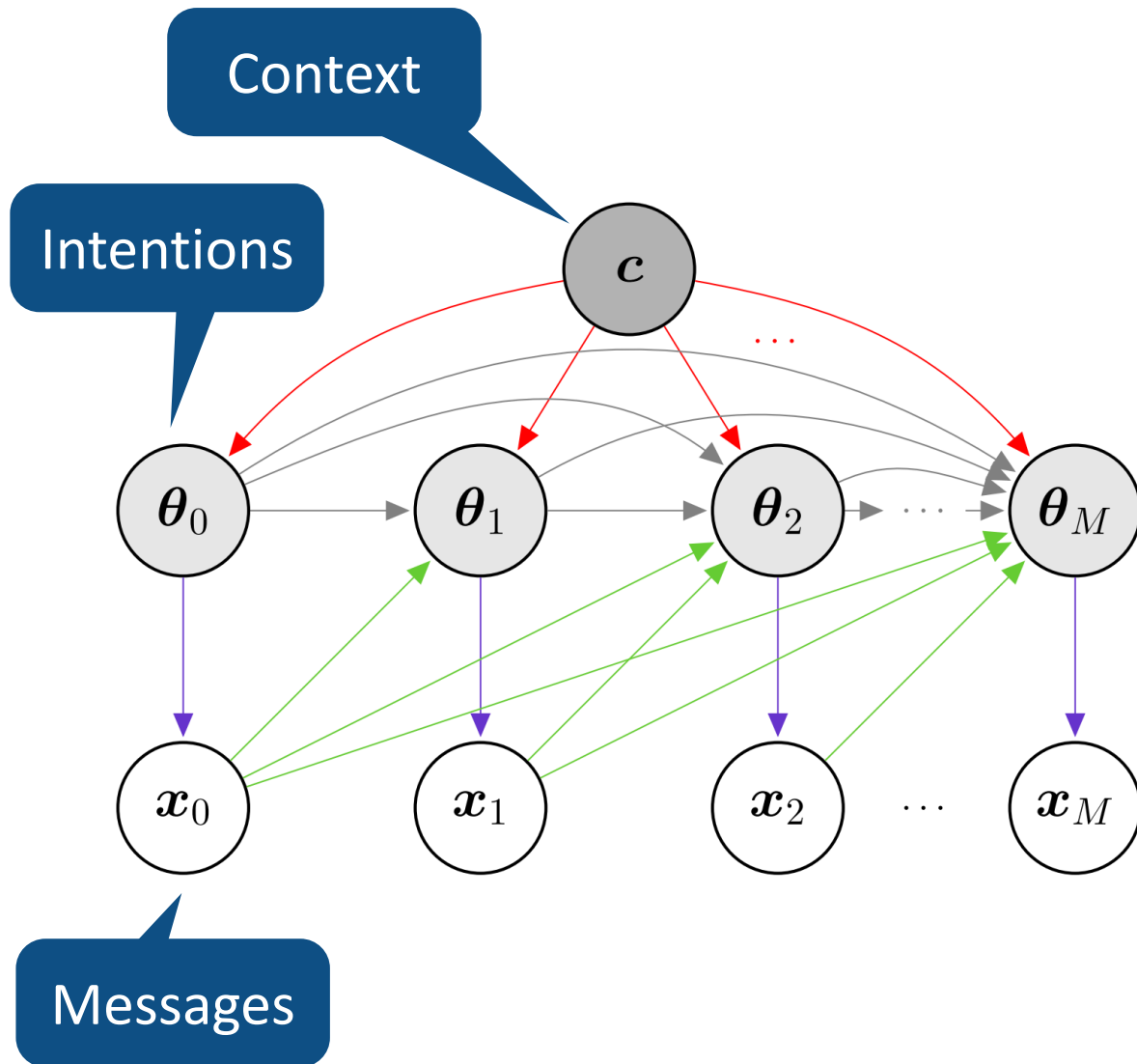


Why Does Structured Prompting Work? Probabilistic Graphical Models [1]



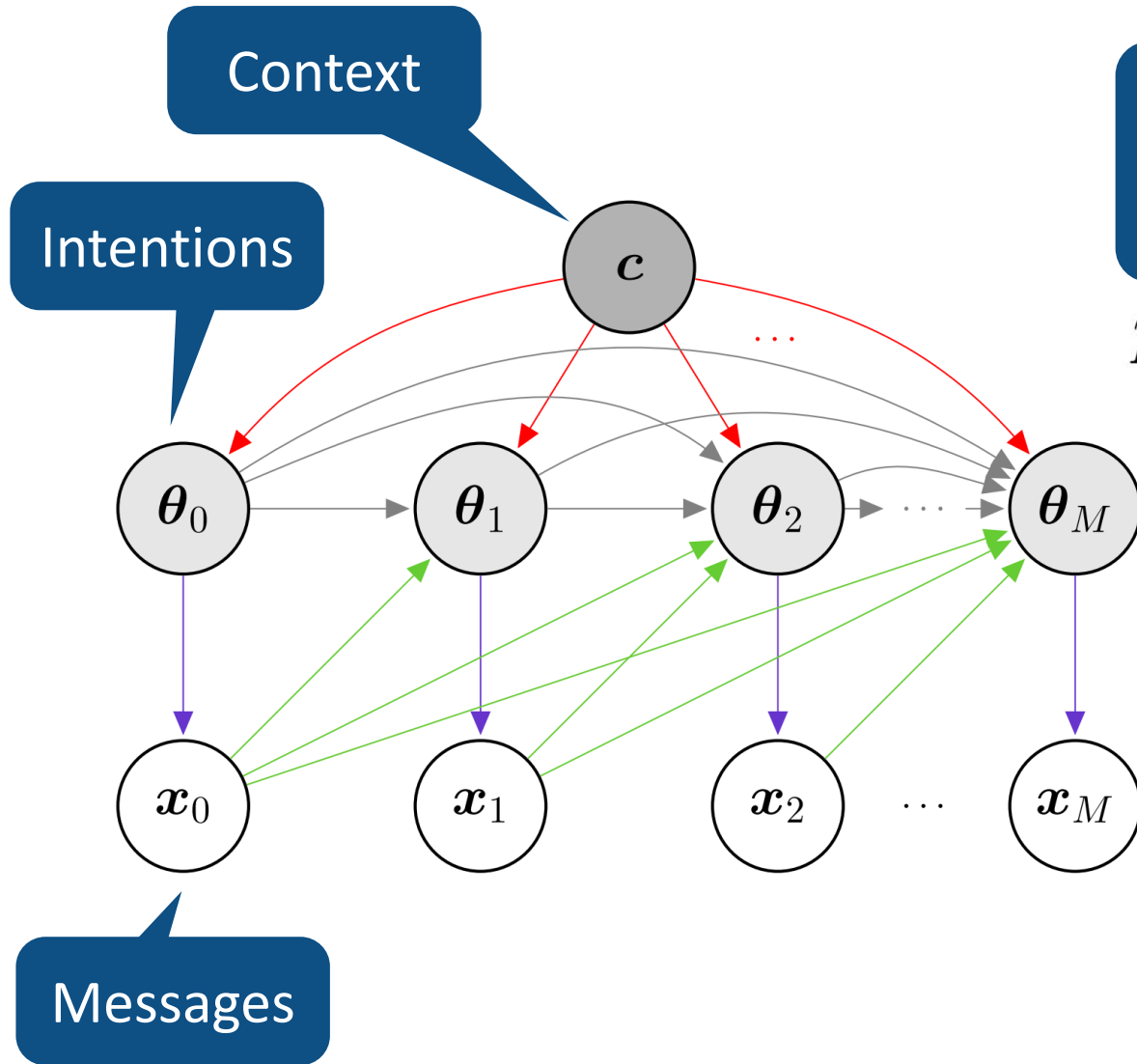
[1] R. Tatunov et al. *Why Can Large Language Models Generate Correct Chain-of-Thoughts?* Arxiv, 30 October 2023.

Why Does Structured Prompting Work? Probabilistic Graphical Models [1]



[1] R. Tatunov et al. *Why Can Large Language Models Generate Correct Chain-of-Thoughts?* Arxiv, 30 October 2023.

Why Does Structured Prompting Work? Probabilistic Graphical Models [1]

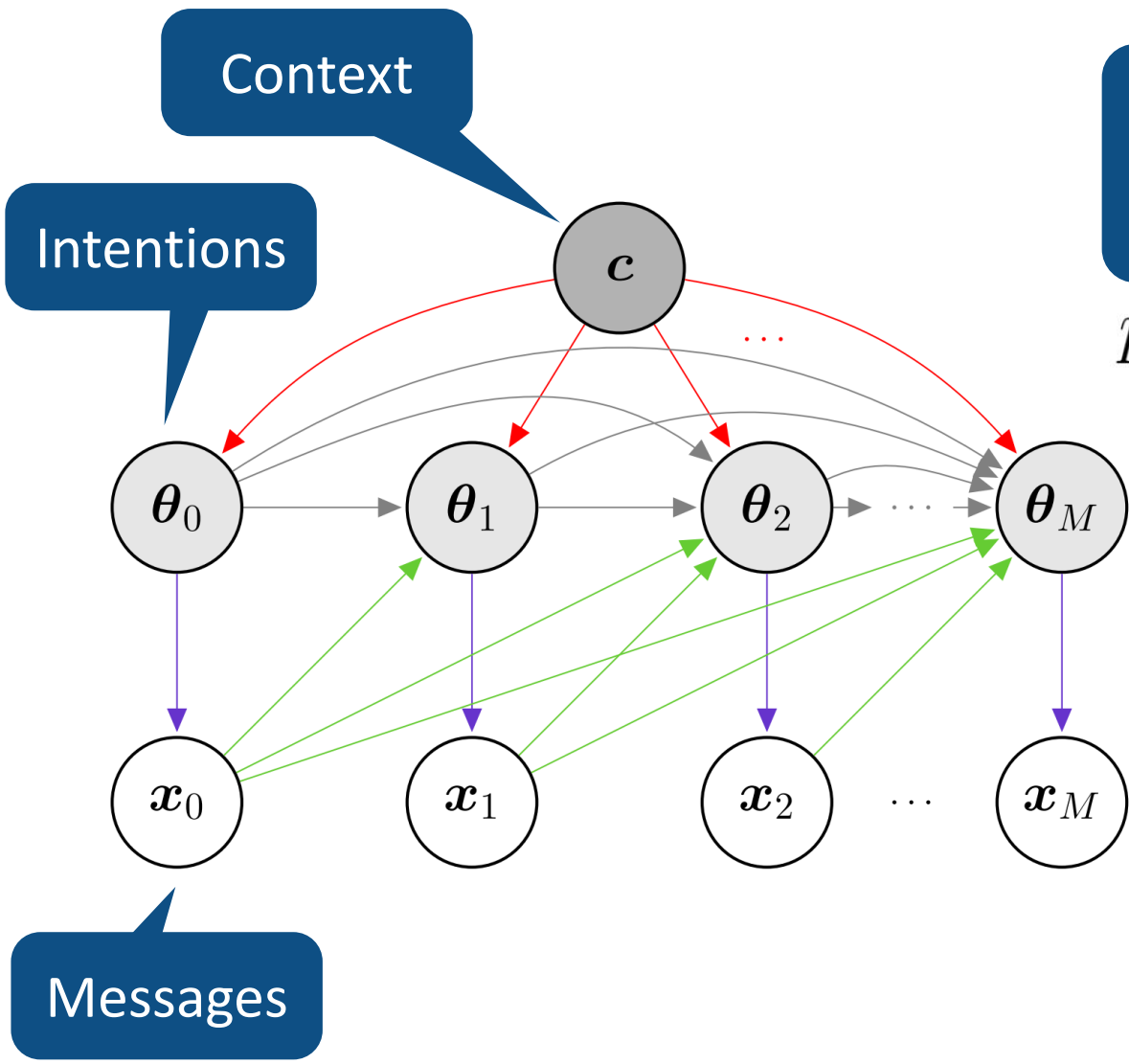


Likelihood of generating a chain of thoughts, basing on a pre-prompt with input I and with N CoT in-context examples, but without access to the true context.

$$p_{LLM} \equiv p_{LLM}(\text{CoT} | I, \text{CoT-Examples}(N))$$

[1] R. Tatunov et al. *Why Can Large Language Models Generate Correct Chain-of-Thoughts?* Arxiv, 30 October 2023.

Why Does Structured Prompting Work? Probabilistic Graphical Models [1]



Likelihood of generating a chain of thoughts, basing on a pre-prompt with input I and with N CoT in-context examples, but without access to the true context.

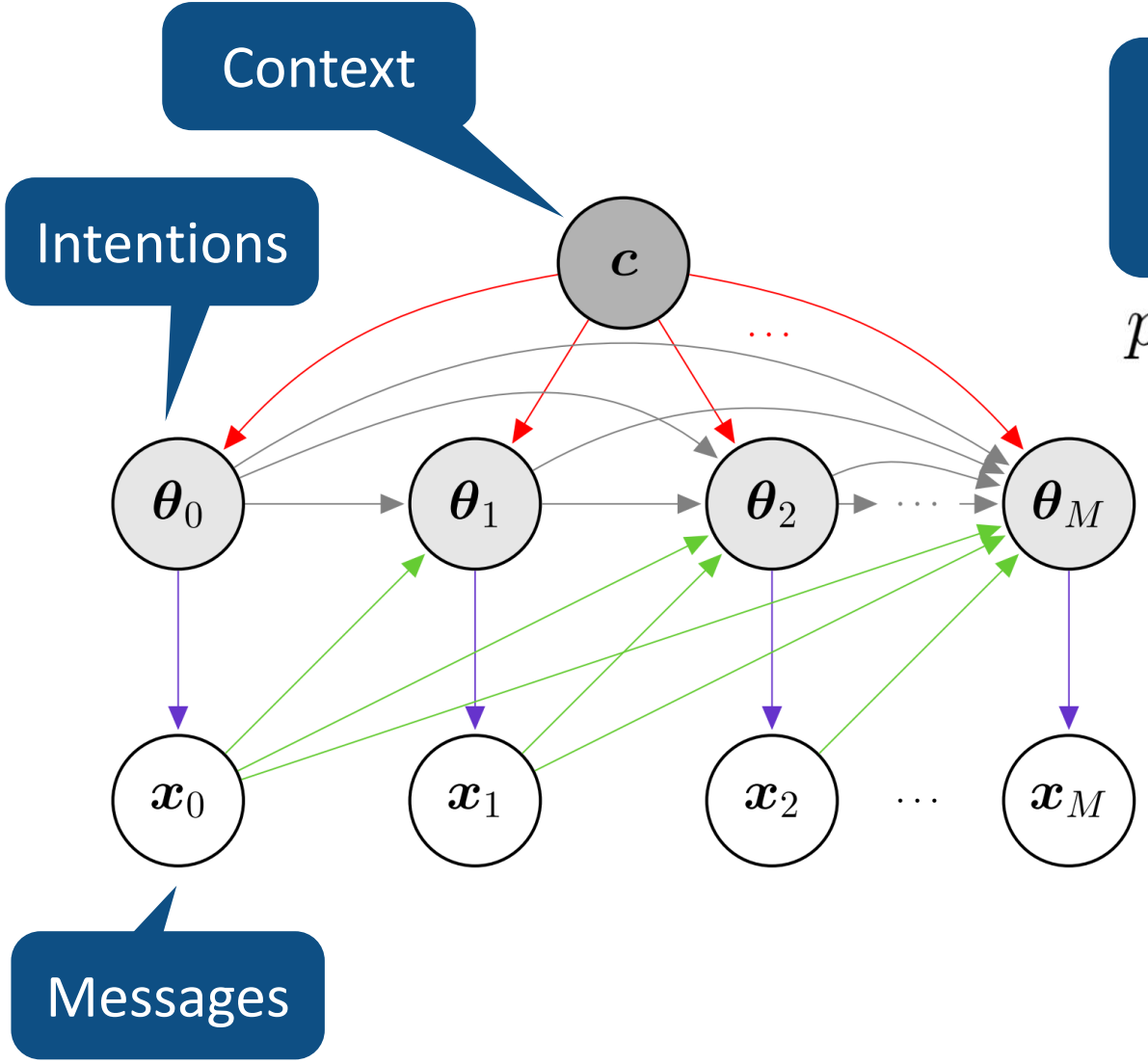
$$p_{LLM} \equiv p_{LLM}(\text{CoT} | I, \text{CoT-Examples}(N))$$

Likelihood of generating the same chain of thoughts as above, but using the true language (and context) conditioned on the same input I

$$p_{\text{True}} \equiv p_{\text{True}}(\text{CoT} | I, \text{True-Context})$$

[1] R. Tatunov et al. *Why Can Large Language Models Generate Correct Chain-of-Thoughts?* Arxiv, 30 October 2023.

Why Does Structured Prompting Work? Probabilistic Graphical Models [1]



Likelihood of generating a chain of thoughts, basing on a pre-prompt with input I and with N CoT in-context examples, but without access to the true context.

$$p_{LLM} \equiv p_{LLM}(\text{CoT} | I, \text{CoT-Examples}(N))$$

Likelihood of generating the same chain of thoughts as above, but using the true language (and context) conditioned on the same input I

$$p_{\text{True}} \equiv p_{\text{True}}(\text{CoT} | I, \text{True-Context})$$

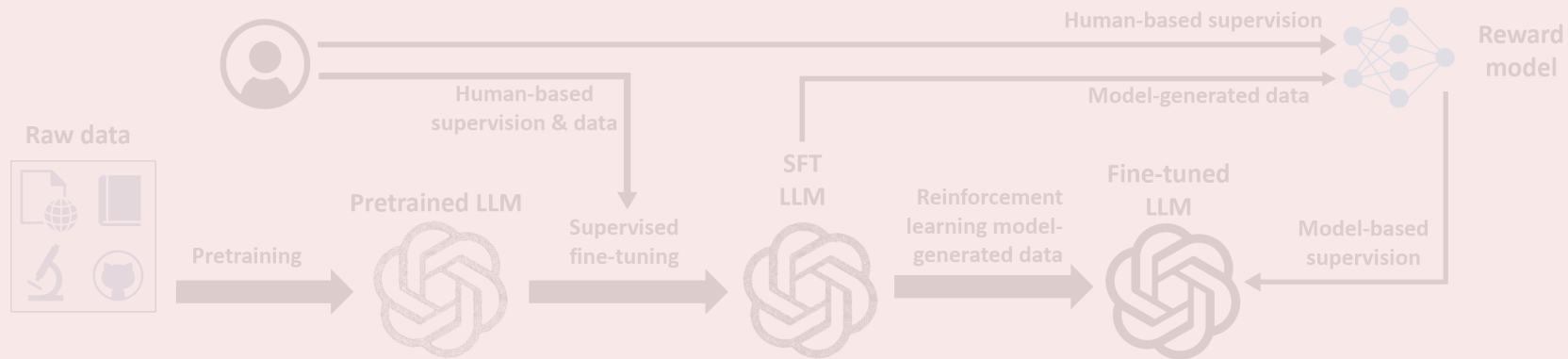
$$|p_{LLM} - p_{\text{True}}| \leq \rho^N$$

„A function of the language ambiguities“, < 1

[1] R. Tatunov et al. *Why Can Large Language Models Generate Correct Chain-of-Thoughts?* Arxiv, 30 October 2023.

The Emergence of the „Generative AI Ecosystem”

Training related



Inference related

Prompting Structures

Tools

Psychology

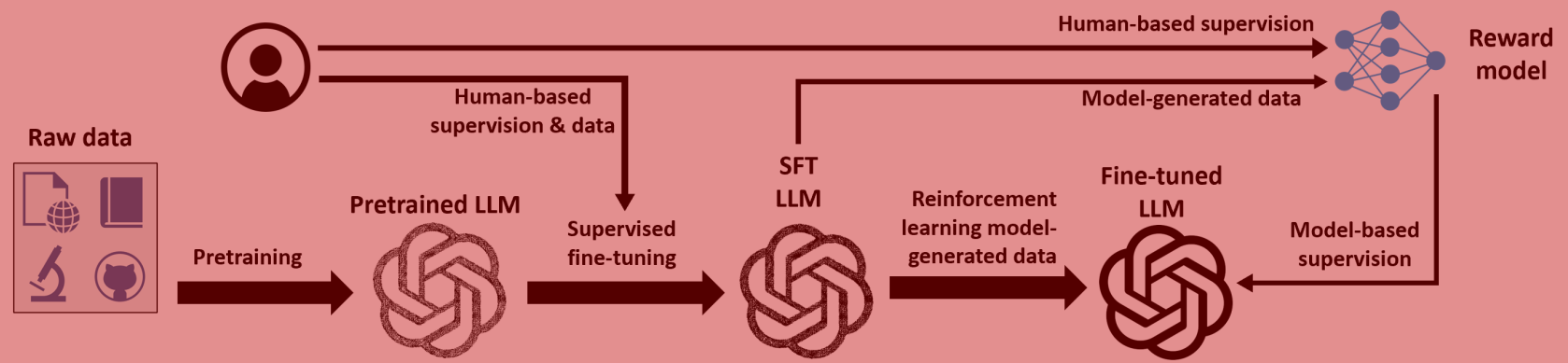
Web

Replies („thoughts”) ↑ Prompts ↓

Retrieval

The Emergence of the „Generative AI Ecosystem”: Training

Training related



Inference related

Prompting Structures

Tools

Psychology

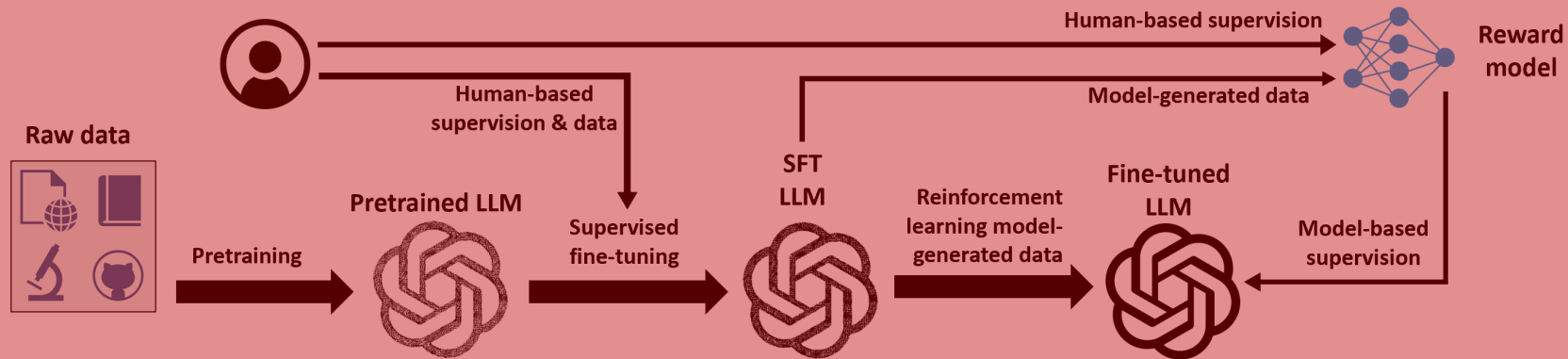
Replies („thoughts”)
↓
↑
Prompts

Web

Retrieval

The Emergence of the „Generative AI Ecosystem”: Training

Training related



Inference related

Prompting Structures

Tools

Psychology

Web

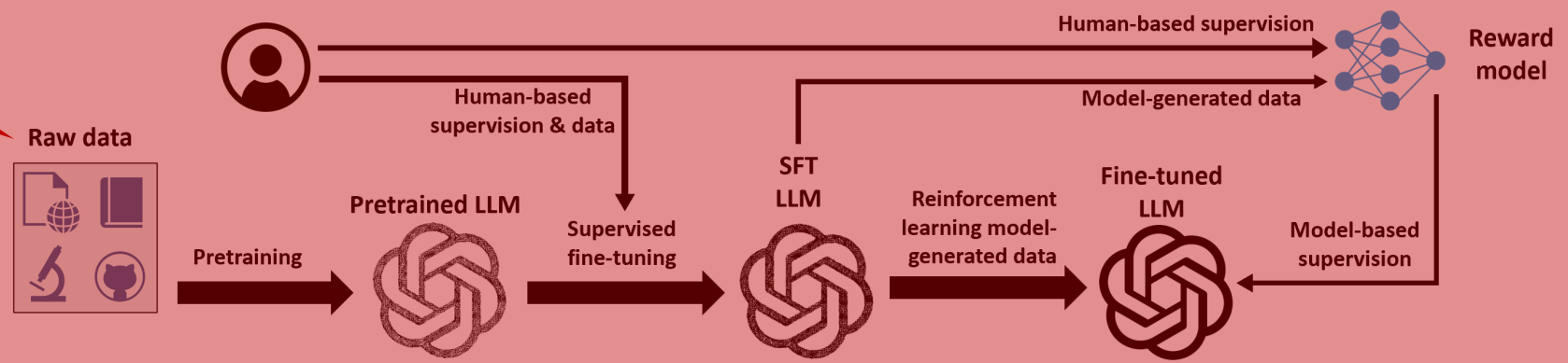
Replies („thoughts”) ↑ Prompts ↓

Retrieval

The Emergence of the „Generative AI Ecosystem”: Training

Training related

Graphs as modality



Inference related

Prompting Structures

Tools

Psychology

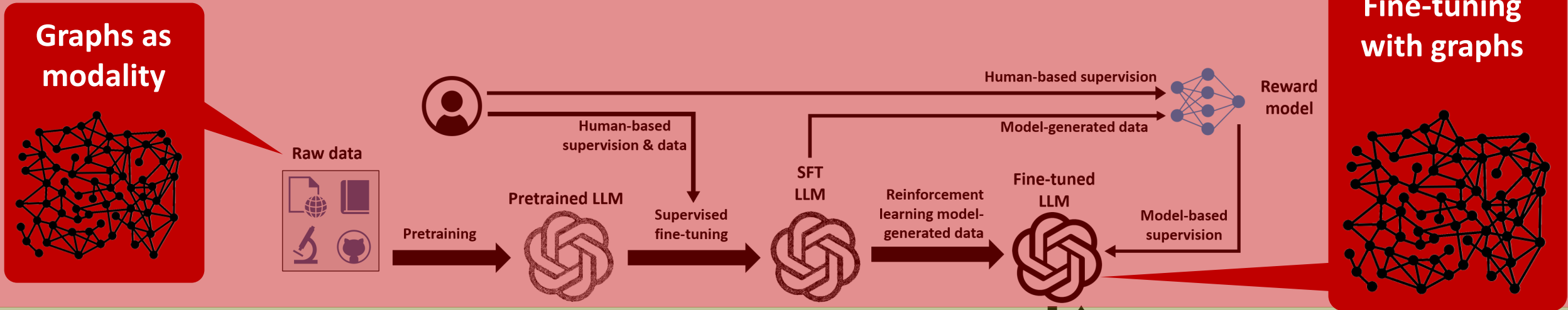
Web

Replies („thoughts”) ↑ Prompts ↓

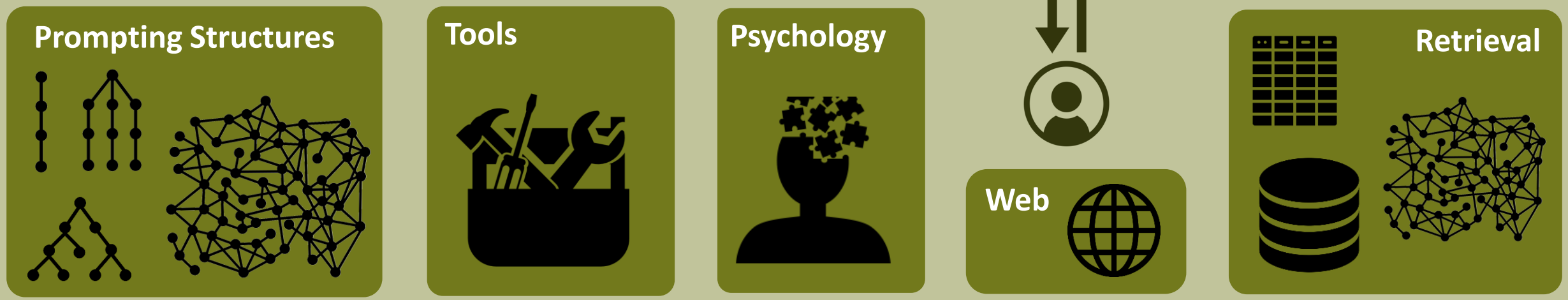
Retrieval

The Emergence of the „Generative AI Ecosystem”: Training

Training related

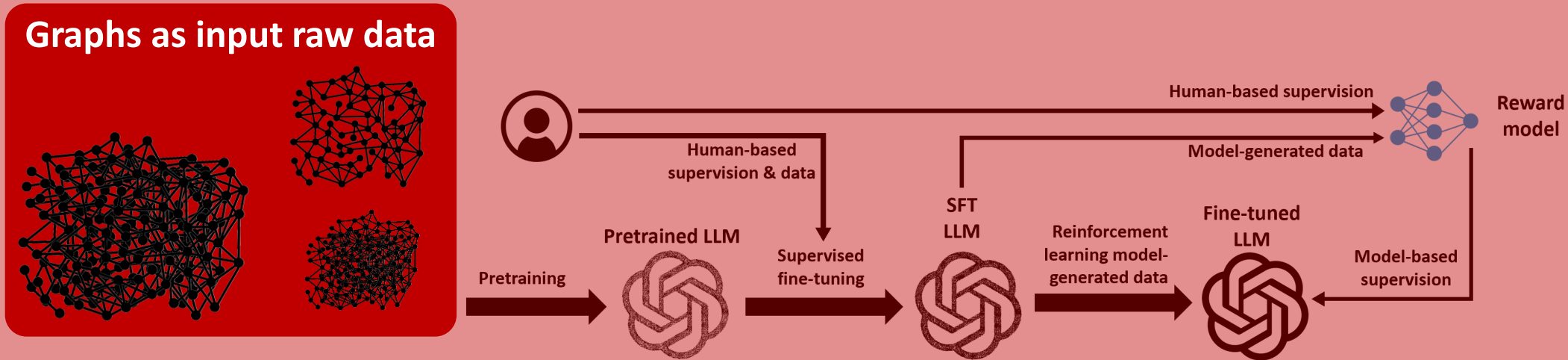


Inference related

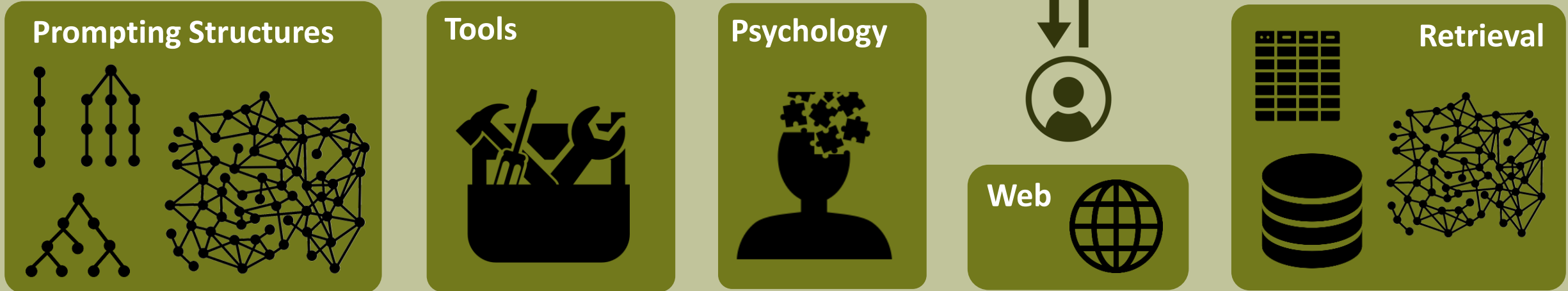


Graph Foundation Models & Graph Prompting

Training related



Inference related



M. BESTA, T. HOEFLER

WITH N. BLACH, A. KUBICEK, R. GERSTENBERGER, AND MANY OTHERS

Graph of Thoughts: Solving Elaborate Problems with Large Language Models



M. BESTA, T. HOEFLER

WITH N. BLACH, A. KUBICEK, R. GERSTENBERGER, AND MANY OTHERS

Graph of Thoughts: Solving Elaborate Problems with Large Language Models

Thank you



Prompting Example: Sorting



Hello. I want to sort the following input sequence of numbers: {input}

Prompting Example: Sorting

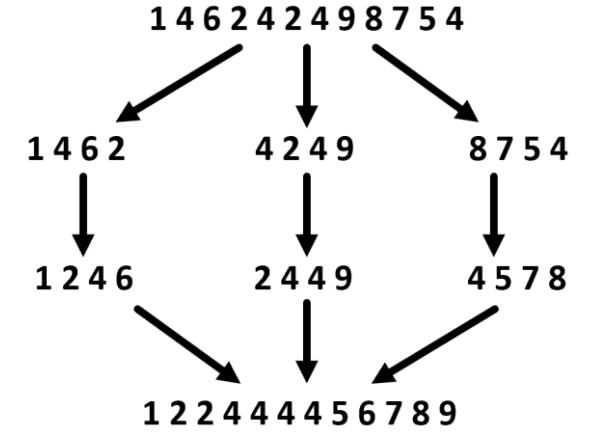


1

<Instruction> Split the following list of 64 numbers into 4 lists of 16 numbers each, the first list should contain the first 16 numbers, the second list the second 16 numbers, the third list the third 16 numbers and the fourth list the fourth 16 numbers. Only output the final 4 lists in the following format without any additional text or thoughts!

```

{
  "List 1": [3, 4, 3, 5, 7, 8, 1, ...],
  "List 2": [2, 9, 2, 4, 7, 1, 5, ...],
  "List 3": [6, 9, 8, 1, 9, 2, 4, ...],
  "List 4": [9, 0, 7, 6, 5, 6, 6, ...]
} </Instruction>
  
```



Prompting Example: Sorting



1

<Instruction> Split the following list of 64 numbers into 4 lists of 16 numbers each, the first list should contain the first 16 numbers, the second list the second 16 numbers, the third list the third 16 numbers and the fourth list the fourth 16 numbers. Only output the final 4 lists in the following format without any additional text or thoughts!

```


    {{
      "List 1": [3, 4, 3, 5, 7, 8, 1, ...],
      "List 2": [2, 9, 2, 4, 7, 1, 5, ...],
      "List 3": [6, 9, 8, 1, 9, 2, 4, ...],
      "List 4": [9, 0, 7, 6, 5, 6, 6, ...]
    }}
  

```

<Example>

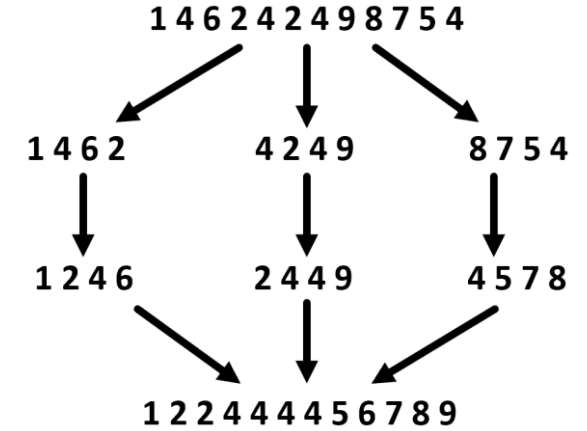
Input: [3, 1, 9, 3, 7, 5, 5, 4, 8, 1, 5, 3, 3, 2, 3, 0, 9, 7, 2, 2, 4, 4, 8, 5, 0, 8, 7, 3, 3, 8, 7, 0, 9, 5, 1, 6, 7, 6, 8, 9, 0, 3, 0, 6, 3, 4, 8, 0, 6, 9, 8, 4, 1, 2, 9, 0, 4, 8, 8, 9, 9, 8, 5, 9]

Output:

```


    {{
      "List 1": [3, 1, 9, 3, 7, 5, 5, 4, 8, 1, 5, 3, 3, 2, 3, 0],
      "List 2": [9, 7, 2, 2, 4, 4, 8, 5, 0, 8, 7, 3, 3, 8, 7, 0],
      "List 3": [9, 5, 1, 6, 7, 6, 8, 9, 0, 3, 0, 6, 3, 4, 8, 0],
      "List 4": [6, 9, 8, 4, 1, 2, 9, 0, 4, 8, 8, 9, 9, 8, 5, 9]
    }}
  

```



Prompting Example: Sorting



1

<Instruction> Split the following list of 64 numbers into 4 lists of 16 numbers each, the first list should contain the first 16 numbers, the second list the second 16 numbers, the third list the third 16 numbers and the fourth list the fourth 16 numbers. Only output the final 4 lists in the following format without any additional text or thoughts!

```


  {{
    "List 1": [3, 4, 3, 5, 7, 8, 1, ...],
    "List 2": [2, 9, 2, 4, 7, 1, 5, ...],
    "List 3": [6, 9, 8, 1, 9, 2, 4, ...],
    "List 4": [9, 0, 7, 6, 5, 6, 6, ...]
  }}
  </Instruction>
  

```

<Example>

Input: [3, 1, 9, 3, 7, 5, 5, 4, 8, 1, 5, 3, 3, 2, 3, 0, 9, 7, 2, 2, 4, 4, 8, 5, 0, 8, 7, 3, 3, 8, 7, 0, 9, 5, 1, 6, 7, 6, 8, 9, 0, 3, 0, 6, 3, 4, 8, 0, 6, 9, 8, 4, 1, 2, 9, 0, 4, 8, 8, 9, 9, 8, 5, 9]

Output:

```


  {{
    "List 1": [3, 1, 9, 3, 7, 5, 5, 4, 8, 1, 5, 3, 3, 2, 3, 0],
    "List 2": [9, 7, 2, 2, 4, 4, 8, 5, 0, 8, 7, 3, 3, 8, 7, 0],
    "List 3": [9, 5, 1, 6, 7, 6, 8, 9, 0, 3, 0, 6, 3, 4, 8, 0],
    "List 4": [6, 9, 8, 4, 1, 2, 9, 0, 4, 8, 8, 9, 9, 8, 5, 9]
  }}
  

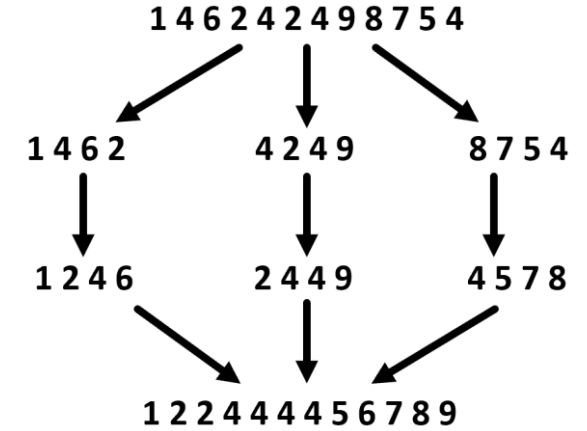
```

}}

</Example>

Input: {input}

The input
thought t



Prompting Example: Sorting



1

<Instruction> Split the following list of 64 numbers into 4 lists of 16 numbers each, the first list should contain the first 16 numbers, the second list the second 16 numbers, the third list the third 16 numbers and the fourth list the fourth 16 numbers. Only output the final 4 lists in the following format without any additional text or thoughts!

```


    {{
      "List 1": [3, 4, 3, 5, 7, 8, 1, ...],
      "List 2": [2, 9, 2, 4, 7, 1, 5, ...],
      "List 3": [6, 9, 8, 1, 9, 2, 4, ...],
      "List 4": [9, 0, 7, 6, 5, 6, 6, ...]
    }}
  

```

<Example>

Input: [3, 1, 9, 3, 7, 5, 5, 4, 8, 1, 5, 3, 3, 2, 3, 0, 9, 7, 2, 2, 4, 4, 8, 5, 0, 8, 7, 3, 3, 8, 7, 0, 9, 5, 1, 6, 7, 6, 8, 9, 0, 3, 0, 6, 3, 4, 8, 0, 6, 9, 8, 4, 1, 2, 9, 0, 4, 8, 8, 9, 9, 8, 5, 9]

Output:

```

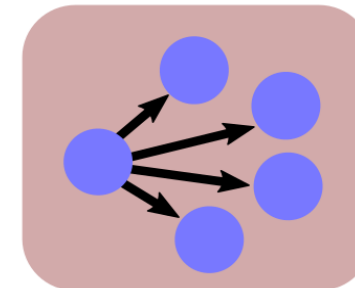
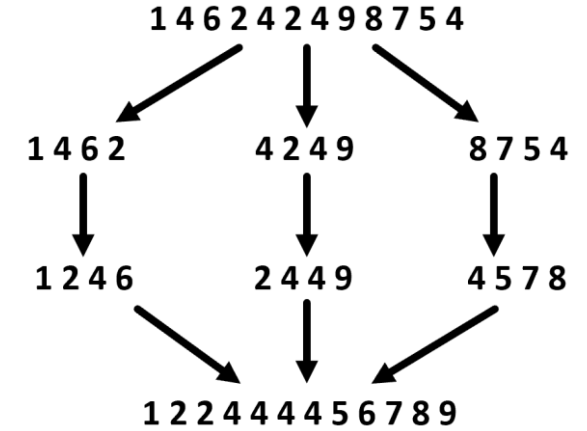

    {{
      "List 1": [3, 1, 9, 3, 7, 5, 5, 4, 8, 1, 5, 3, 3, 2, 3, 0],
      "List 2": [9, 7, 2, 2, 4, 4, 8, 5, 0, 8, 7, 3, 3, 8, 7, 0],
      "List 3": [9, 5, 1, 6, 7, 6, 8, 9, 0, 3, 0, 6, 3, 4, 8, 0],
      "List 4": [6, 9, 8, 4, 1, 2, 9, 0, 4, 8, 8, 9, 9, 8, 5, 9]
    }}
  

```

</Example>

Input: {input}

The input thought t

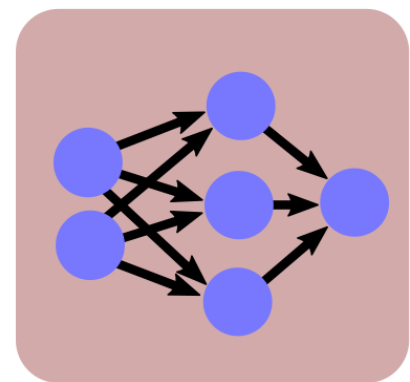
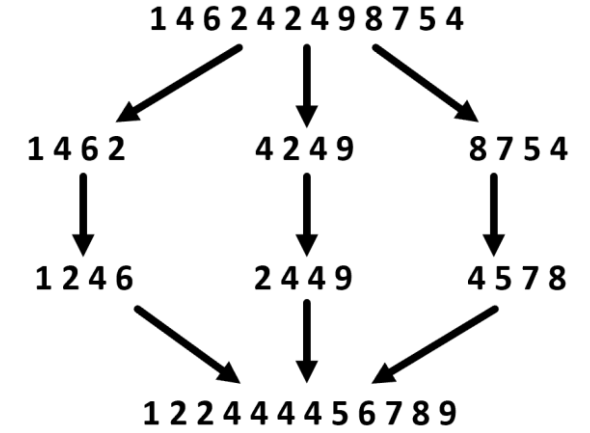


Prompting Example: Sorting

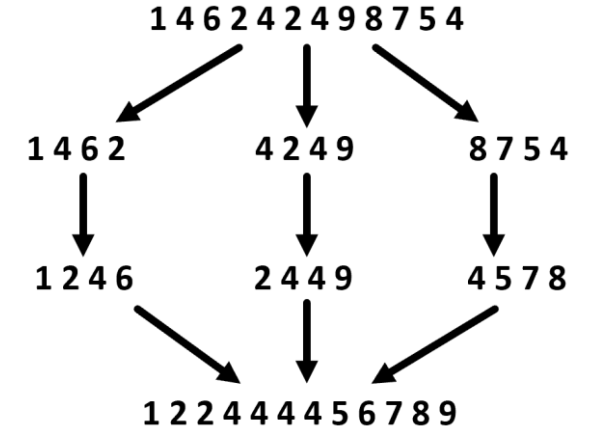
>_

2

*<Instruction> Merge the following 2 sorted lists of length {length1} each, into one sorted list of length {length2} using a merge sort style approach. Only output the final merged list without any additional text or thoughts!
</Instruction>*



Prompting Example: Sorting



>_

3

<Instruction> Sort the following list of numbers in ascending order. Output only the sorted list of numbers, no additional text. **</Instruction>**

<Example>

Input: [3, 7, 0, 2, 8, 1, 2, 2, 2, 4, 7, 8, 5, 5, 3, 9, 4, 3, 5, 6, 6, 4, 4, 5, 2, 0, 9, 3, 3, 9, 2, 1]

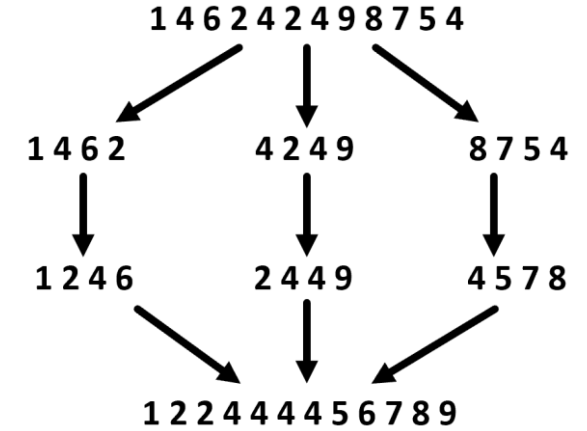
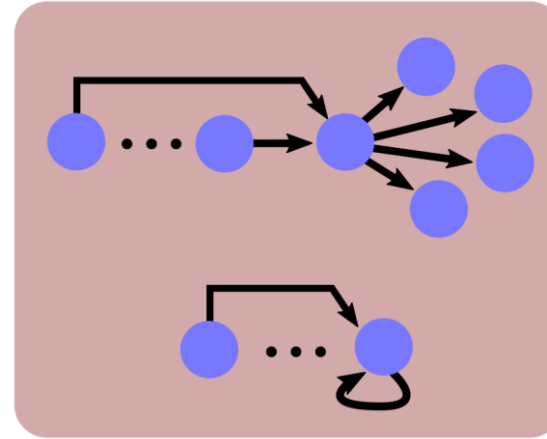
Output: [0, 0, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 9]

</Example>

Input: {input}

The input thought t 

Prompting Example: Sorting



>_

3

<Instruction> Sort the following list of numbers in ascending order.
Output only the sorted list of numbers, no additional text. **</Instruction>**

<Example>

Input: [3, 7, 0, 2, 8, 1, 2, 2, 2, 4, 7, 8, 5, 5, 3, 9, 4, 3, 5, 6, 6, 4, 4, 5,
2, 0, 9, 3, 3, 9, 2, 1]

Output: [0, 0, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5,
6, 6, 7, 7, 8, 8, 9, 9, 9]

</Example>

Input: {input}

The input
thought t

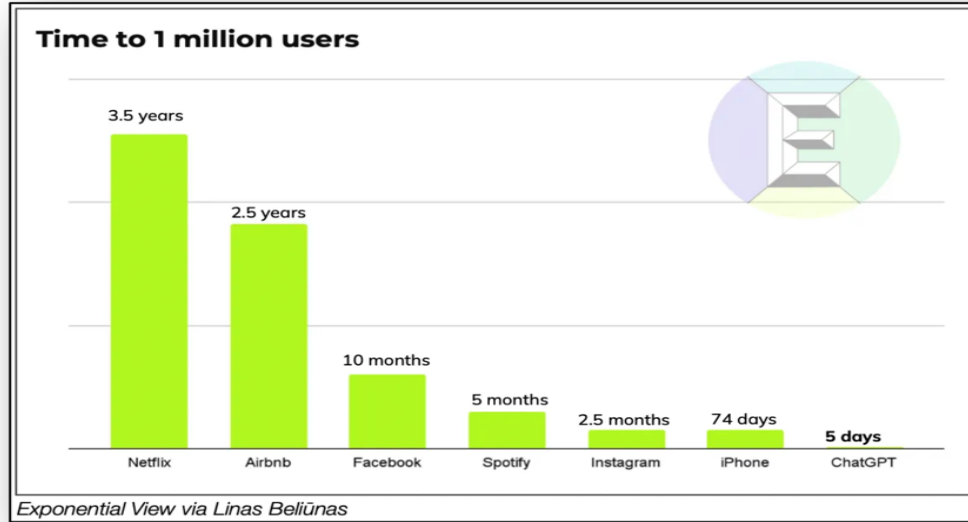


Generative AI Revolution

Generative AI Revolution



Generative AI Revolution



Generative AI Revolution

