NEURAL INFORMATION PROCESSING SYSTEMS

MIX LAB

UC San Diego

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

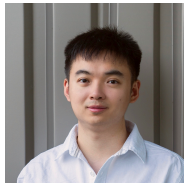
ToolkenGPT

Augmenting Frozen Language Models with Massive Tools via Tool Embeddings

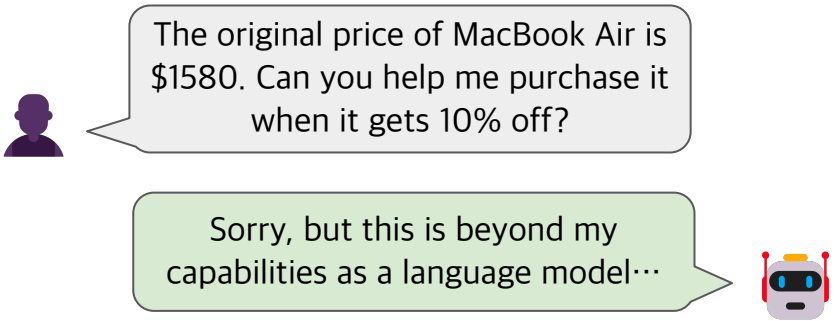

Shibo Hao


Tianyang Liu


Zhen Wang

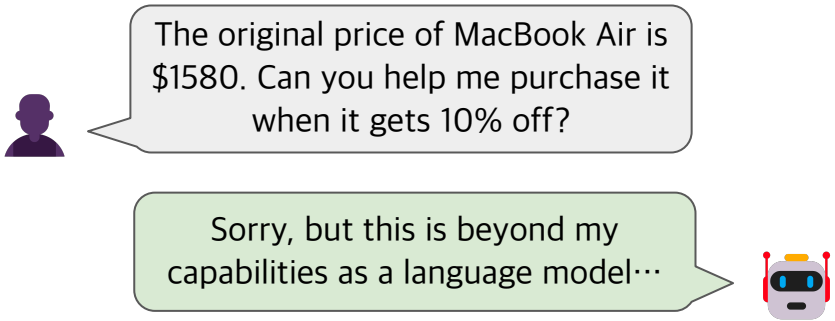

Zhiting Hu

LLMs fail on complex real-world tasks



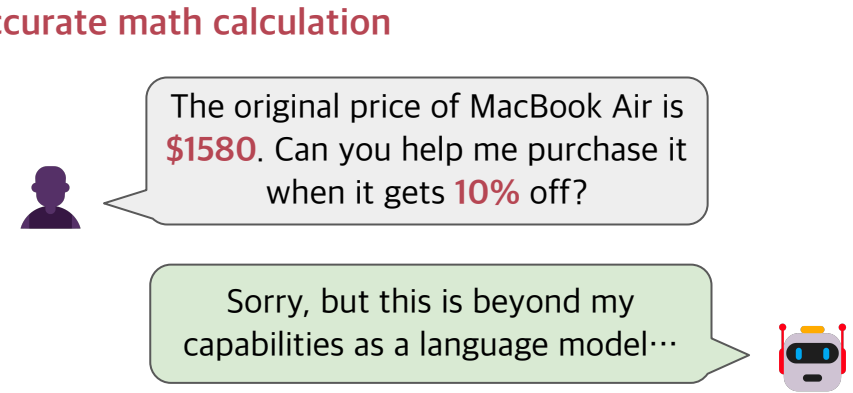
LLMs fail on complex real-world tasks

Lacking the abilities for



LLMs fail on complex real-world tasks

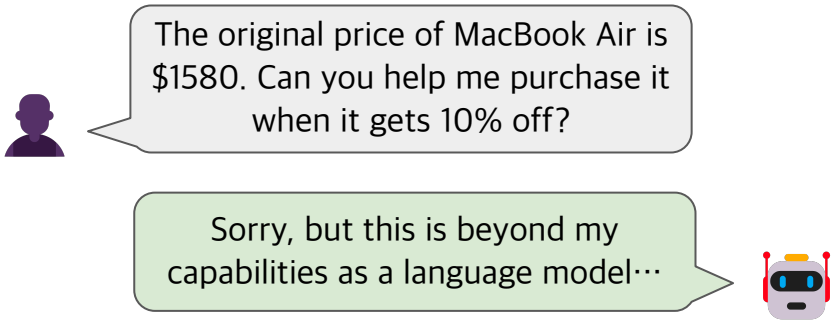
Lacking the abilities for



LLMs fail on complex real-world tasks

Lacking the abilities for

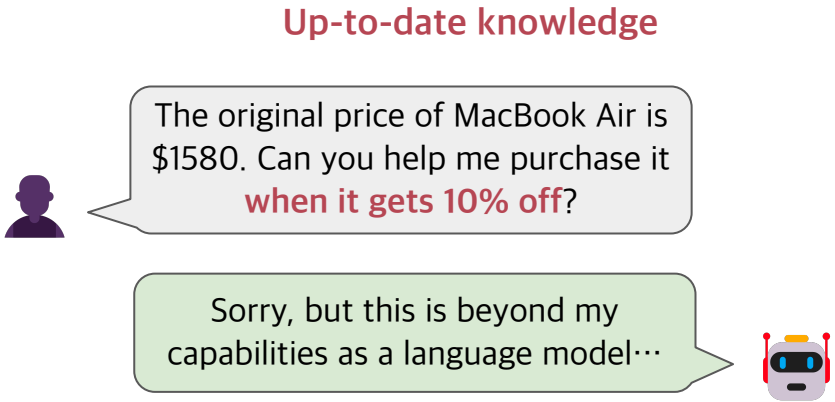
- Accurate math calculation



LLMs fail on complex real-world tasks

Lacking the abilities for

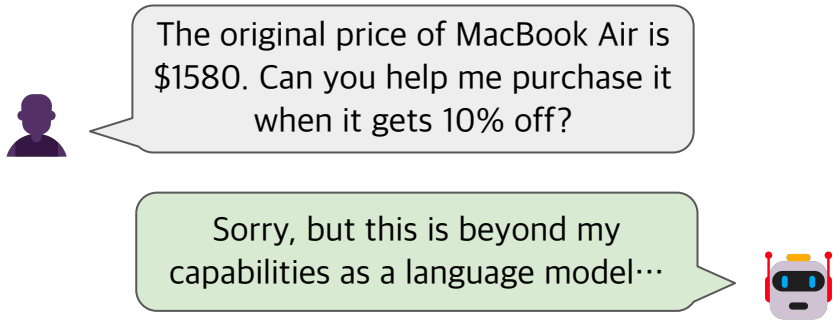
- Accurate math calculation



LLMs fail on complex real-world tasks

Lacking the abilities for

- Accurate math calculation
- Accessing up-to-date knowledge

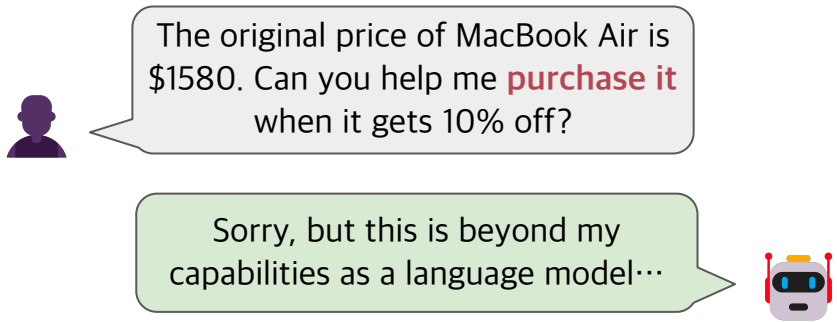


LLMs fail on complex real-world tasks

Lacking the abilities for

- Accurate math calculation
- Accessing up-to-date knowledge

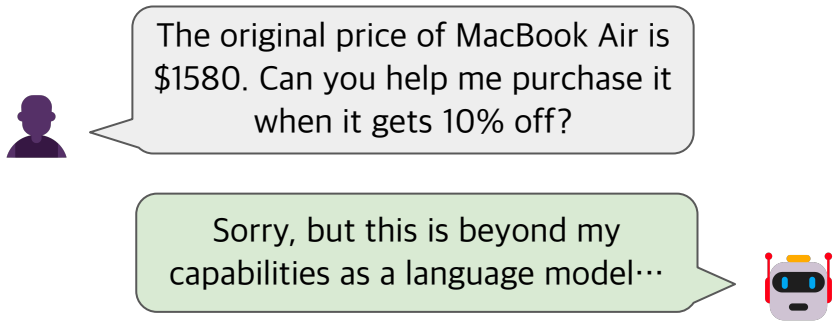
Real-world actions



LLMs fail on complex real-world tasks

Lacking the abilities for

- Accurate math calculation
- Accessing up-to-date knowledge
- Taking real-world actions



How to extend the abilities of LLMs?

Augmenting language models with tools will help unlock those abilities!

- Accurate math calculation
- Accessing up-to-date knowledge
- Taking real-world actions
- …


Calculator


Database


API/Robot

The original price of MacBook Air is $1580. Can you help me purchase it when it gets 10% off?

Sorry, but this is beyond my capabilities as a language model…

Teaching LLMs to Use Tools

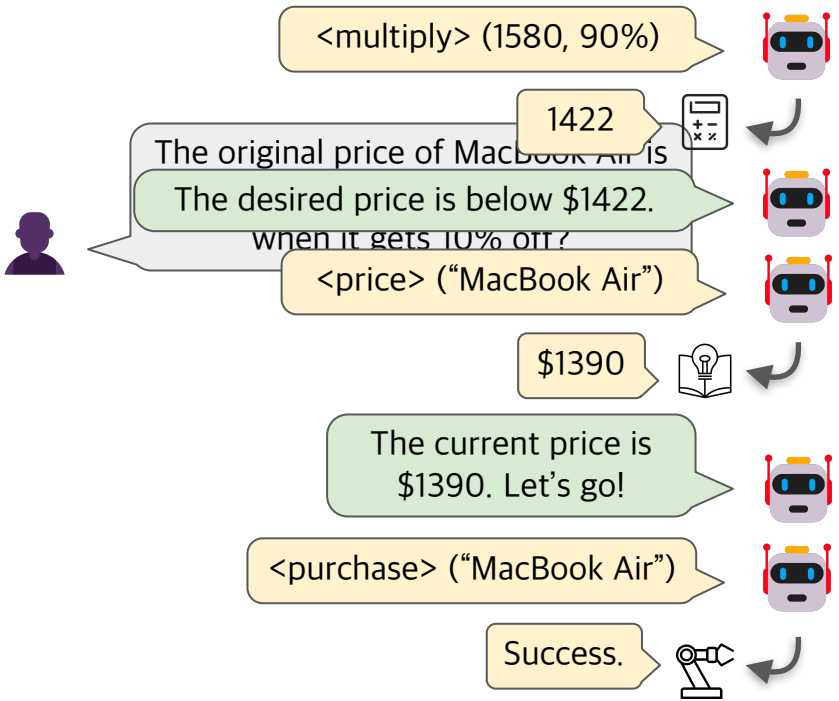
Augmenting language models with tools will help unlock those abilities!

- Accurate math calculation
- Accessing up-to-date knowledge
- Taking real-world actions
- …

Calculator

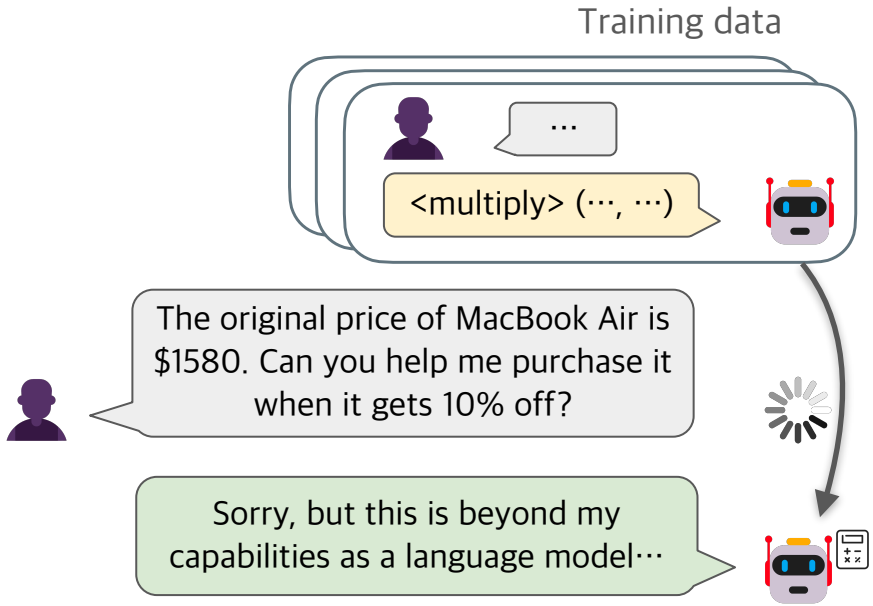
Database

API/Robot



Previous method #1: **Fine-tuning**

Train the LLM with the demonstrations of tool calling



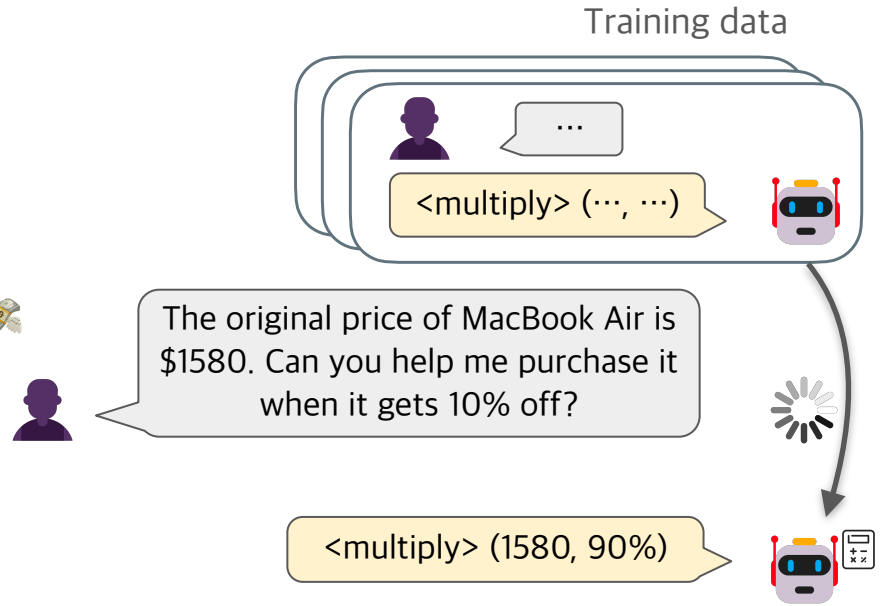
Talm: Tool augmented language models [Parisi et al., 2022]
Toolformer: Language models can teach themselves to use tools [Schick et al., 2023]

Previous method #1: **Fine-tuning**

Train the LLM with the demonstrations of tool calling

But …

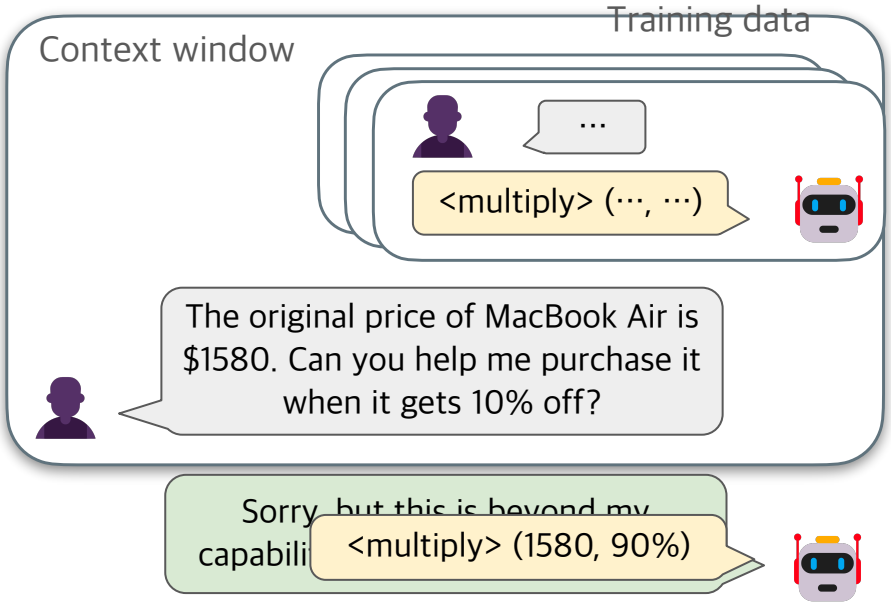
- **Not Frozen LLMs**: Fine-tuning an LLM is expensive 💸
- **Not Plug-and-play**: Once we want to add, delete or update a tool, the LLM needs to be **re-trained** 🔄



Talm: Tool augmented language models [Parisi et al., 2022]
Toolformer: Language models can teach themselves to use tools [Schick et al., 2023]

Previous method #2: **In-context Learning**

Prompting LLMs with demonstrations of tool calling



But …

- **Shallow Understanding**: Can only learn from surface text instead of large-scale data 🤔
- **Limited tools**: struggles with a large tool set 🧰

ReAct: Synergizing Reasoning and Acting in Language Models [Yao et al., 2023]
Gorilla: Large language model connected with massive apis [Patil et al., 2023]

Teaching LLMs to Use Tools

Is there a method to overcome all the limitations mentioned above?

	Fine-tuning	In-context learning
• **Frozen LMs**: No need to fine-tune the LLM	✗	✓
• **Massive Tools**: Work well with a large tool set	✓	✗
• **Plug & Play**: Flexible to add / delete / update a tool	✗	✓
• **Deep Understanding**: Learn better with more training data	✓	✗

Teaching LLMs to Use Tools

Is there a method to overcome all the limitations mentioned above?

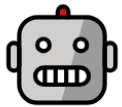
	Fine-tuning	In-context learning	ToolkenGPT
• **Frozen LMs**: No need to fine-tune the LLM	✗	✓	✓
• **Massive Tools**: Work well with a large tool set	✓	✗	✓
• **Plug & Play**: Flexible to add / delete / update a tool	✗	✓	✓
• **Deep Understanding**: Learn better with more training data	✓	✗	✓

We propose  ToolkenGPT to tackle these challenges

Background: Next Token Prediction

Recall how a standard LLM predicts the next token…

LLM



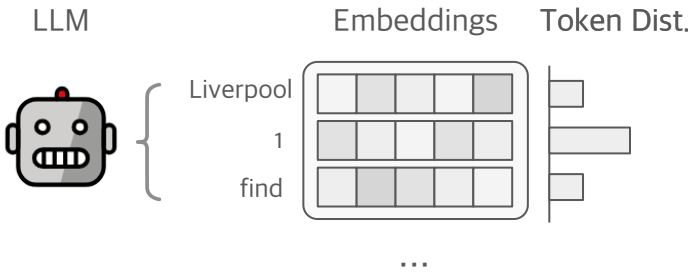
Example: Solving a math word problem

Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is ____

Background: Next Token Prediction

Recall how a standard LLM predicts the next token…



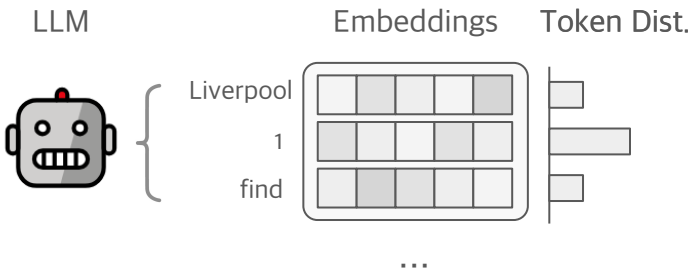
Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is ____

Background: Next Token Prediction

Recall how a standard LLM predicts the next token…

What if we have the embeddings of ⚙️ tools?



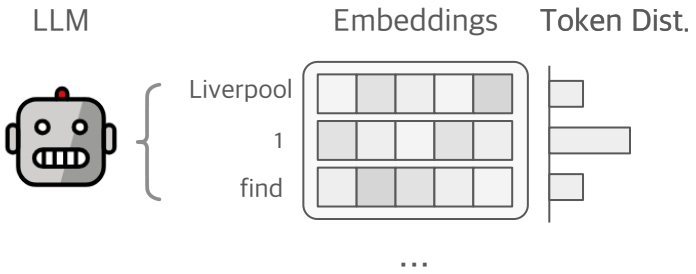
Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is ____

Background: Next Token Prediction

Recall how a standard LLM predicts the next token…

What if we have the embeddings of ⚙️ tools?

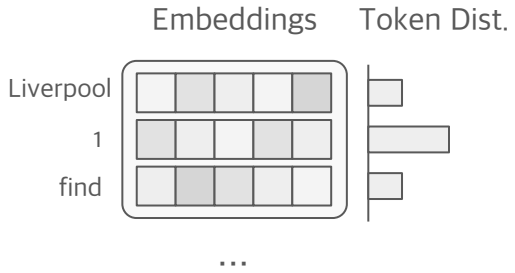


"Tool as token"

Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is ____

Step 1: Next token/toolken prediction

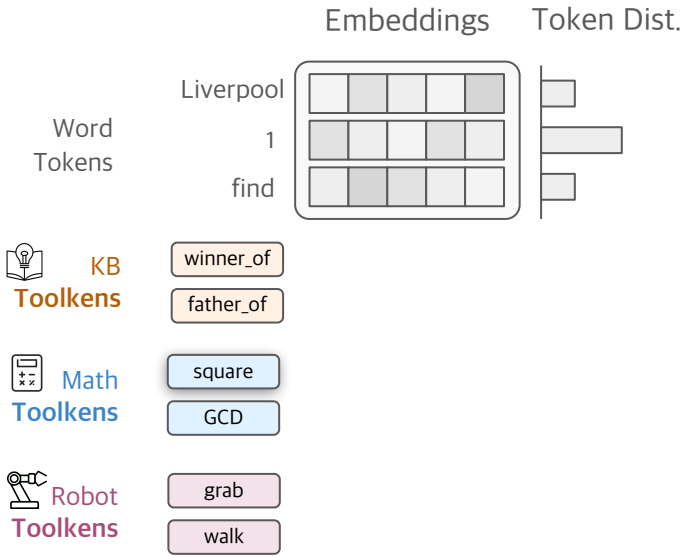


Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is ____

Step 1: Next token/toolken prediction

Adding **Toolkens** to the vocabulary

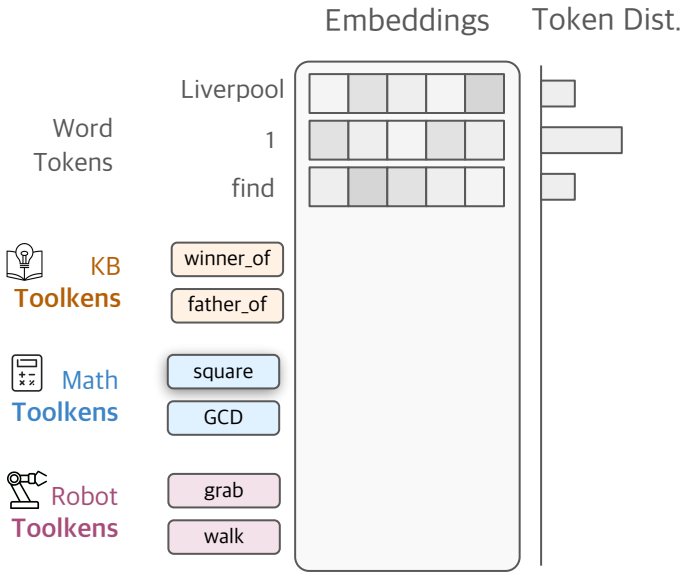


Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is ____

Step 1: Next token/toolken prediction

Adding **Toolkens** to the vocabulary

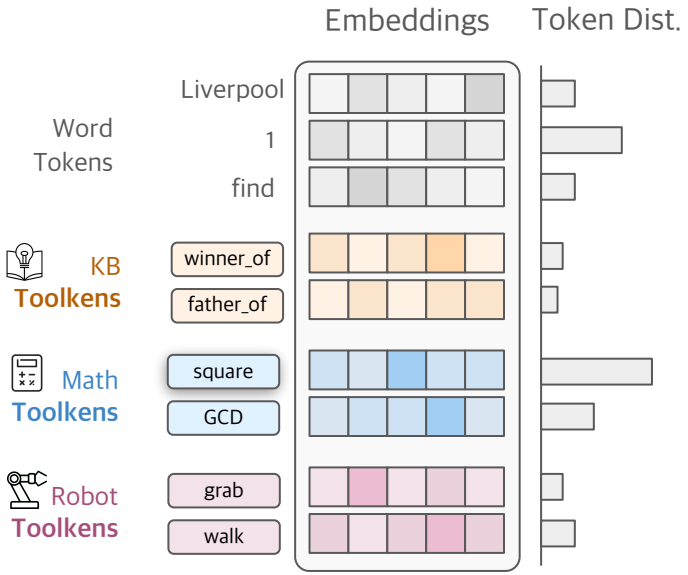


Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is ____

Step 1: Next token/toolken prediction

Adding **Toolkens** to the vocabulary

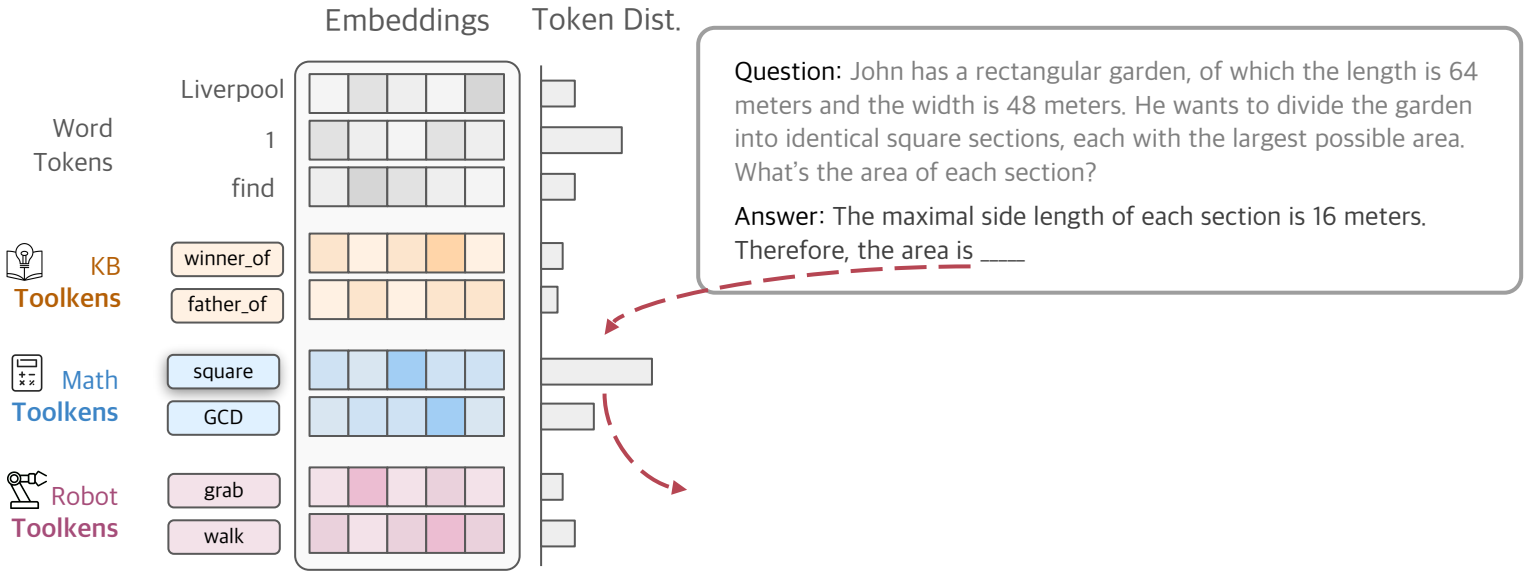


Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is ____

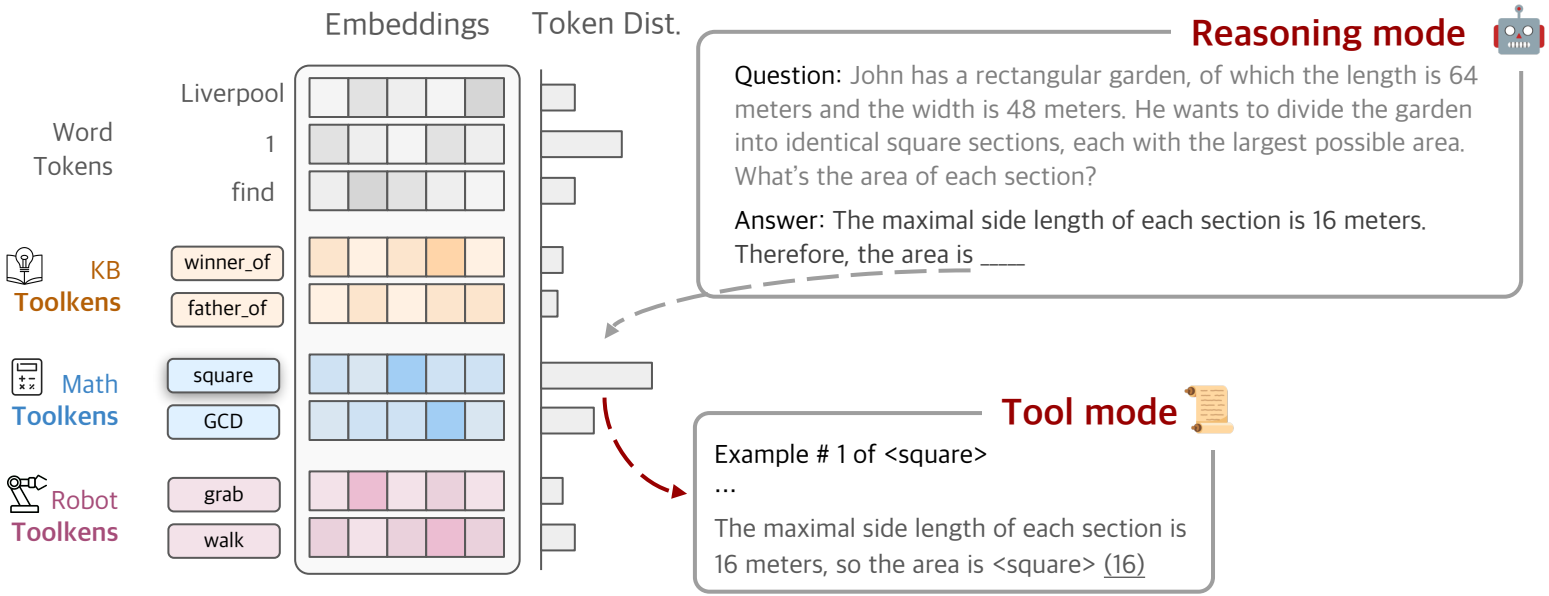
Step 1: Next token/toolken prediction

Adding **Toolkens** to the vocabulary



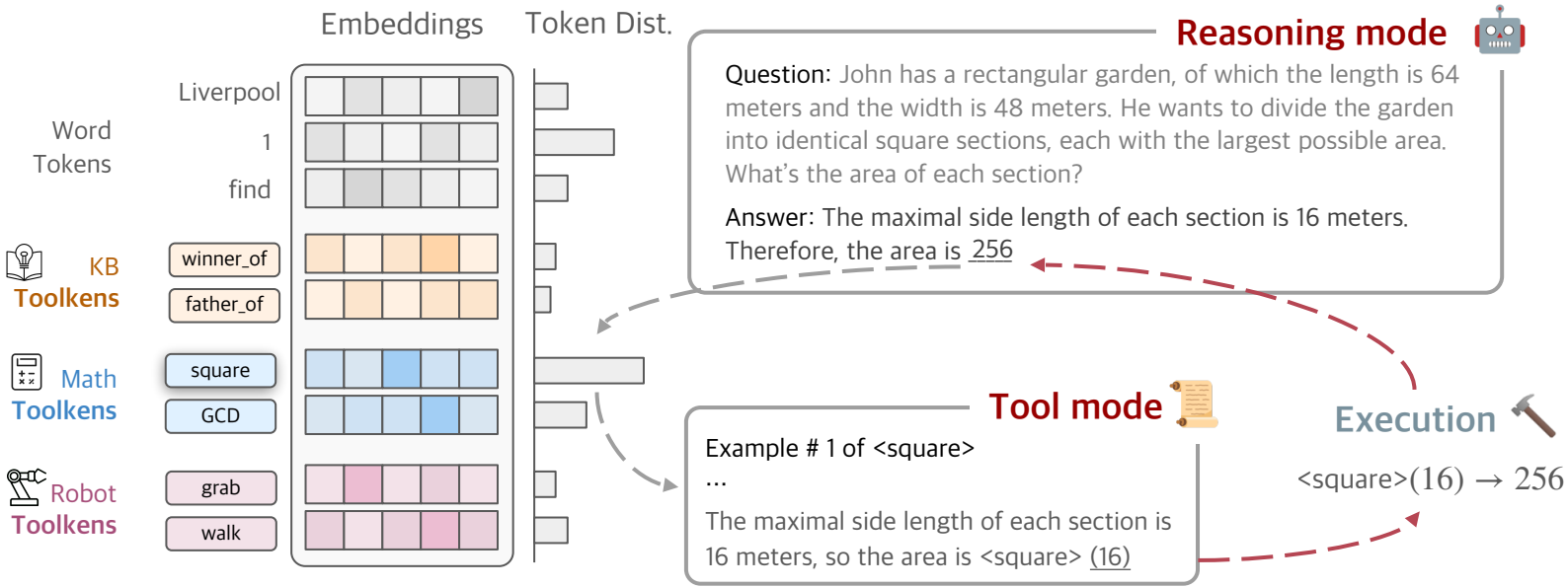
Step 2: Argument prediction in a separate tool mode

Generating arguments with **in-context learning**



Step 3: Execute the tool call and return the result

Finally, the tool call is **executed** and the result is **sent back** to the reasoning mode



Training toolken embedding – Objective

Training objective: Next token / toolken prediction

Input sequence s	The	area	is	2	5	6	square	feet
Target sequence s'	The	area	is	<square>	[mask]	[mask]	square	feet

Training toolken embedding – Objective

Training objective: Next token / toolken prediction

Input sequence s	The	area	is	2	5	6	square	feet
Target sequence s'	The	area	is	<square>	[mask]	[mask]	square	feet

Training toolken embedding – Objective

Training objective: Next token / toolken prediction

Input sequence s	The	area	is	2	5	6	square	feet
Target sequence s'	The	area	is	<square>	[mask]	[mask]	square	feet

Training toolken embedding – Objective

Training objective: Next token / toolken prediction

Input sequence s	The	area	is	2	5	6	square	feet
Target sequence s'	The	area	is	<square>	[mask]	[mask]	square	feet

Training toolken embedding – Objective

Training objective: Next token / toolken prediction

Input sequence s	The	area	is	2	5	6	square	feet
Target sequence s'	The	area	is	<square>	[mask]	[mask]	square	feet

Training toolken embedding – Objective

Training objective: Next token / toolken prediction

Input sequence s	The	area	is	2	5	6	square	feet
Target sequence s'	The	area	is	<square>	[mask]	[mask]	square	feet

Training toolken embedding – Objective

Training objective: Next token / toolken prediction

Input sequence s	The	area	is	2	5	6	square	feet
Target sequence s'	The	area	is	<square>	[mask]	[mask]	square	feet

Training toolken embedding – Objective

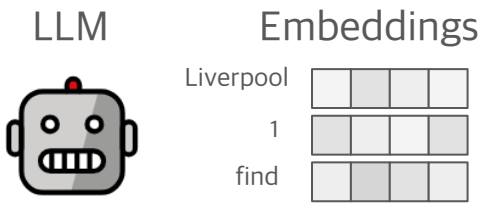
Training objective: Next token / toolken prediction

Input sequence s	The	area	is	2	5	6	square	feet
Target sequence s'	The	area	is	<square>	[mask]	[mask]	square	feet

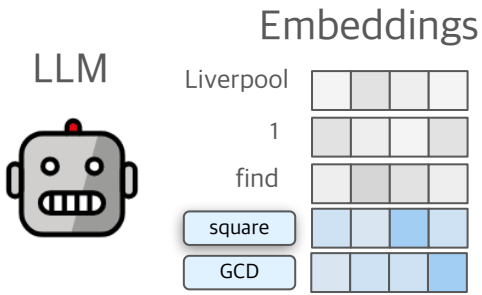
Training Data:
- Demonstration data
- **Synthetic data**

Self-instruct: Aligning language model with self generated instructions. [Wang et al., 2022]
Toolformer: Language models can teach themselves to use tools [Schick et al., 2023]

Training toolken embedding – Optimization

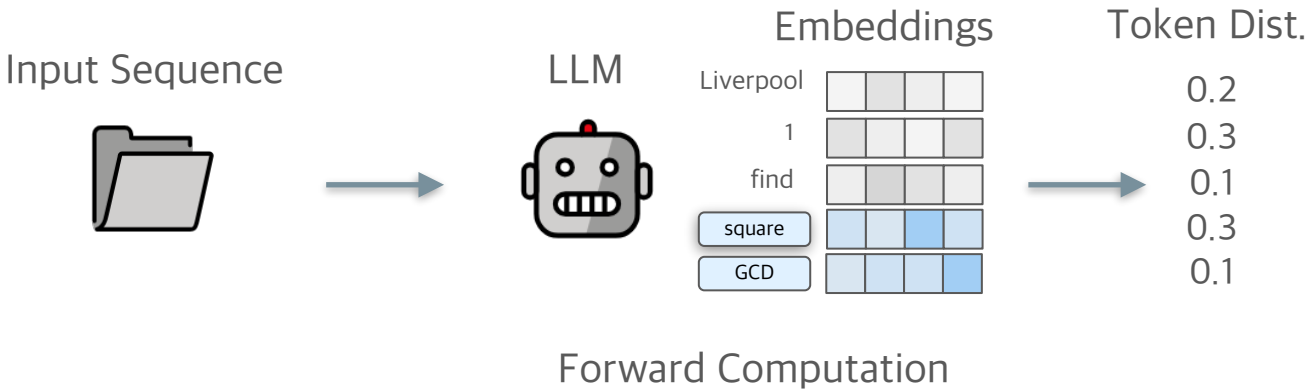


Training toolken embedding – Optimization

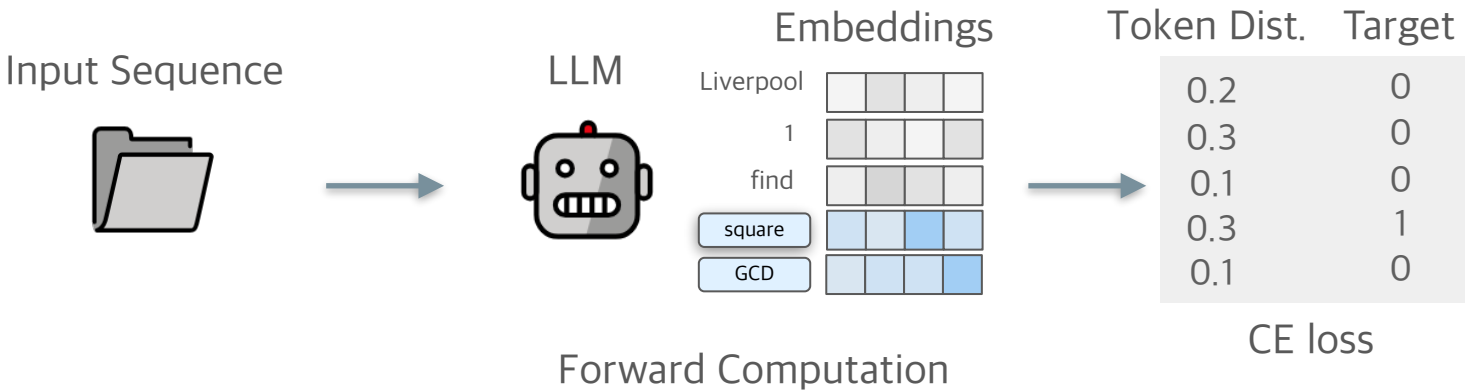


Initialize the toolken embeddings

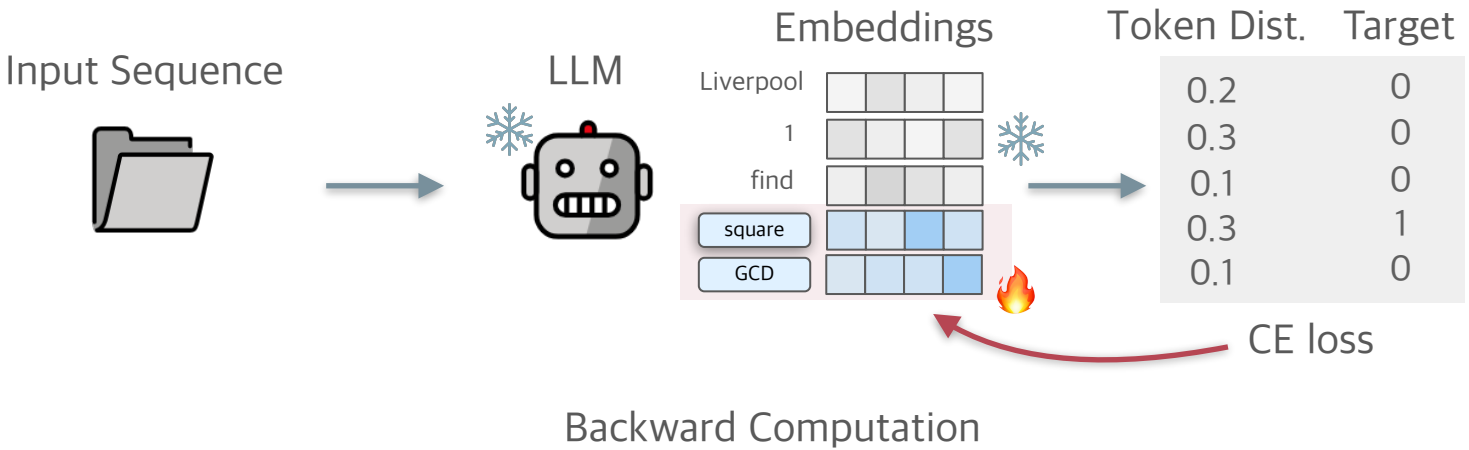
Training toolken embedding - Optimization



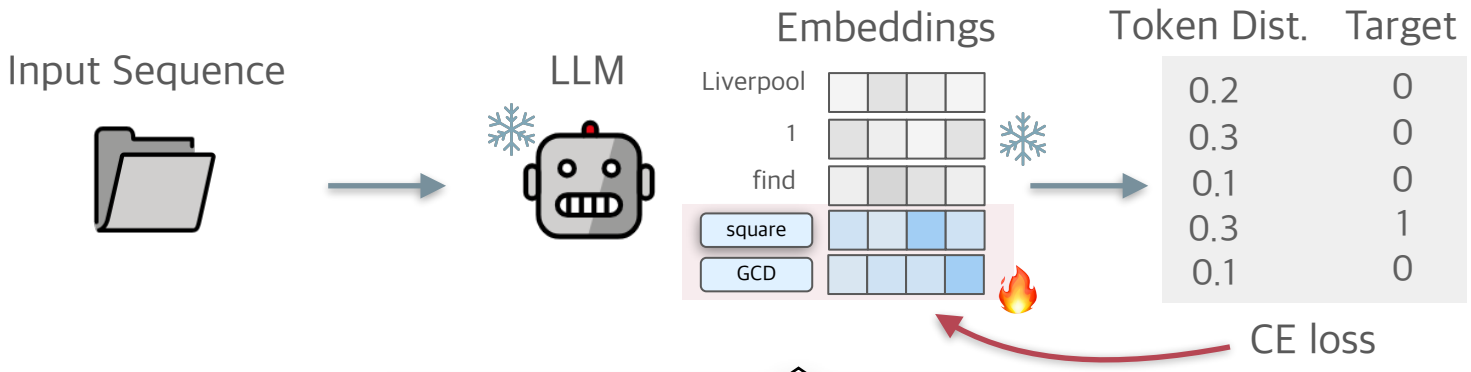
Training toolken embedding - Optimization



Training toolken embedding - Optimization



Training toolken embedding - Optimization

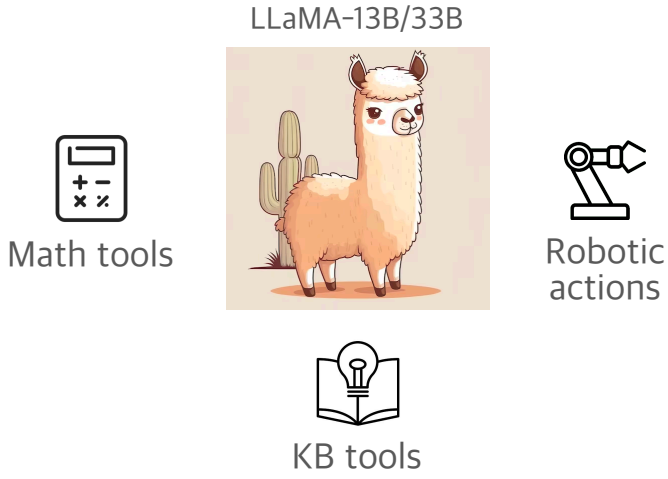


- Disentangled representations of tools
- No gradients flow through LLM

Plug-and-play of tools

Training speed & memory ≈ LLM inference

Experiments



Experiments – Math Reasoning

LLaMA-13B/33B

Math tools

Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer:

Experiments – Math Reasoning

LLaMA-13B/33B

Math tools

Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is

Experiments – Math Reasoning

LLaMA-13B/33B

Math tools

Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is GCD(64, 48)

Experiments – Math Reasoning

LLaMA-13B/33B

Math tools

Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16

Experiments – Math Reasoning

LLaMA-13B/33B

Math tools

Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is

Experiments – Math Reasoning

LLaMA-13B/33B

Math tools

Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is square (16)

Experiments – Math Reasoning

LLaMA-13B/33B

Math tools

Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is 256

Experiments – Math Reasoning

LLaMA-13B/33B

Math tools

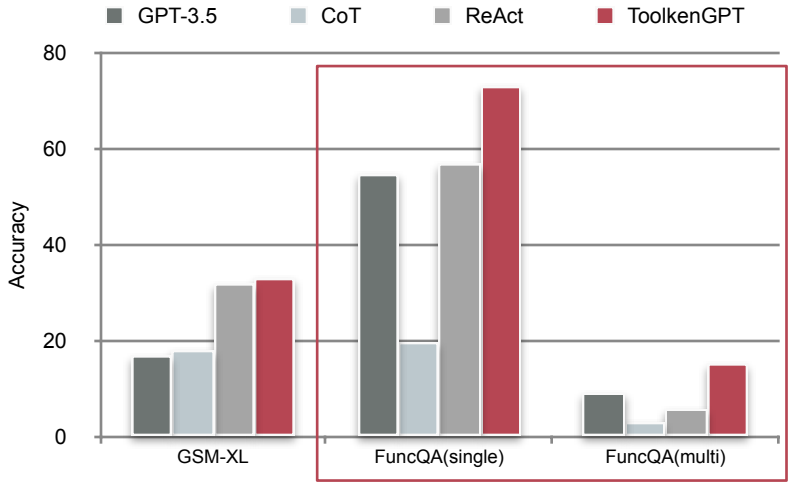
Question: John has a rectangular garden, of which the length is 64 meters and the width is 48 meters. He wants to divide the garden into identical square sections, each with the largest possible area. What's the area of each section?

Answer: The maximal side length of each section is 16 meters. Therefore, the area is 256 square meters.

Experiments – Math Reasoning

LLaMA-13B/33B

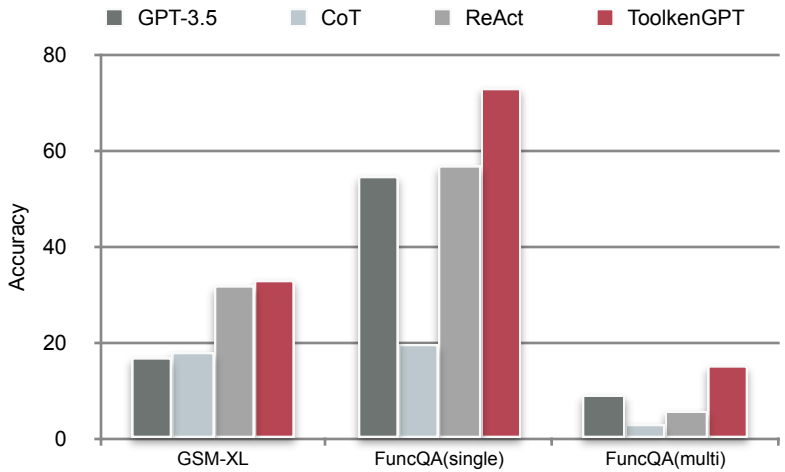
Math tools



Datasets that requires uncommon and complex tools

- Outperforms other tool learning baselines, especially better at **more complex math tools**.

Experiments – Math Reasoning

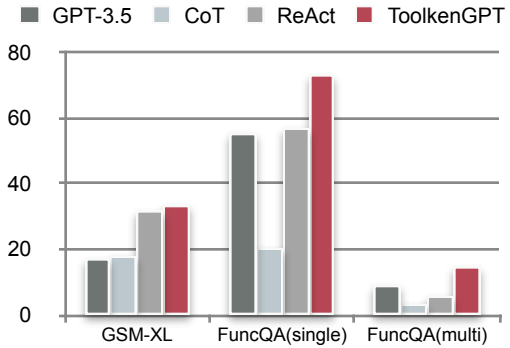


- Outperforms other tool learning baselines, especially better at **more complex math tools**.
- **Beats GPT-3.5** with LLaMA-33B

LLaMA-13B/33B

Math tools

Experiments



- Outperforms other tool learning baselines, especially better at **more complex math tools**.
- **Beats GPT-3.5** with LLaMA-33B

Math tools

LLaMA-13B/33B

KB tools

Robotic actions

Experiments – Knowledge-based QA

LLaMA-13B/33B

KB tools

Question: Which team is the winner of 2005-06 FA CUP?

Answer:

Experiments – Knowledge-based QA

LLaMA-13B/33B

KB tools

Question: Which team is the winner of 2005-06 FA CUP?

Answer: The winner is

Experiments – Knowledge-based QA

LLaMA-13B/33B

KB tools

Question: Which team is the winner of 2005-06 FA CUP?

Answer: The winner is winner_of (2005-06 FA CUP)

Experiments – Knowledge-based QA

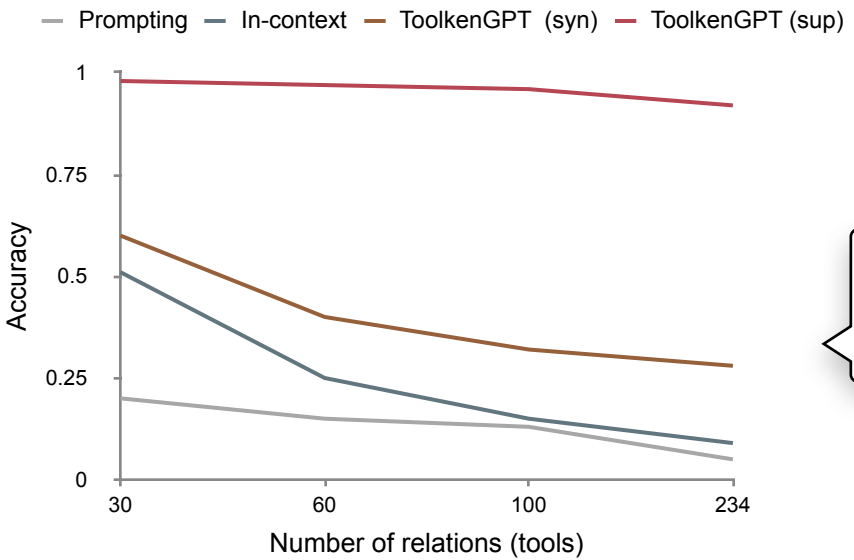
LLaMA-13B/33B

KB tools

Question: Which team is the winner of 2005-06 FA CUP?

Answer: The winner is Liverpool

Experiments - Knowledge-based QA



LLaMA-13B/33B

KB tools

- ToolkenGPT with only **synthetic data** beats all baselines
- Scales to **> 200 tools**

Experiments

- Outperforms other tool learning baselines, especially better at **more complex math tools**.
- **Beats GPT-3.5** with LLaMA-33B

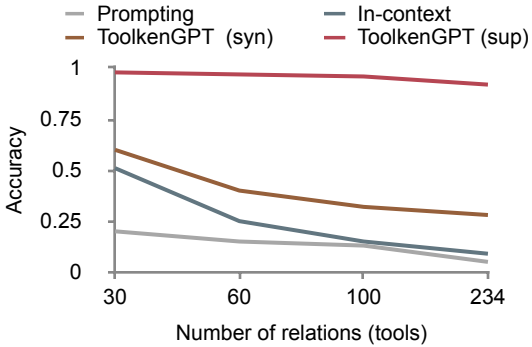
- ToolkenGPT with only **synthetic data** beats all baselines
- Scales to **> 200 tools**

Math tools

LLaMA-13B/33B

Robotic actions

KB tools



Experiments – Embodied Plan Generation

LLaMA-13B/33B

Robotic actions

Work: Go to office, sit at desk, turn on computer, enter password, open application and begin work

Experiments – Embodied Plan Generation

LLaMA-13B/33B

Robotic actions

Work: Go to office, sit at desk, turn on computer, enter password, open application and begin work

Plan:

[WALK] <office>

[WALK] <desk>

[FIND] <desk>

[SIT] <desk>

Generation with In-context Learning

Experiments – Embodied Plan Generation

LLaMA-13B/33B

Robotic actions

Work: Go to office, sit at desk, turn on computer, enter password, open application and begin work

Plan:

[WALK] <office>

[WALK] <desk>

[FIND] <desk>

[SIT] <desk>

Generation with In-context Learning

“Object not found” — Plan not grounded to the environment

Experiments – Embodied Plan Generation

LLaMA-13B/33B

Robotic actions

Work: Go to office, sit at desk, turn on computer, enter password, open application and begin work

Plan:

Generation with In-context Learning

[WALK] <office>

[WALK] <desk>

"Object not found" — Plan not grounded to the environment

[FIND] <desk>

[SIT] <desk>

"Desk not sittable": Doesn't understand the action "[SIT]"

Experiments – Embodied Plan Generation

LLaMA-13B/33B

Robotic actions

Work: Go to office, sit at desk, turn on computer, enter password, open application and begin work

Plan:

walk	home_office
walk	desk
find	home_office
find	chair

Generation with In-context Learning

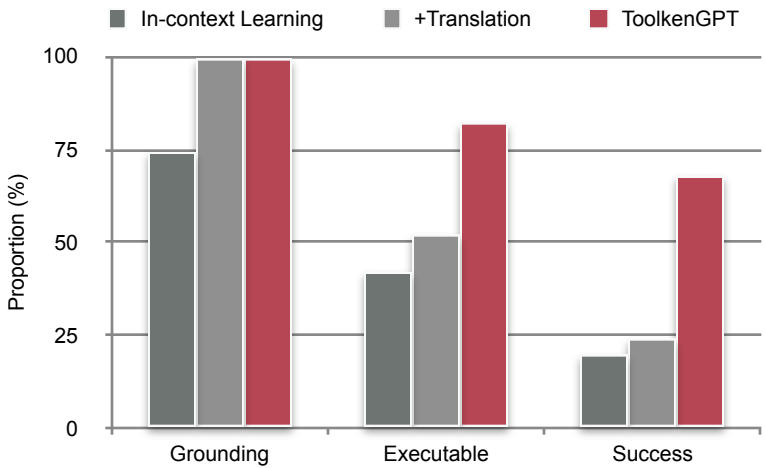
"Object not found" — Plan not grounded to the environment

✅ **Valid actions and objects = Toolken vocabulary**

"Desk not sittable" — Doesn't understand the action "[SIT]"

✅ **Learn from training data!**

Experiments – Embodied Plan Generation



- Naturally solved the **grounding** problem in embodied planning
- Higher success rate due to **deeper understandings** of tools

LLaMA-13B/33B

Robotic actions

Summary and Future Work

ToolkenGPT: Embedding the tools as tokens

- Frozen LLM / Massive tools / Plug & Play / Deeper understanding
- Superior performance in diverse domains

arXiv

GitHub

Summary and Future Work

ToolkenGPT: Embedding the tools as tokens

- Frozen LLM / Massive tools / Plug & Play / Deeper understanding
- Superior performance in diverse domains

Future work:

- **Planning for multi-step tool using** to solve more complex tasks

Reasoning with Language Model is Planning with World Model

Shibo Hao*♣ Yi Gu*♣ Haodi Ma♢ Joshua Jiahua Hong♣
Zhen Wang♣♠ Daisy Zhe Wang♢ Zhiting Hu♣

♣UC San Diego, ♢University of Florida
♠Mohamed bin Zayed University of Artificial Intelligence
{s5hao, yig025, jjhong, zhw085, zhh019}@ucsd.edu
{ma.haodi, daisyw}@ufl.edu

EMNLP 23'
GenPlan@NeurIPS 23'

Summary and Future Work

ToolkenGPT: Embedding the tools as tokens

- Frozen LLM / Massive tools / Plug & Play / Deeper understanding
- Superior performance in diverse domains

Future work:

- **Planning for multi-step tool using** to solve more complex tasks
- **Embedding stronger tools**?

arXiv

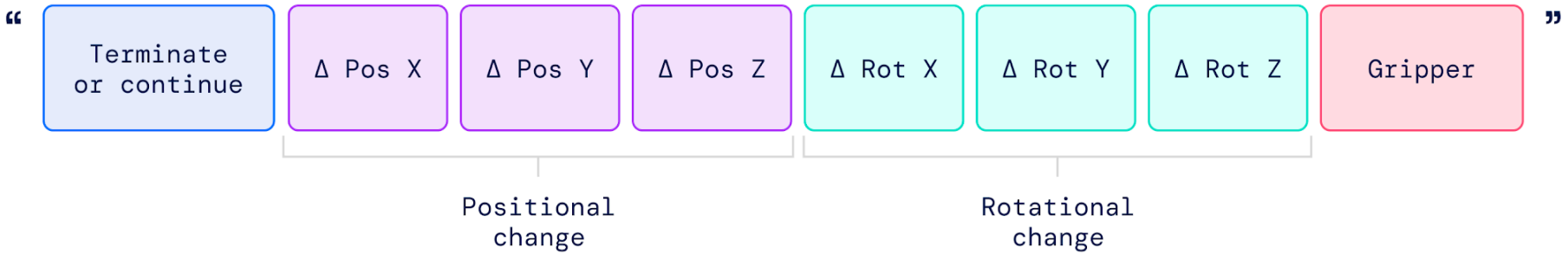
GitHub

Summary and Future Work

RT-2

By Google DeepMind

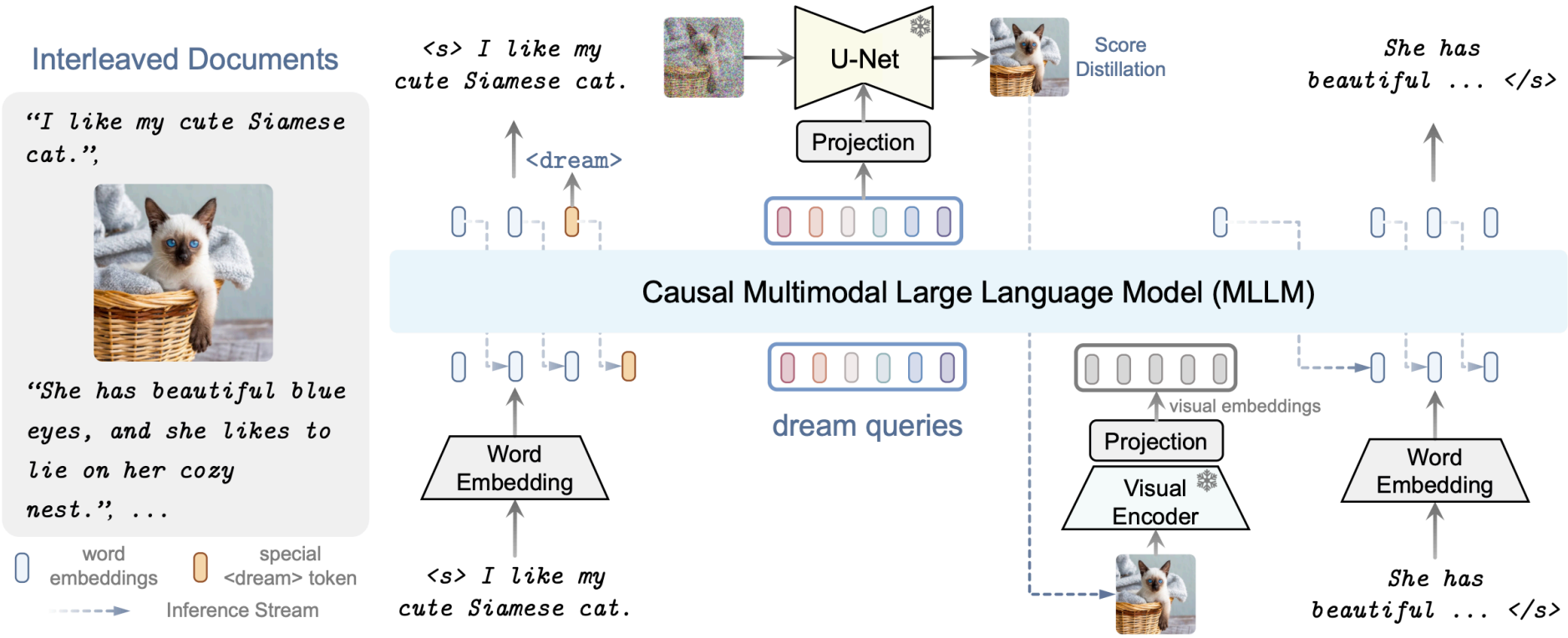
“To control a robot, it must be trained to output actions. We address this challenge by **representing actions as tokens in the model’s output - similar to language tokens** - and describe actions as strings that can be processed by standard natural language tokenizer”



Summary and Future Work

DreamLLM

[Dong et al., 2023]



Summary and Future Work

ToolkenGPT: Embedding the tools as tokens

- Frozen LLM / Massive tools / Plug & Play / Deeper understanding
- Superior performance in diverse domains

Future work:

- **Planning for multi-step tool using** to solve more complex tasks
- **Embedding stronger tools, ⋯ or even multiple LLM agents?**

arXiv

GitHub