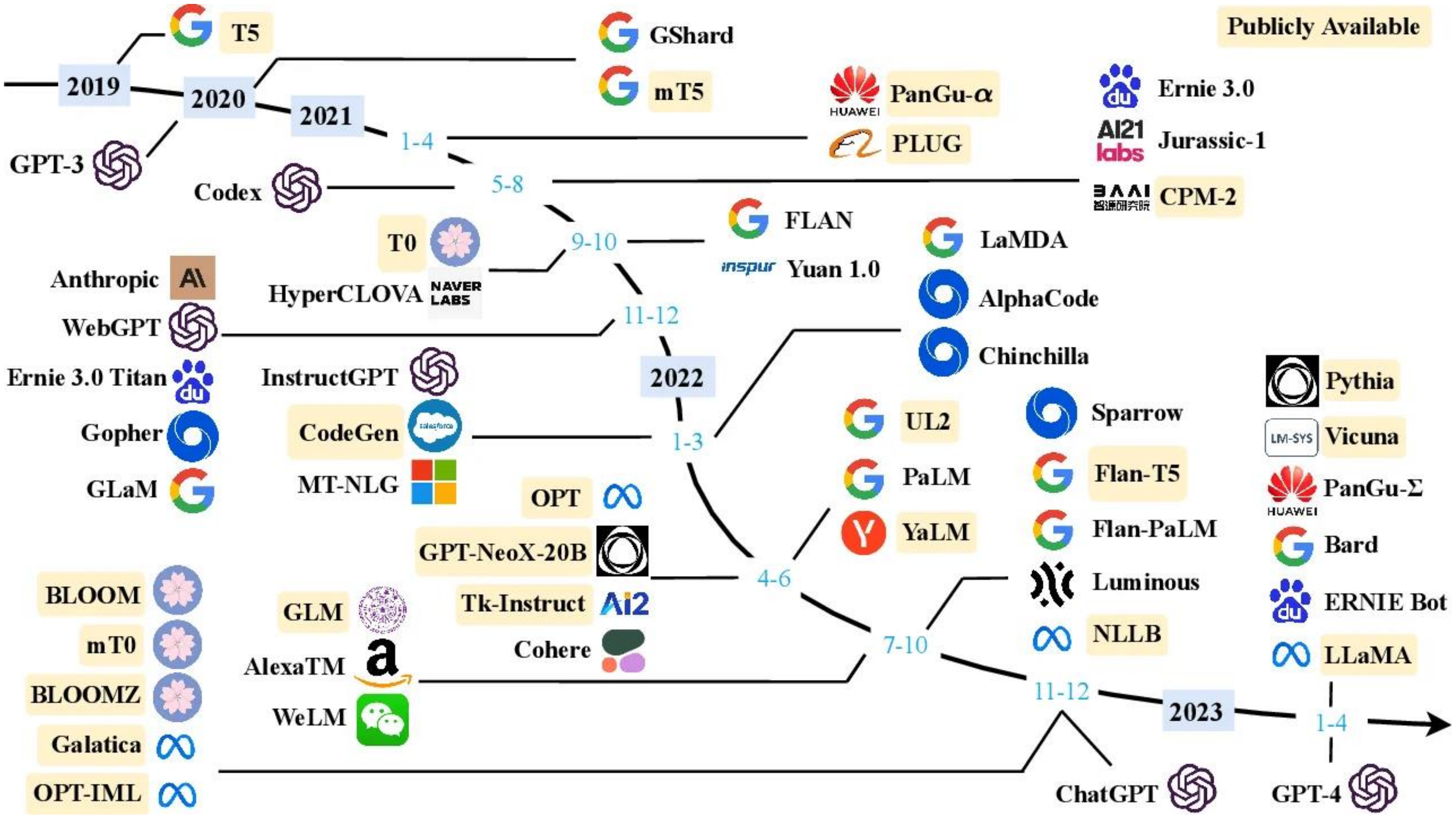
A Primer on Large Language Models

Soham Dan
Microsoft
soham.dan@microsoft.com

Outline of the Presentation

- Language Models
- Types of Large Language Models
- Transformer Architecture
- Typical LLM Workflow
- In-Context Learning
- *Reasoning* by LLMs
- Tool Usage by LLMs
- LLM Agents

Explosion of Large Language Models



Publicly Available

2019
T5
GPT-3

2020
GShard
mT5

2021
1-4
PanGu-α
PLUG

5-8
Codex
Ernie 3.0
Jurassic-1
CPM-2

9-10
T0
HyperCLOVA
FLAN
Yuan 1.0

11-12
Anthropic
WebGPT
Ernie 3.0 Titan
Gopher
GLaM

2022
1-3
InstructGPT
CodeGen
MT-NLG
LaMDA
AlphaCode
Chinchilla

4-6
OPT
GPT-NeoX-20B
Tk-Instruct
Cohere
UL2
PaLM
YaLM

7-10
GLM
AlexaTM
WeLM
Sparrow
Flan-T5
Flan-PaLM
Luminous
NLLB

11-12
BLOOM
mT0
BLOOMZ
Galatica
OPT-IML
ChatGPT

2023
1-4
Pythia
Vicuna
PanGu-Σ
Bard
ERNIE Bot
LLaMA
GPT-4

What are Language Models ?

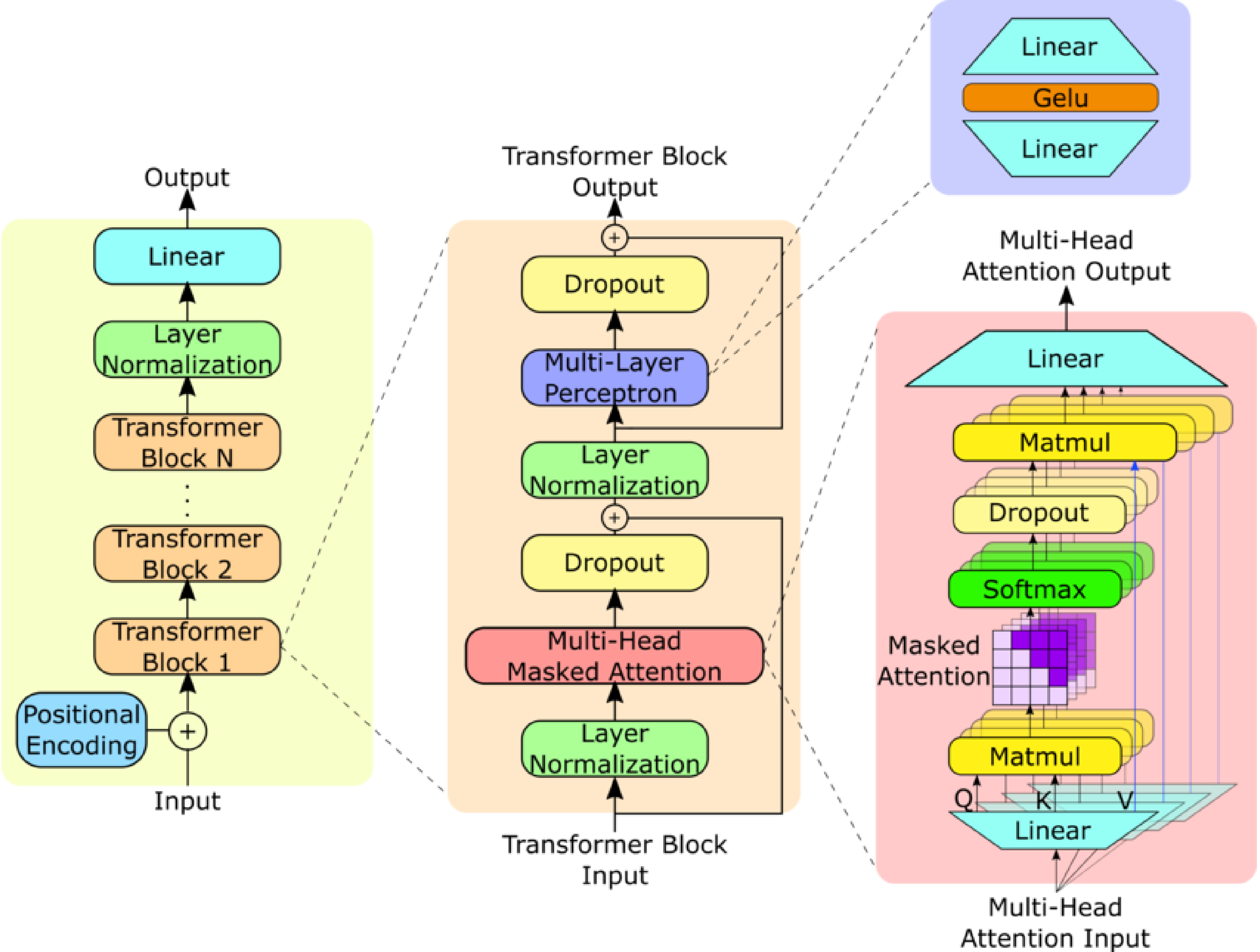
Language Modeling: How *likely* is the sequence $w_1 w_2 \ldots w_T$?

A language model estimates the probability $P(w_1, w_2, \ldots, w_T)$ of a sequence by factorizing it as:

$$P(w_1, w_2, \ldots, w_T) = \prod_t P\left(w_t \mid w_{\{<t\}}\right)$$

- Various ways to model this conditional probability: n-gram models, feed-forward neural networks, recurrent neural networks, transformers.

Transformer Architecture



- **Input Embedding:** Converts discrete tokens into continuous vector representations.

- **Positional Encoding:** Since transformers lack a natural notion of sequence order, positional encodings are added.

- **Self-Attention Mechanism:** Computes a weighted sum of all token representations in the sequence, where the weights are determined by the similarity between tokens.

- **Output Projection:** After processing through the stacked layers, a final linear projection maps the model's output to a probability distribution over the vocabulary.

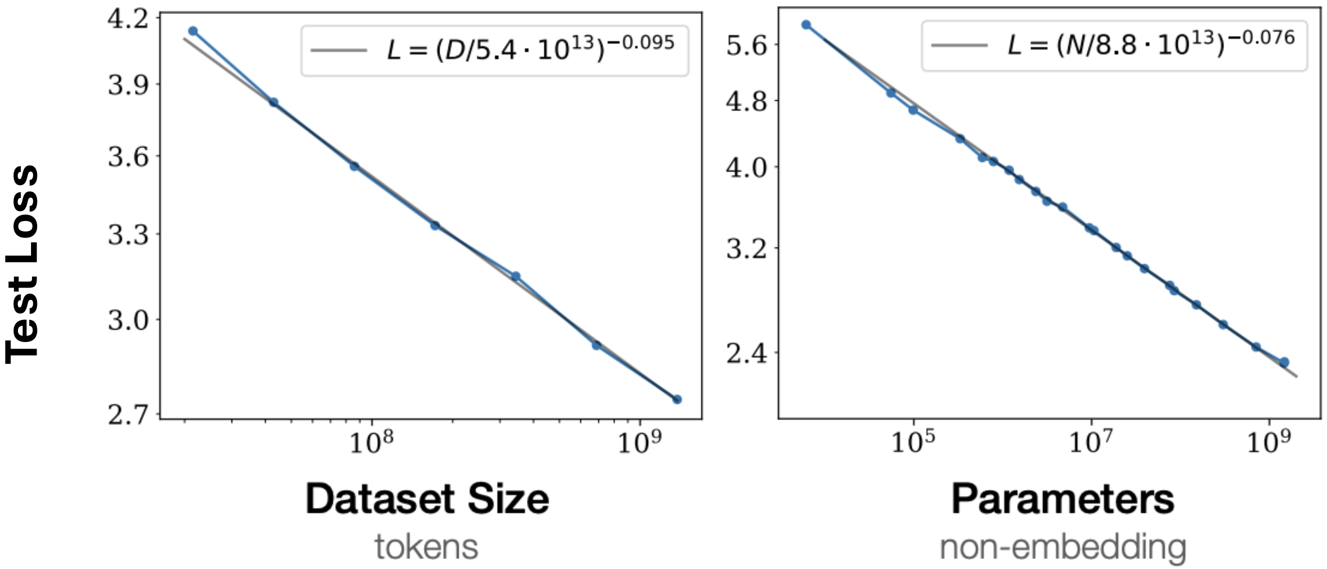
Attention is All You Need, Vaswani et al, 2017.

Large Language Models

- Decoder-only LLMs: Decode (Generate) words one at a time for text generation.

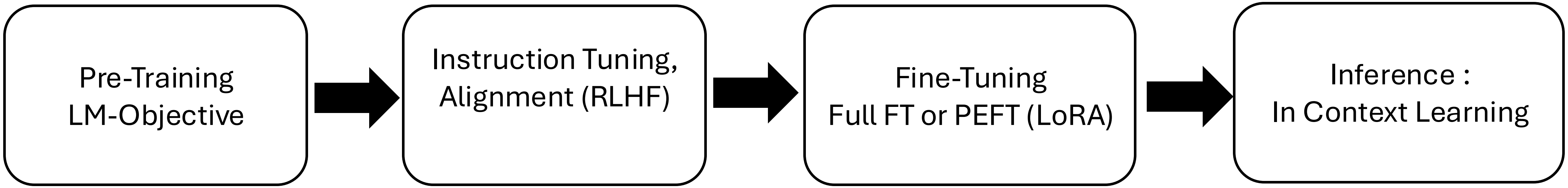
Examples: GPT-family, most recent LLMs

- Basic building block is the Transformer



Scaling Laws for Neural Language Models, Kaplan et al. 2020

Large Language Model Workflow

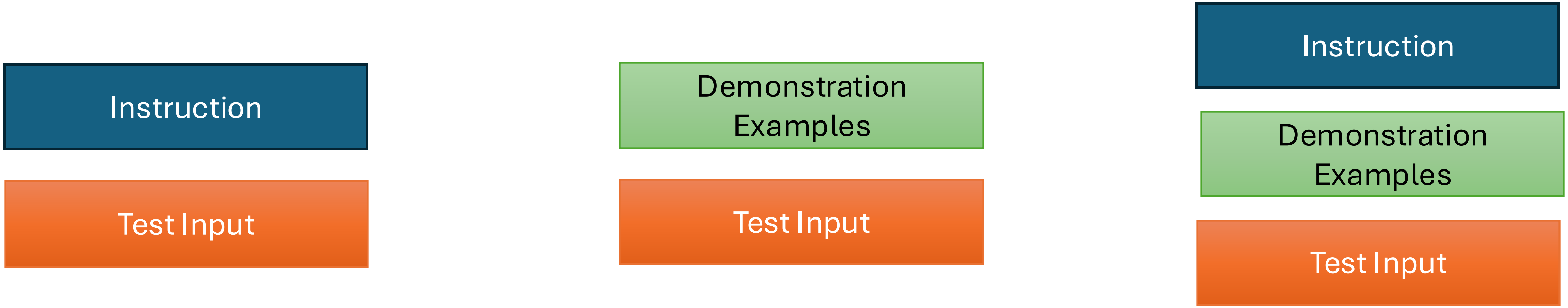


Pre-Training
LM-Objective

Instruction Tuning,
Alignment (RLHF)

Fine-Tuning
Full FT or PEFT (LoRA)

Inference :
In Context Learning



Instruction

Test Input

Zero Shot
Instruction

Demonstration
Examples

Test Input

Few Shot

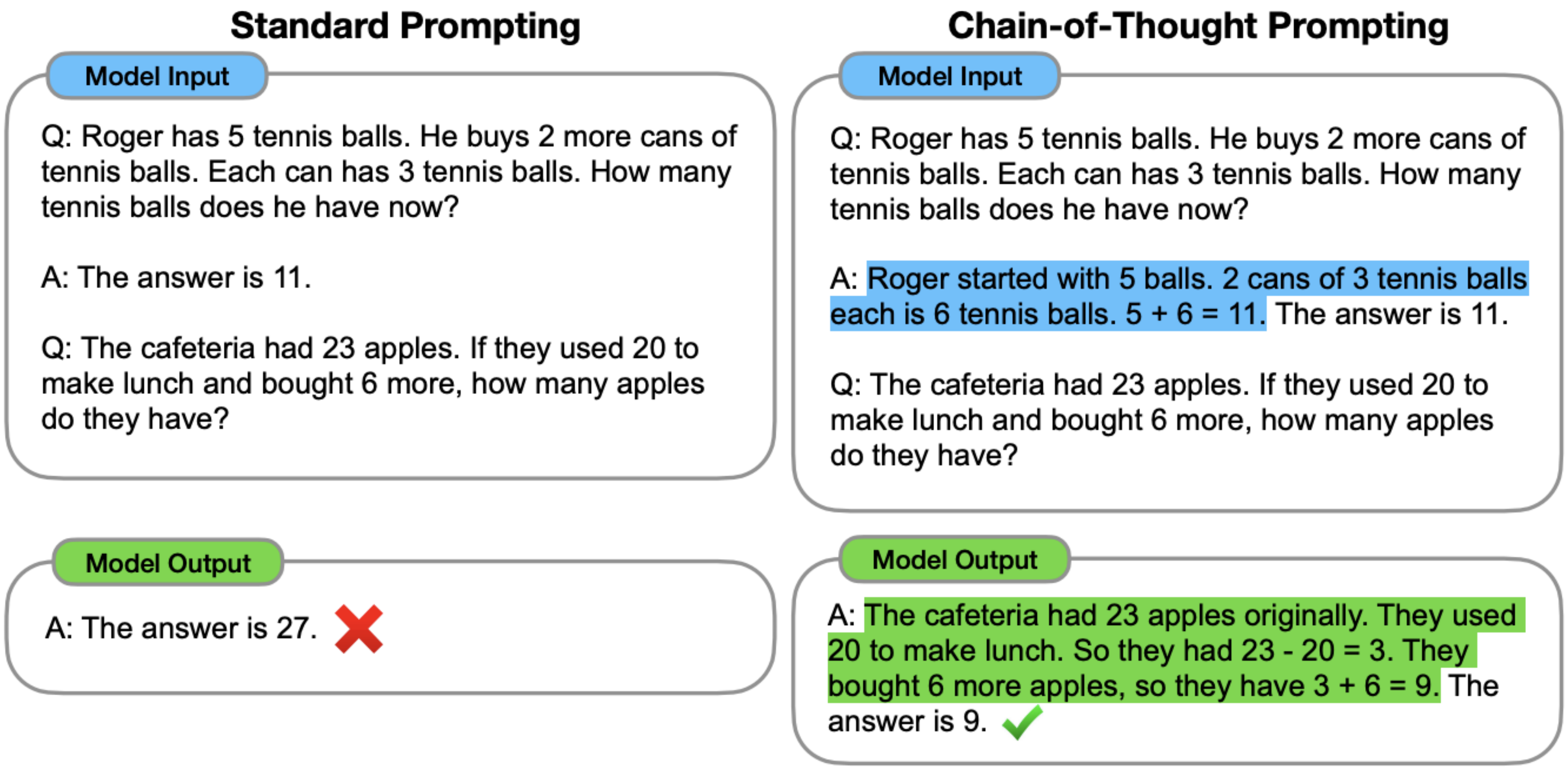
Instruction

Demonstration
Examples

Test Input

Few Shot with
Instruction

Basic Reasoning : Chain of Thought Prompting



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?

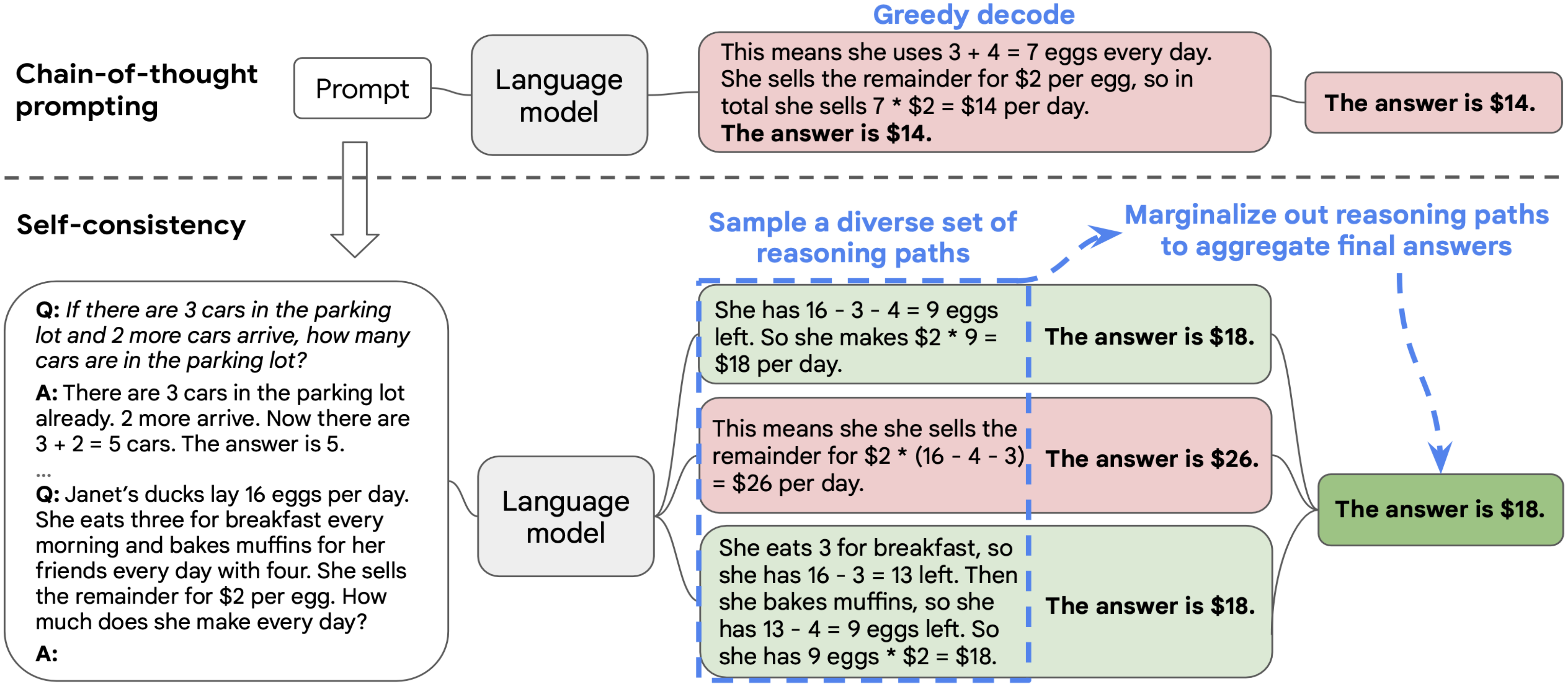
Model Output

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. ✓

- Chain of Thought
- multiple Chain of Thought
- Tree of Thought
- Graph of Thought ...

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Wei et al, 2022.

Self Consistencv on CoT



Chain-of-thought prompting

Prompt

Language model

Greedy decode

This means she uses 3 + 4 = 7 eggs every day. She sells the remainder for $2 per egg, so in total she sells 7 * $2 = $14 per day. **The answer is $14.**

The answer is $14.

Self-consistency

Q: *If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?*

A: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5.

...

Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for $2 per egg. How much does she make every day?

A:

Language model

Sample a diverse set of reasoning paths

She has 16 - 3 - 4 = 9 eggs left. So she makes $2 * 9 = $18 per day. **The answer is $18.**

This means she she sells the remainder for $2 * (16 - 4 - 3) = $26 per day. **The answer is $26.**

She eats 3 for breakfast, so she has 16 - 3 = 13 left. Then she bakes muffins, so she has 13 - 4 = 9 eggs left. So she has 9 eggs * $2 = $18. **The answer is $18.**

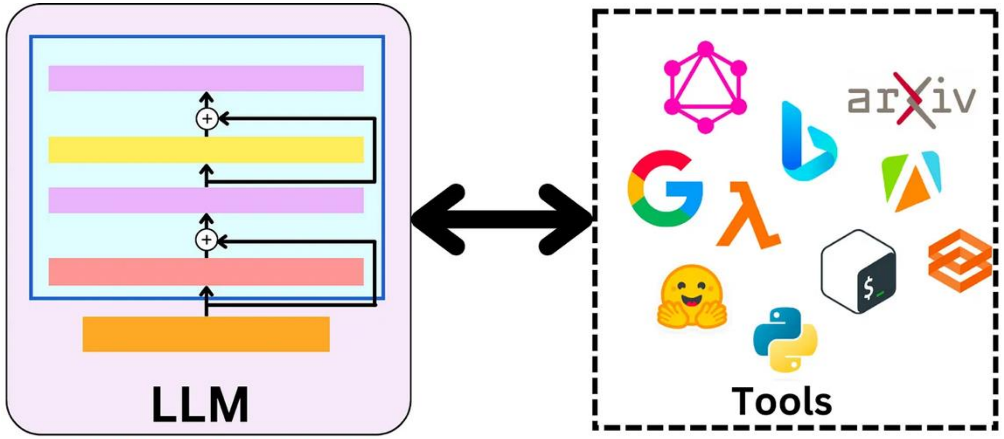
Marginalize out reasoning paths to aggregate final answers

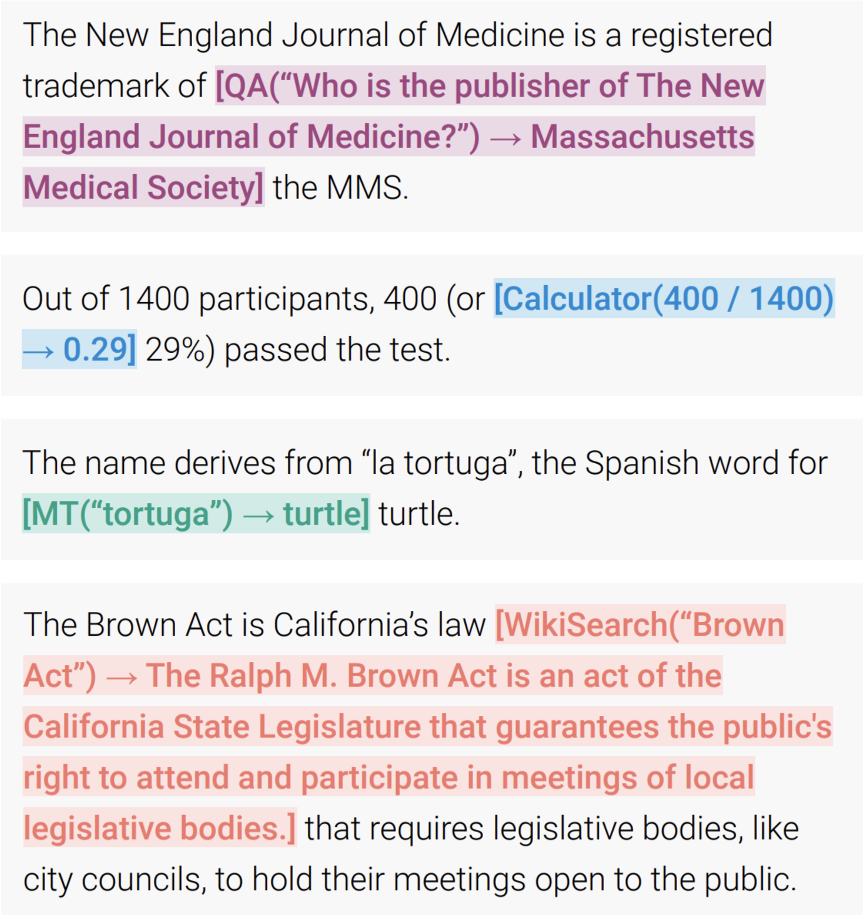
The answer is $18.

Chain of Thoughts can be used to fine-tune LLMs (eg: DeepSeek-V3 fine-tuned on synthetic CoTs)

Self-consistency improves chain of thought reasoning in language models, Wang et al, 2023

Tool Usage by LLMs




The New England Journal of Medicine is a registered trademark of **[QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society]** the MMS.

Out of 1400 participants, 400 (or **[Calculator(400 / 1400) → 0.29]** 29%) passed the test.

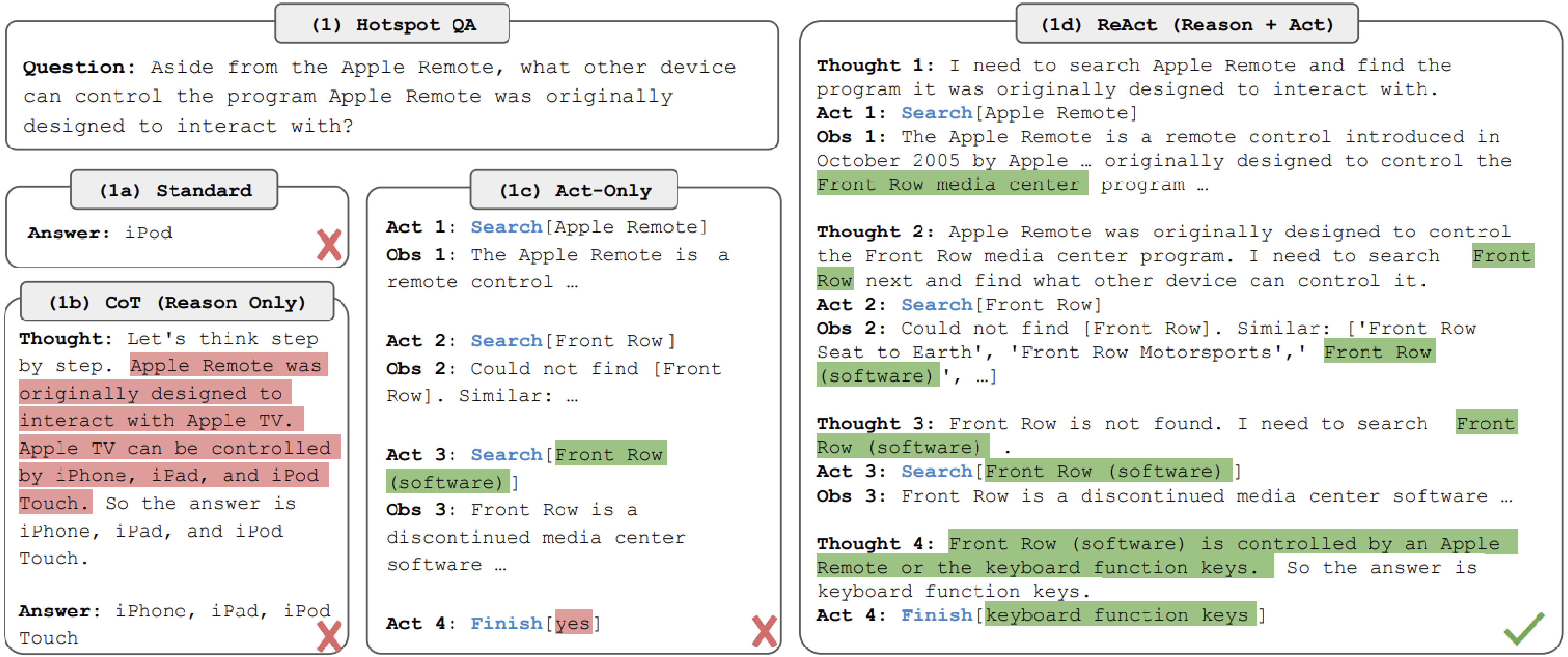
The name derives from "la tortuga", the Spanish word for **[MT("tortuga") → turtle]** turtle.

The Brown Act is California's law **[WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.]** that requires legislative bodies, like city councils, to hold their meetings open to the public.

RAG: **Retrieval Augmented Generation** involves a Retriever model to retrieve document(s) relevant to a query and a Generator model to generate the answer conditioned on the query and retrieved document.

Toolformer: Language Models Can Teach Themselves to Use Tool, Schick et al, 2023.

Agents: Interleaved Tool Use and Planning



(1) Hotspot QA

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1a) Standard

Answer: iPod ✗

(1b) CoT (Reason Only)

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch ✗

(1c) Act-Only

Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control …

Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: …

Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Act 4: Finish[yes] ✗

(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple … originally designed to control the Front Row media center program …

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports',' Front Row (software) ', …]

Thought 3: Front Row is not found. I need to search Front Row (software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: Finish[keyboard function keys] ✓

React, React + Reflexion, Plan and Solve ...

REACT: Synergizing Reasoning and Acting in Language Models, Yao et al, 2023.

Conclusion

- At large scales interesting emergent behavior of LLMs to do in-context learning, chain-of-thought reasoning and tool usage.

- *Soft reasoning* abilities combined with tool usage results in agents useful for various language and multimodal applications.