salesforce

AI Agents: From Language to Multimodal Reasoning

ICCV 2025 MMRAgI Workshop

Juan Carlos Niebles, Research Director
@jcniebles | www.niebles.net

Conversational AI Agents

salesforce

Work Agent

Hello I'm your trusted Agentforce AI Agent

Can you help route my case for approval?

I have modified this case status to Escalated and created an escalation request with the subject "Modify deliver center to Portland, OR".

[Edit Case]

Personal Agent

Let's start booking my summer vacation travel. Find and book family friendly hotel for my upcoming stay in France..

I found this nice hotel near the venues you are planning to visit. I booked 2 rooms for the full duration of your stay.

[View Booking]

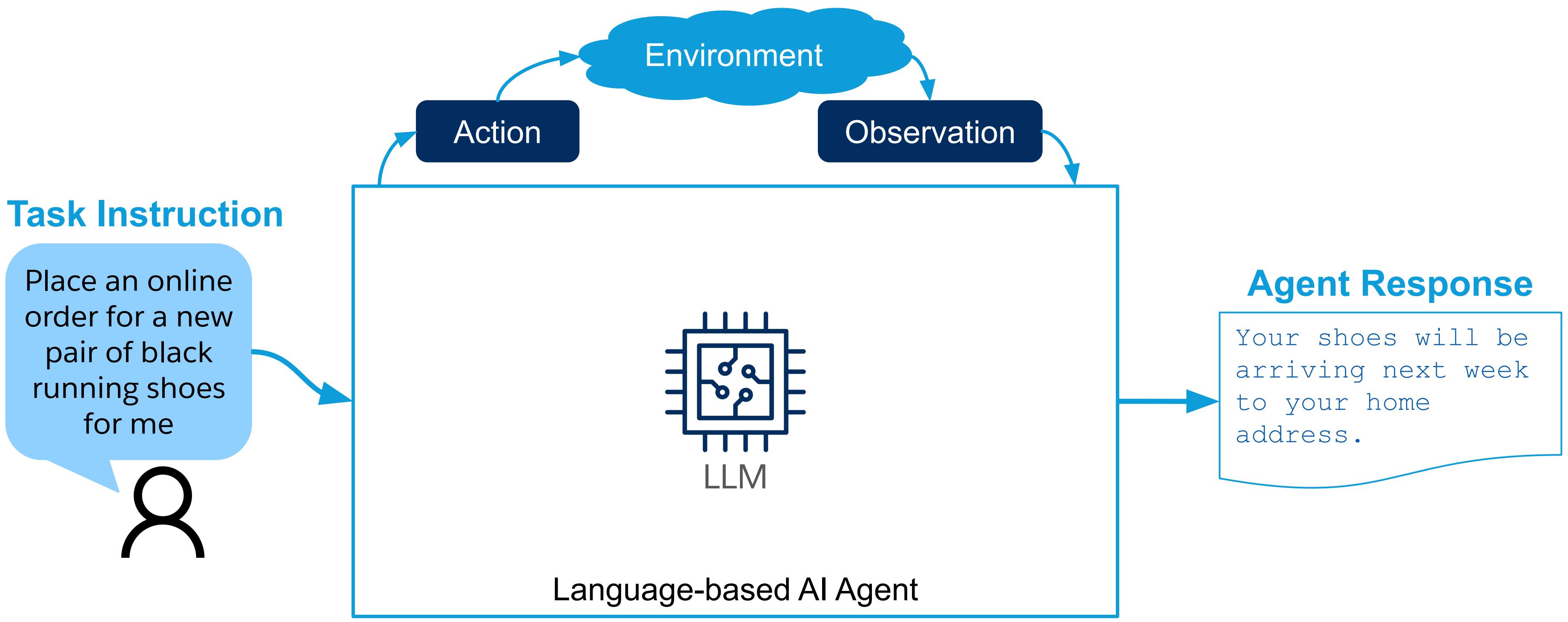
Creative Agent

I'm writing a fiction story about animals in the jungle. What are some adventures these animals might have? Give me some ideas and illustrate them with pictures.

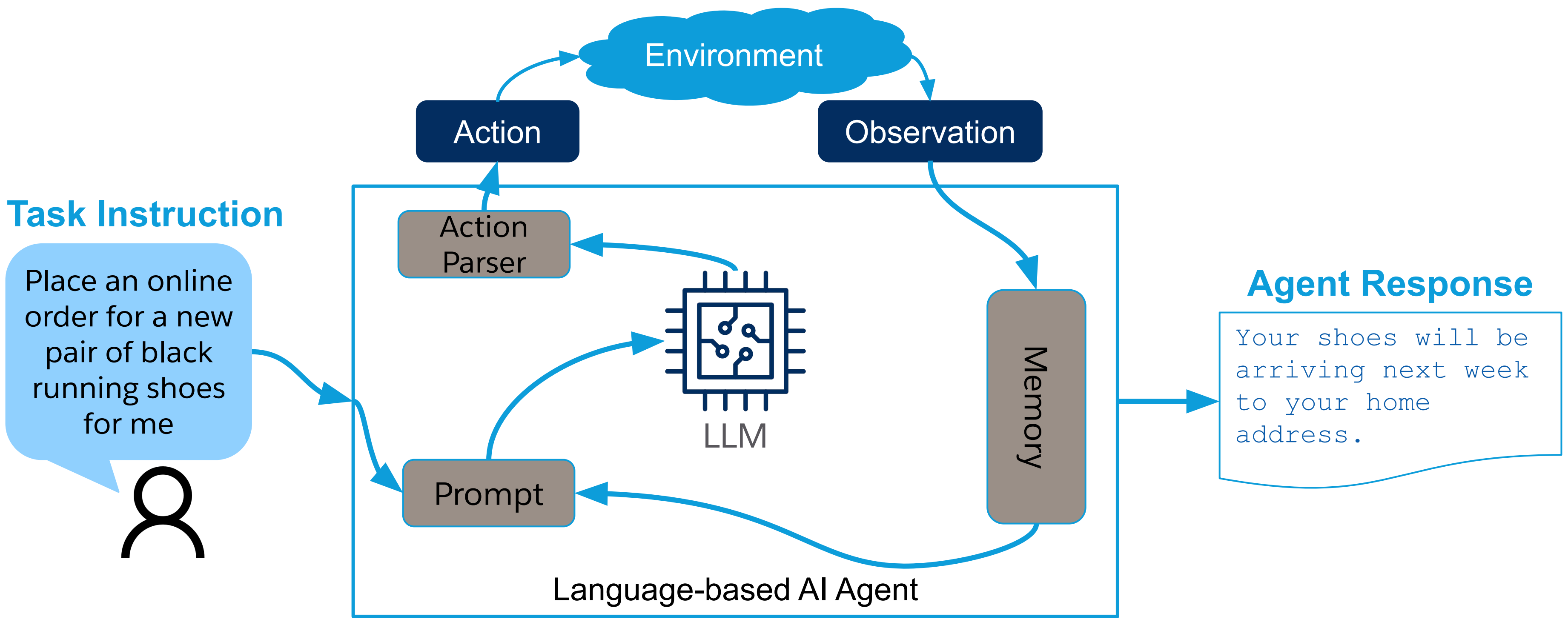
Sure, some ideas:
1. The animals might organizing a party.

2. The animals may need to solve a problem [...]

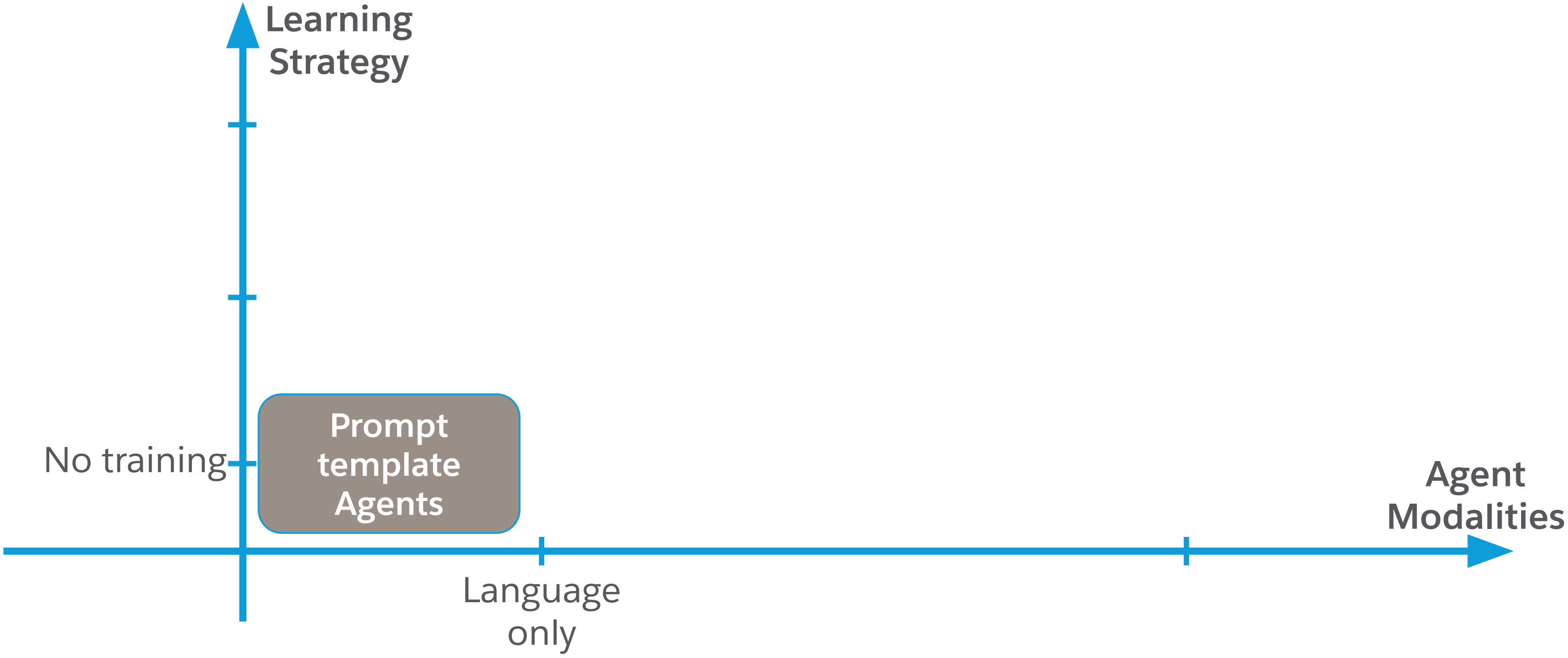
High-level Language-based Agent framework



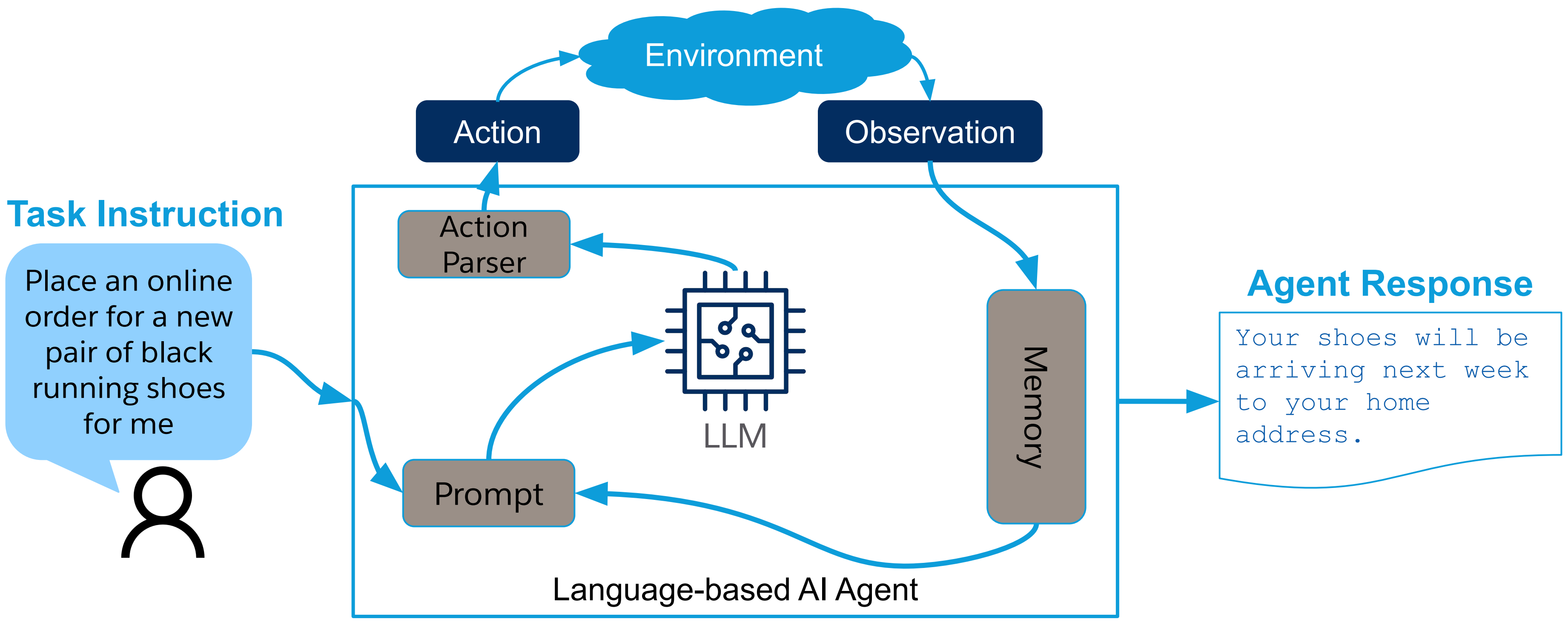
High-level Language-based Agent framework



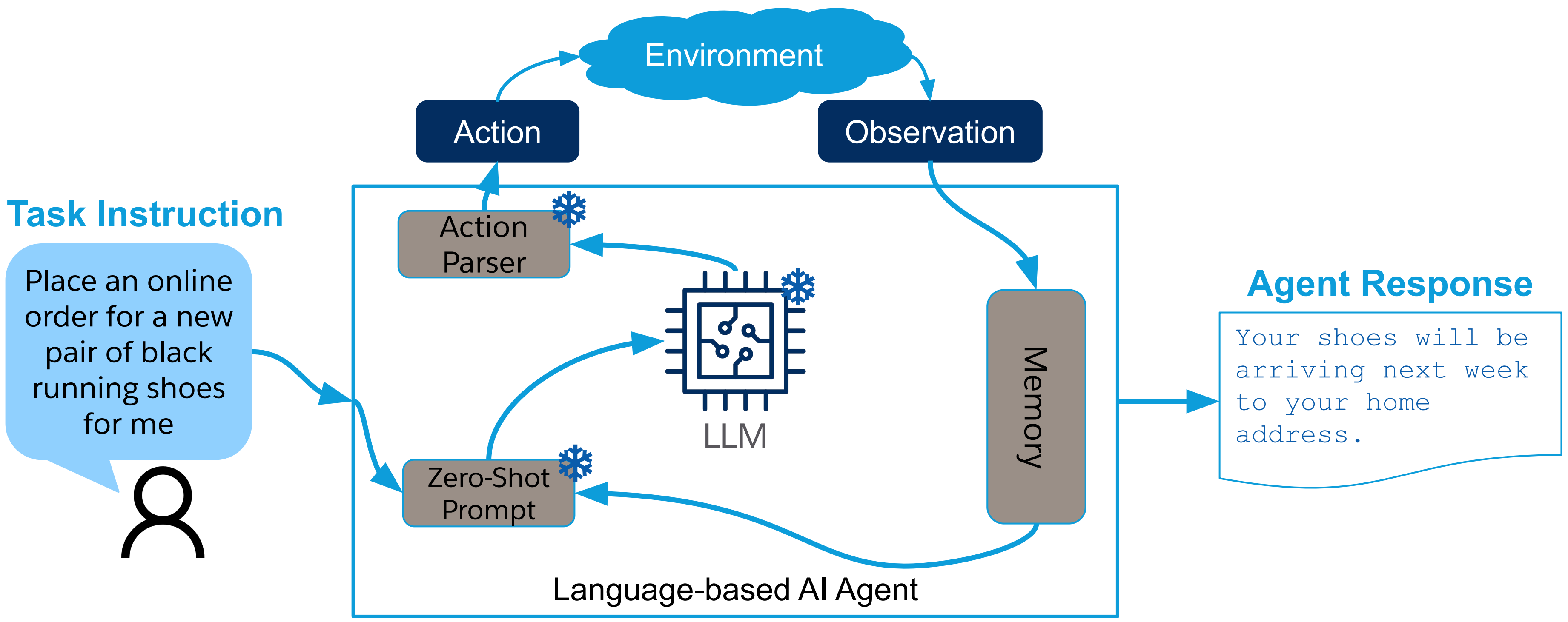
AI Agents: From Language to Multimodality



High-level Language-based Agent framework



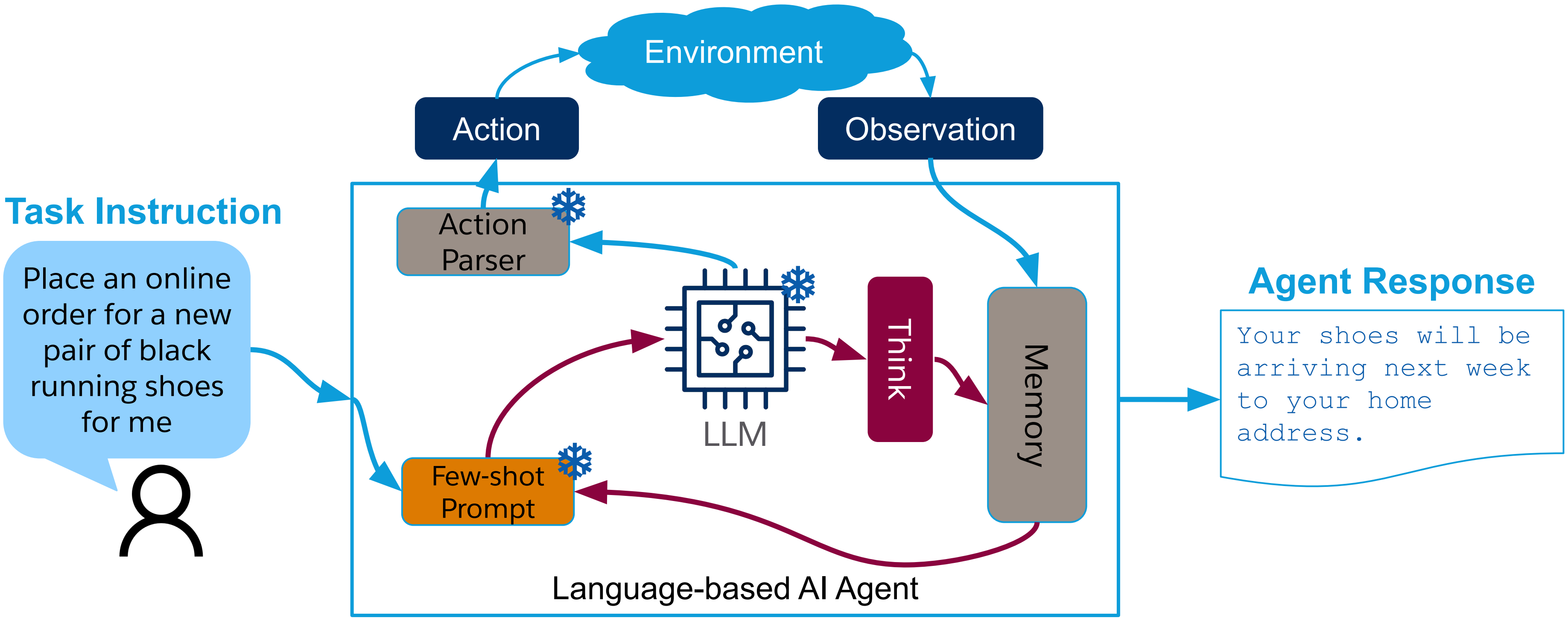
Zero-shot Language-based Agent



[Zhiwei Liu. “BOLAA: Benchmarking and orchestrating LLM-augmented autonomous agents”. ICLR 2024 Workshop on LLM Agents]

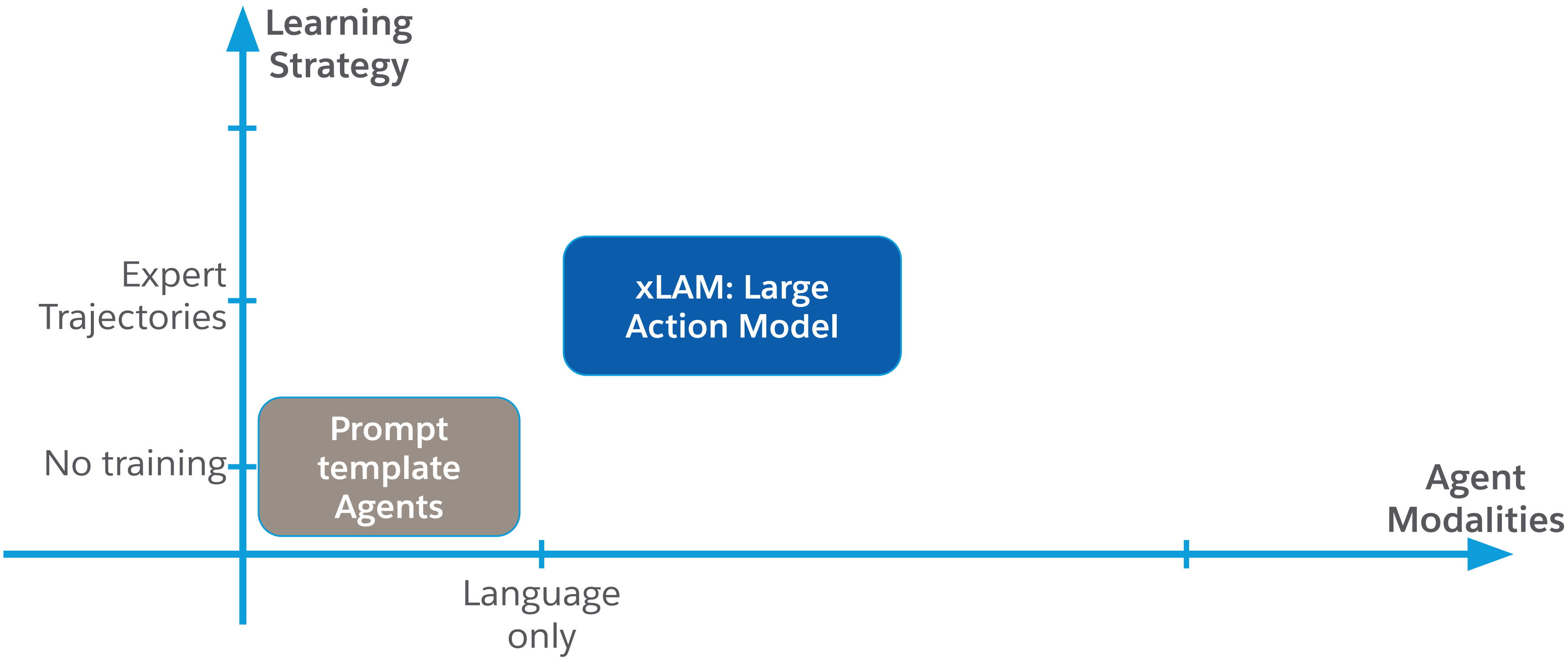
ReAct Agent

https://github.com/ysymyth/ReAct



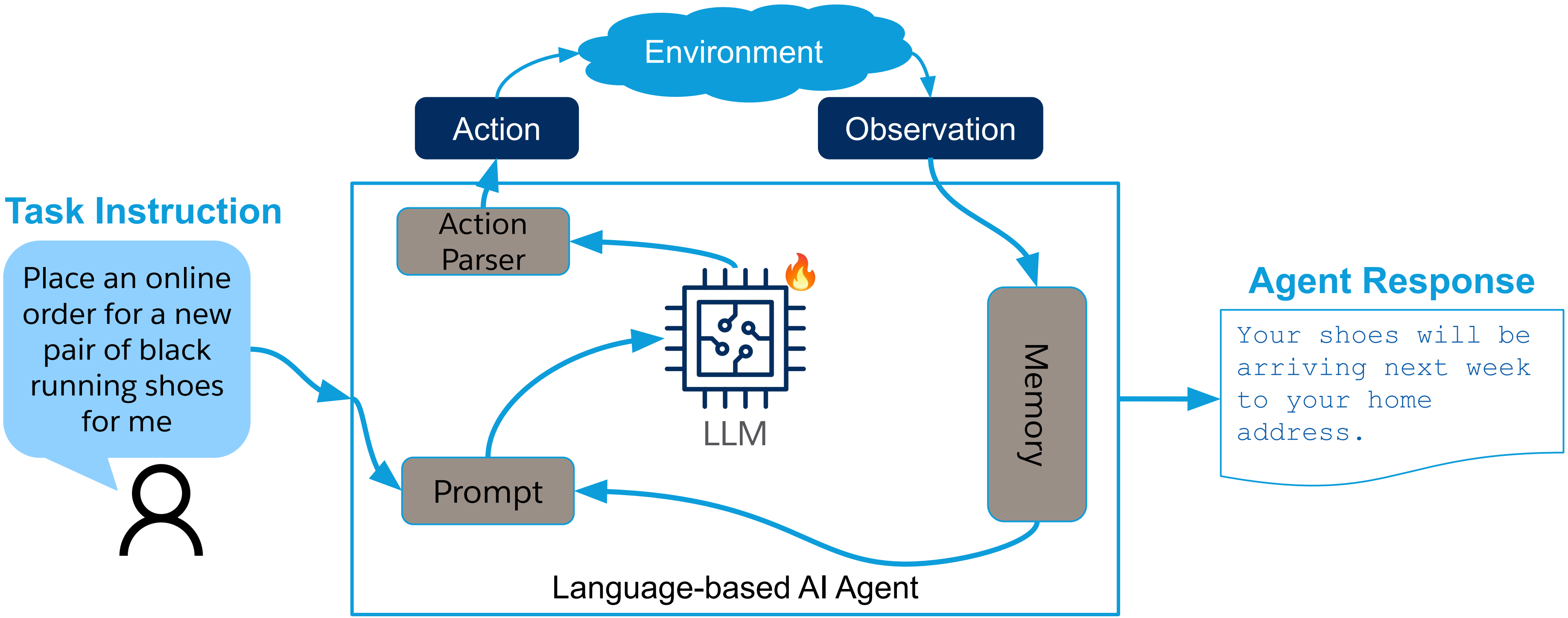
[Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023]

AI Agents: From Language to Multimodality



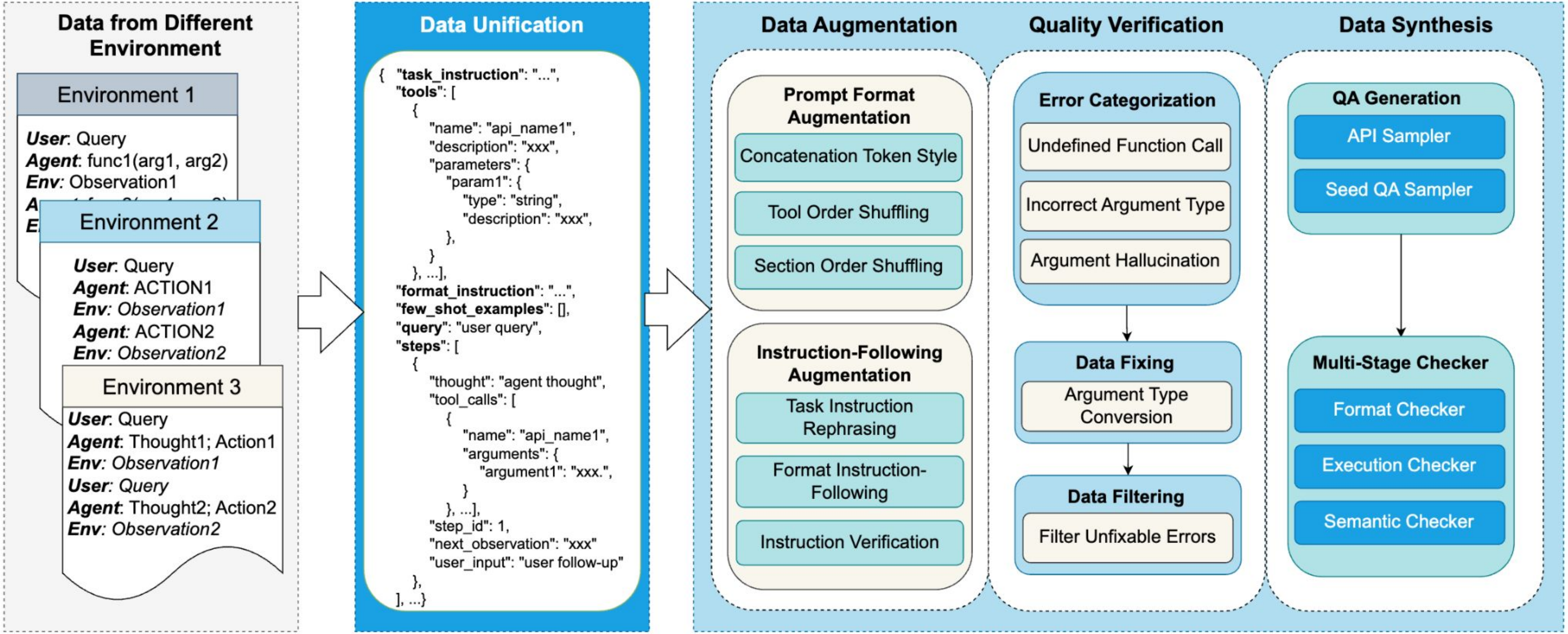
Large Action Models (LAMs)

https://github.com/SalesforceAIResearch/xLAM



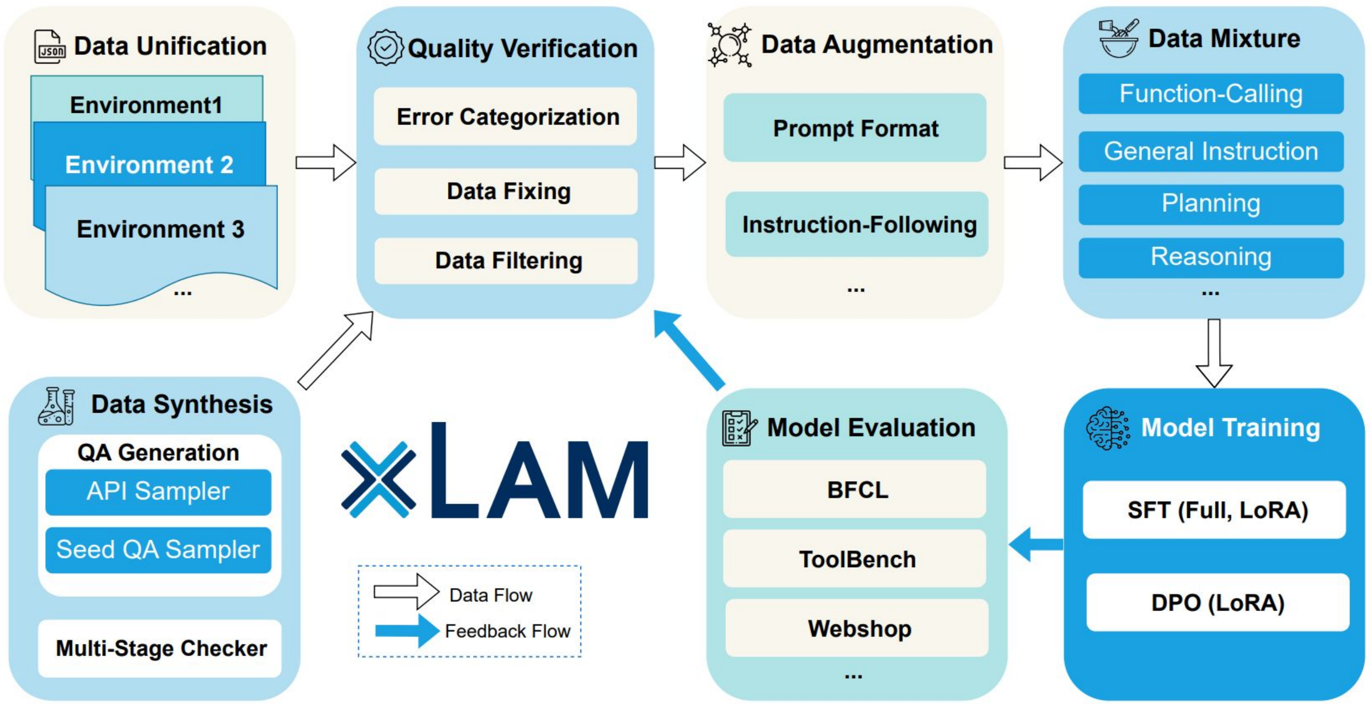
[Zhang et al. xLAM: A Family of Large Action Models to Empower AI Agent Systems.NAACL 2025]

xLAM Data Pipeline



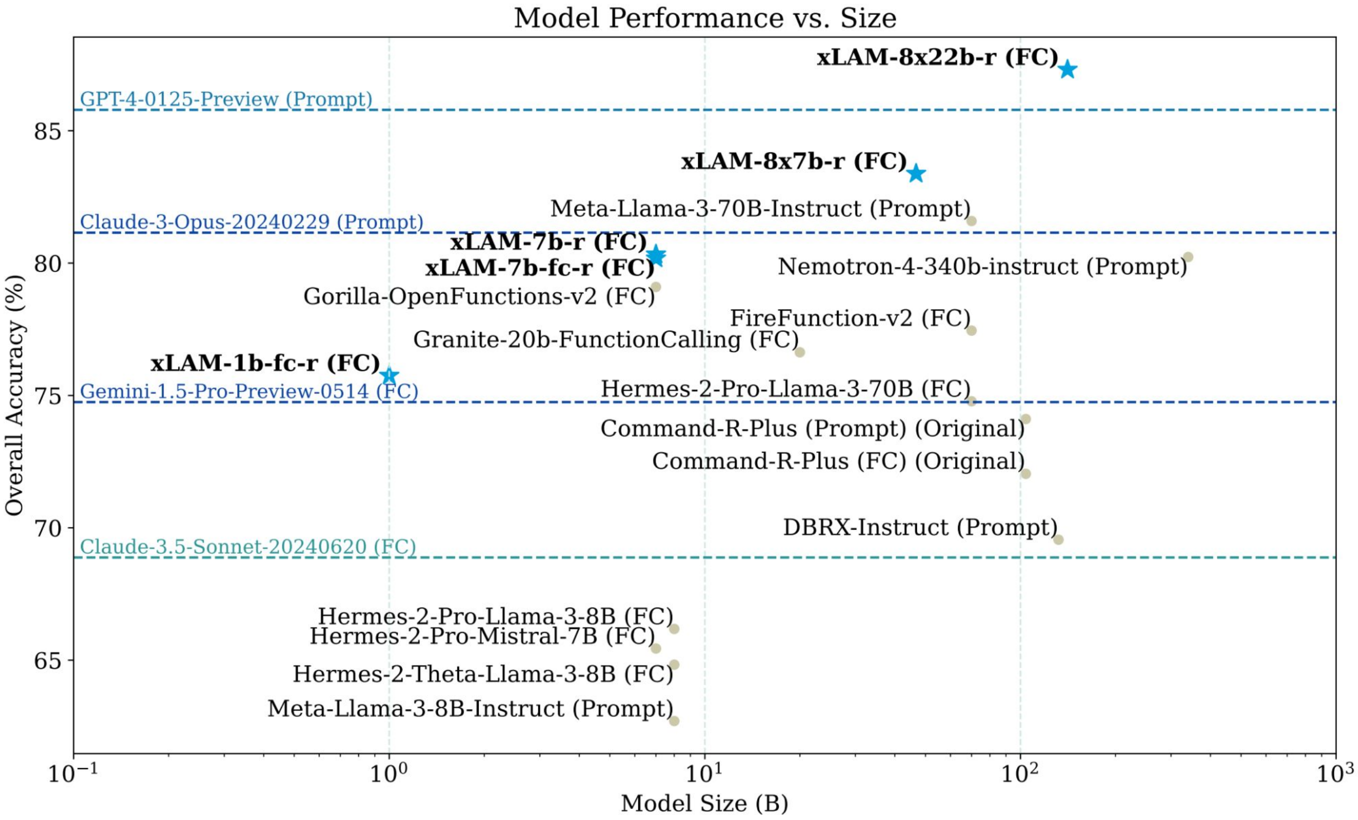
[Zhang et al. xLAM: A Family of Large Action Models to Empower AI Agent Systems.NAACL 2025]

xLAM Training Pipeline



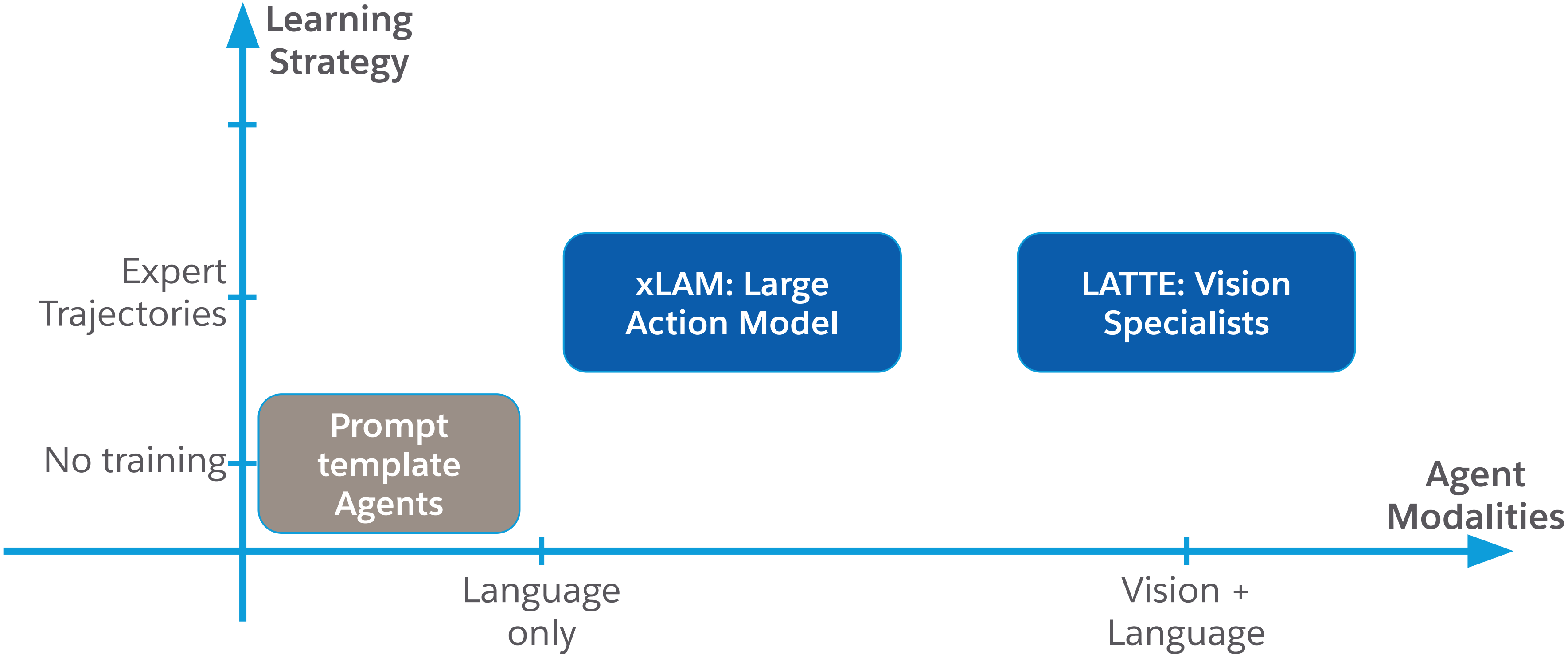
[Zhang et al. xLAM: A Family of Large Action Models to Empower AI Agent Systems.NAACL 2025]

xLAM Performance: Berkely Function Calling Leaderboard v2



[Zhang et al. xLAM: A Family of Large Action Models to Empower AI Agent Systems.NAACL 2025]

AI Agents: From Language to Multimodality



LATTE: A Multimodal AI Agent for Complex VQA

Q: What is the price for tomatoes?
A: 8.0

Mantis-LLaVA: 1.5
LLaVA-OV: 7.00

Finegrained OCR

Q: How many kids are in front of the yellow schoolbus?
A. 5; B. 4; C. 3; D. 6
A: B

Mantis-LLaVA: C
LLaVA-OV: A. There are 5 kids in front of the yellow schoolbus.

Visual grounding & counting

Q: How many gallons of supreme gasoline can I get with $50?
A: 13.7

Mantis-LLaVA: 3.6
LLaVA-OV: 5.2

Multi-step recog. & reasoning

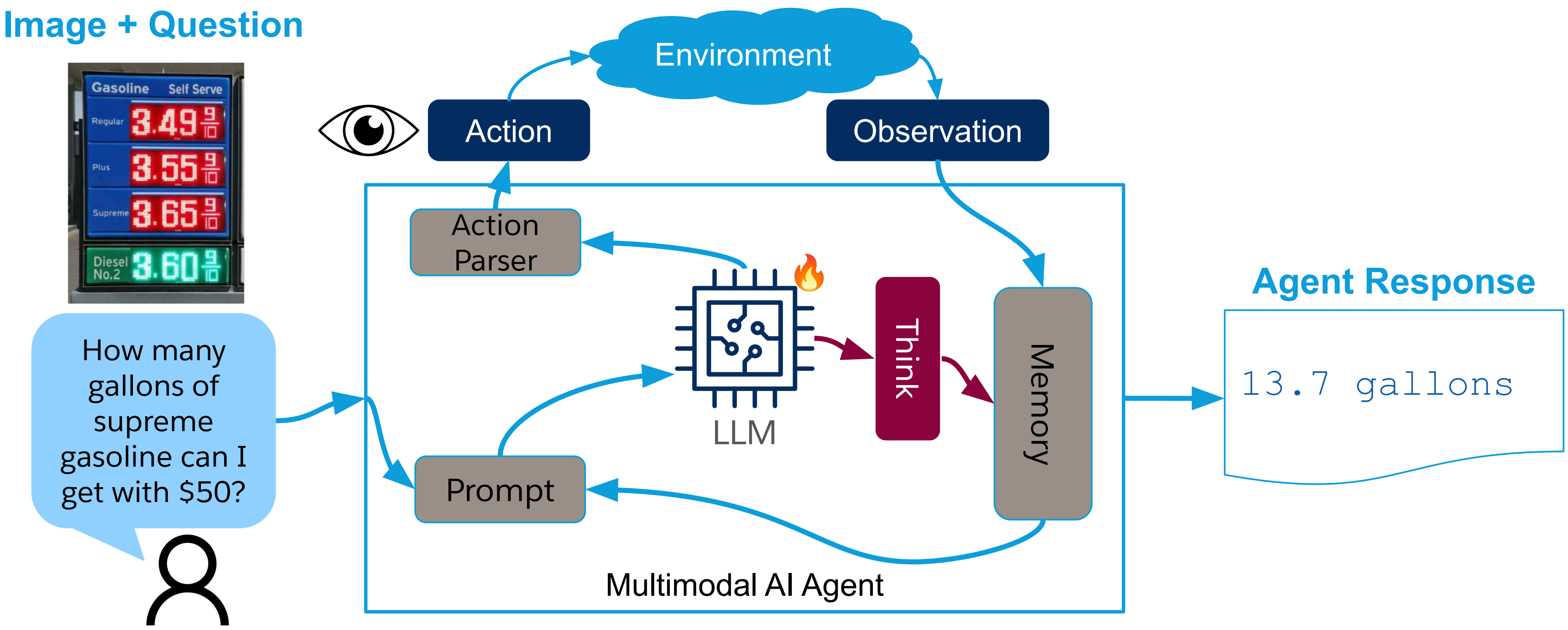
Q: Can you give a short introduction to this painting?
A: The Starry Night is an oil-on-canvas painting by Vincent van Gogh that depicts the view from the east-facing window of his asylum room at Saint-Rémy-de-Provence.

Mantis-LLaVA: The painting is a depiction of a starry night sky with a large starry sky in the background.
LLaVA-OV: Starry Night by Vincent van Gogh.

External knowledge

[Ma et al. LATTE: Learning to Think with Vision Specialists. EMNLP 2025]

LATTE: A Multimodal AI Agent for Complex VQA



LATTE vs MMLLMs

Q: How many gallons of supreme gasoline can I get with $50?
A: 13.7

Mantis-LLaVA-7B: 3.6
LLaVA-OV: 5.2

LATTE

t_0: I need to extract the price of supreme gasoline from the image.
a_0: **OCR**(image-0)
o_0: "Gasoline - Self Serve - Regular - 3.49 9/10 - Plus - 3.55 9/10 - Supreme - 3.65 9/10 - Diesel No. 2 - 3.60 9/10"

t_1: The price of supreme gasoline is $3.65 per gallon...I need to divide $50 by $3.65.
a_1: **Calculate**("50 / 3.65")
o_1: 13.6986

t_2: The calculation shows that you can get approximately 13.7 gallons of supreme gasoline with $50.
a_2: **Terminate**("13.7")

Q: Can the boy reach the highest book?
(A) Yes, (B) No
A: (B)

Mantis-LLaVA: (A)
LLaVA-OV: A. The boy can reach the book.

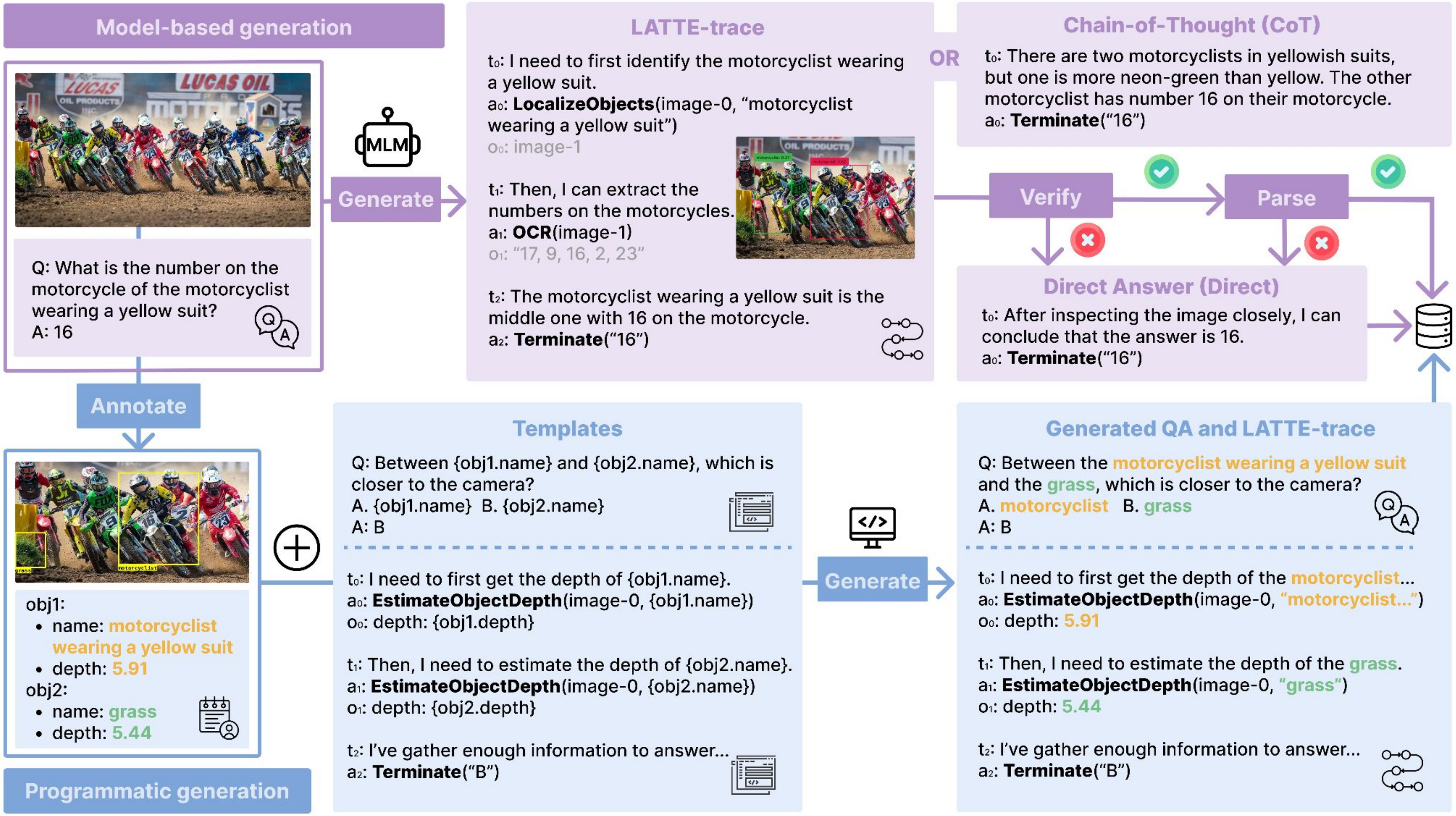
LATTE

t_0: To determine if the boy can reach the highest book, I need to analyze the image for the boy's height and the height of the bookshelf.
a_0: **LocalizeObjects**(image-0, ["boy", "bookshelf"])
o_0: [{'label': 'boy', 'bbox': [0.07, 0.48, 0.18, 0.95], 'score': 0.89},
{'label': 'bookshelf', 'bbox': [0.13, 0.01, 0.32, 0.72], 'score': 0.72}]

t_1: The boy's height is below the bookshelf, indicating he cannot reach the highest book.
a_1: **Terminate**("B")

[Ma et al. LATTE: Learning to Think with Vision Specialists. EMNLP 2025]

LATTE Trajectory Generation



LATTE Results

Substantial gains compared to vanilla instruction-tuning.

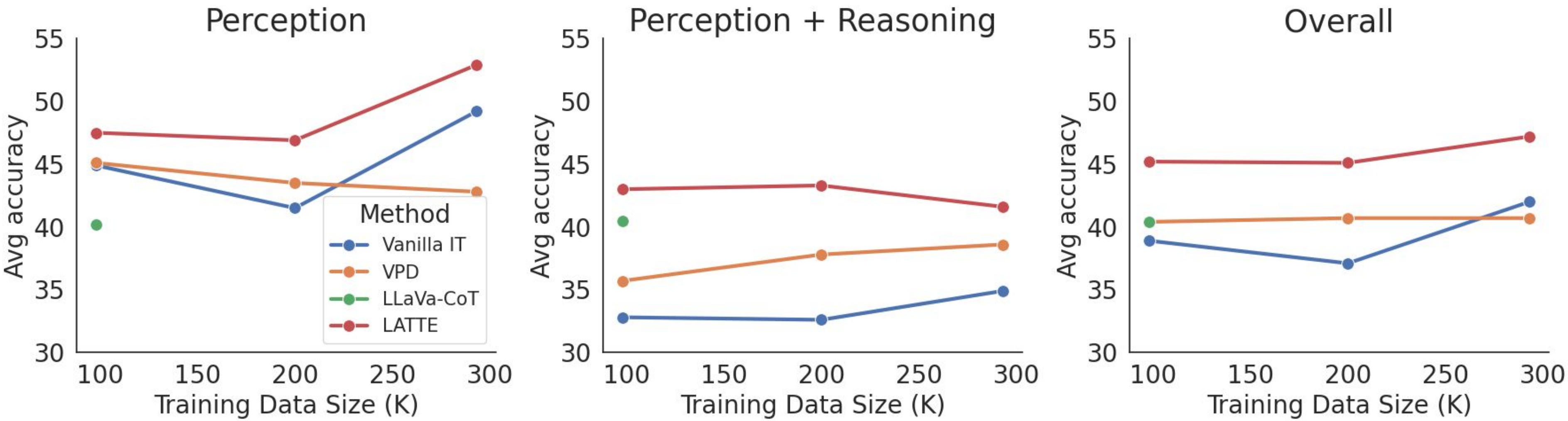
Other distillation baselines result in smaller gains or even degradation on some perception tasks.

Method	Perception				Perception + Reasoning				Overall
	BLINK	CV-Bench	RealWorldQA	Avg	MathVista	MMStar	MMVet	Avg	Avg
Vanilla IT	44.1	49.2	41.4	44.9	31.0	39.7	27.8	32.8	38.9
VPD	41.6	48.8	**44.8**	45.1 (+0.2)	33.0	41.1	32.8	35.7 (+2.8)	40.4 (+1.5)
LLaVa-CoT	42.2	40.4	38.0	40.2 (-4.7)	36.7	**44.6**	40.2	40.5 (+7.7)	40.4 (+1.5)
LATTE	**46.4**	**54.0**	42.0	**47.5** (+2.6)	**36.9**	44.2	**47.9**	**43.0** (+10.2)	**45.2** (+6.4)

[Ma et al. LATTE: Learning to Think with Vision Specialists. EMNLP 2025]

LATTE Results

Consistent gains over baselines across varying training data sizes



[Ma et al. LATTE: Learning to Think with Vision Specialists. EMNLP 2025]

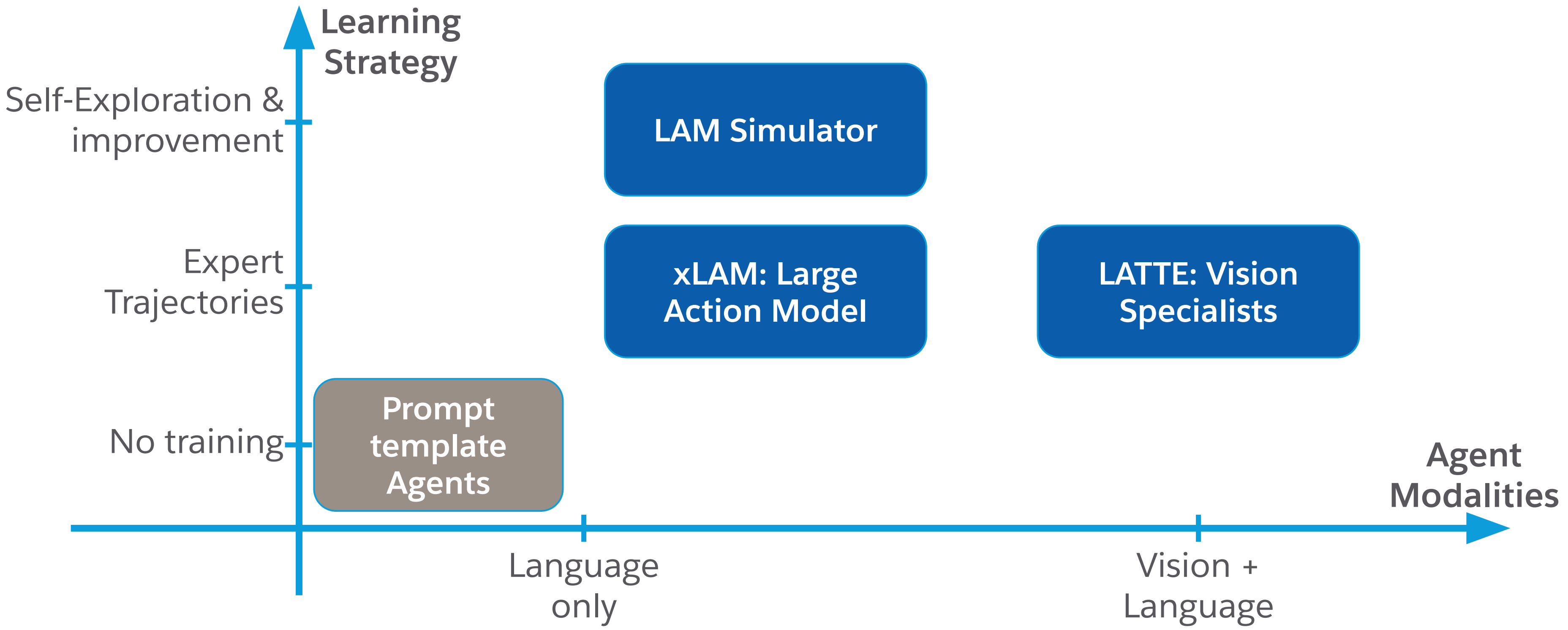
LATTE Results

Gains over baseline across all benchmarks regardless of the base model and checkpoint.

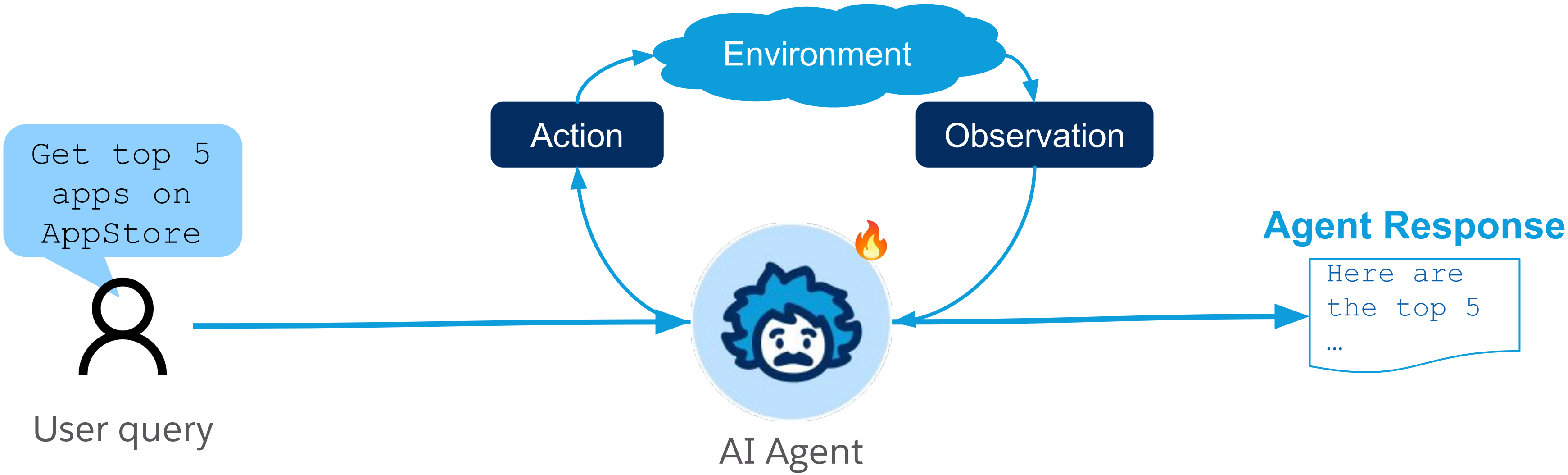
Language / Vision	Starting checkpoint	Method	Perception				Perception + Reasoning				Overall
			CV-Bench	BLINK	RealWorldQA	Avg	MathVista	MMStar	MMVet	Avg	Avg
LLaMA3-8B / CLIP	Mantis Pretrained	Vanilla IT	52.6	45.8	52.3	50.2	33.1	36.7	28.9	32.9	41.6
		LATTE	56.9	49.6	51.1	52.6	36.6	40.8	45.2	40.8	46.7 (+5.1)
LLaMA3-8B / SigLIP	Mantis Pretrained	Vanilla IT	52.3	43.7	51.8	49.3	31.1	40.5	33.0	34.9	42.1
		LATTE	<u>57.2</u>	47.8	53.7	52.9	34.9	44.6	45.2	41.6	47.2 (+5.1)
	Mantis Instruct-tuned	Vanilla IT	50.6	46.7	54.8	50.7	36.2	40.7	29.7	35.5	43.1
		LATTE	51.7	47.3	56.1	51.7	38.9	45.1	<u>50.0</u>	<u>44.7</u>	48.2 (+5.1)
Qwen2-7B / SigLIP	LLaVa-OV Stage 1.5	Vanilla IT	56.8	<u>50.3</u>	<u>57.8</u>	<u>55.0</u>	<u>42.4</u>	<u>50.1</u>	39.3	43.9	<u>49.5</u>
		LATTE	**60.2**	**52.6**	**61.1**	**58.0**	**46.9**	**50.8**	**50.9**	**51.2**	**53.8** (+4.3)

[Ma et al. LATTE: Learning to Think with Vision Specialists. EMNLP 2025]

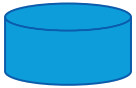
AI Agents: From Language to Multimodality



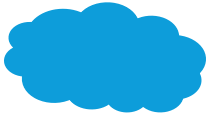
LAM Simulator: Overview



Keys to enable Exploration & Self-improvement:



Parameterized tasks instantiated into *Input user queries*

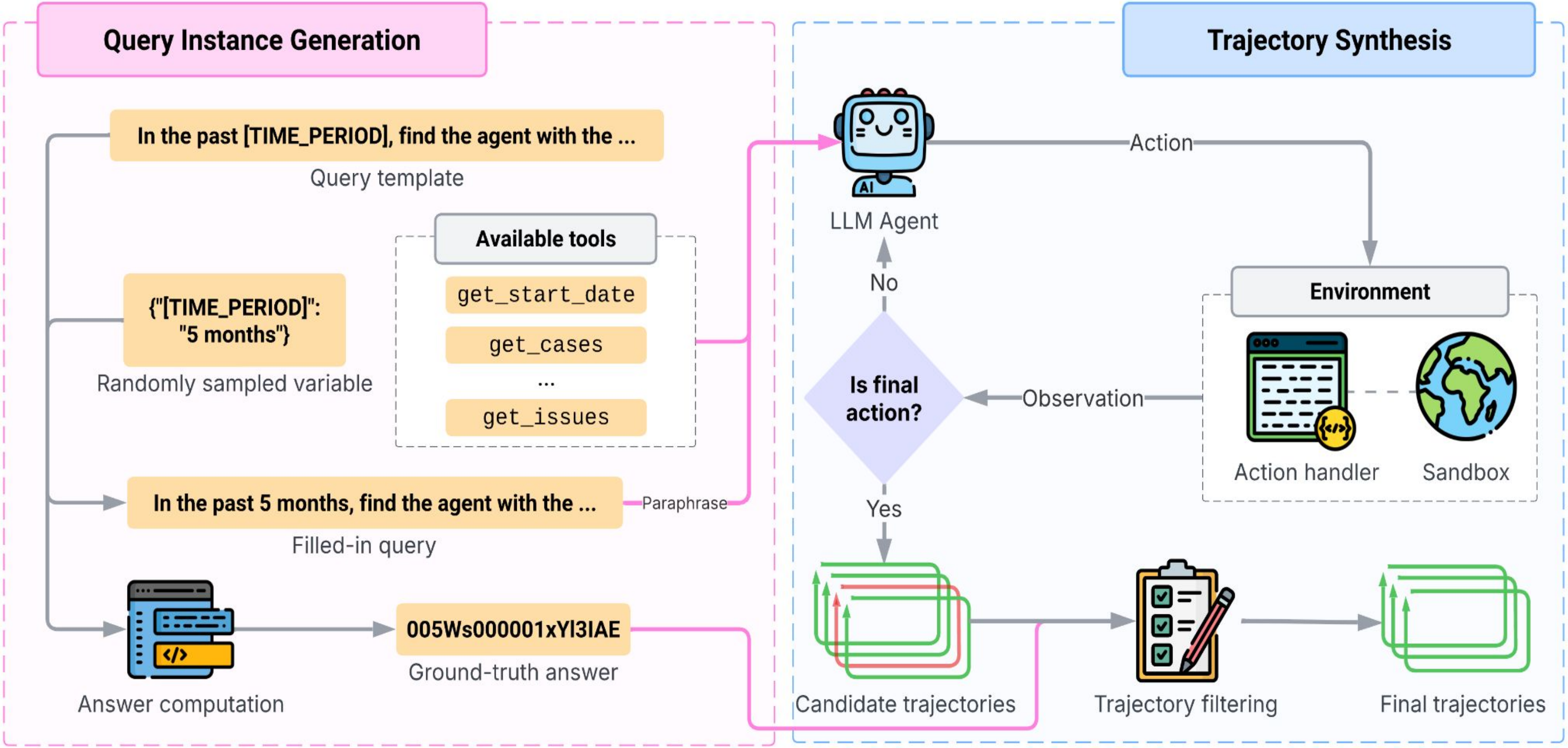


Simulation of Tool/API use

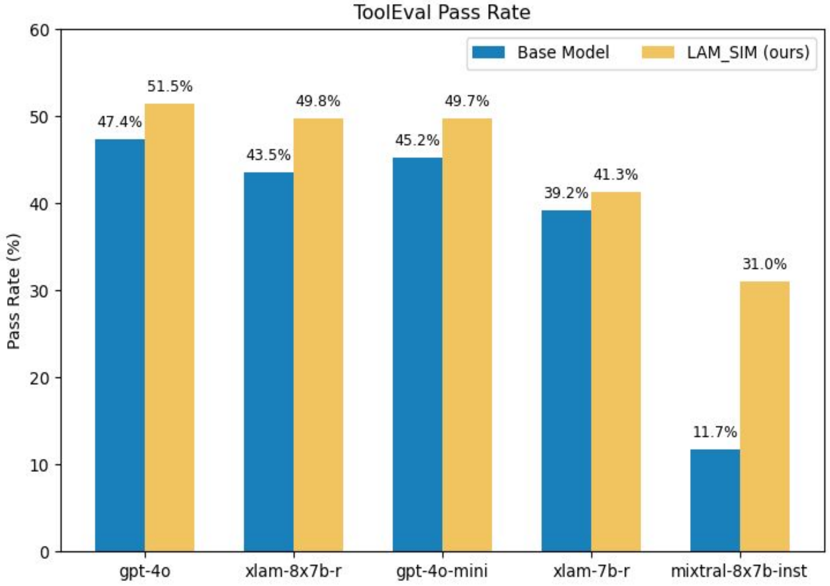
Learn & fine-tune from generated trajectories

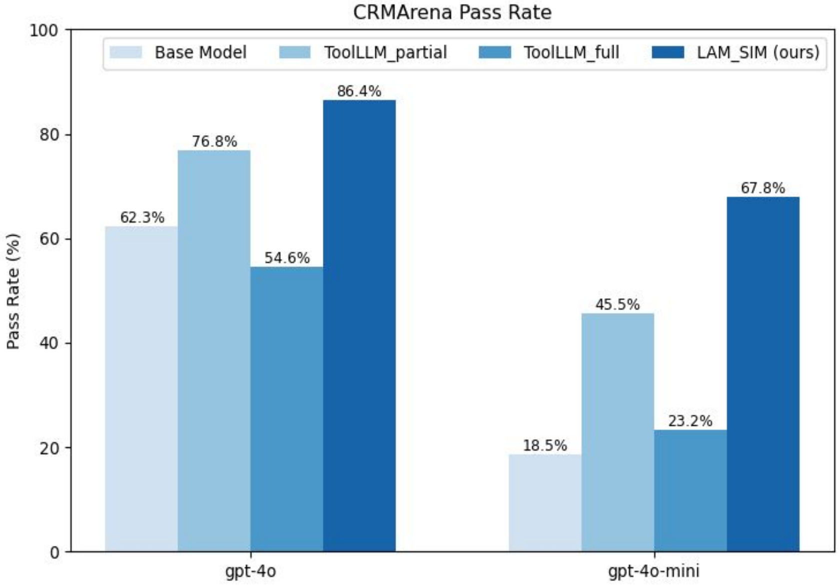
Automatic **Evaluation** of Task completion: **Rewards**

LAM Simulator: Framework design

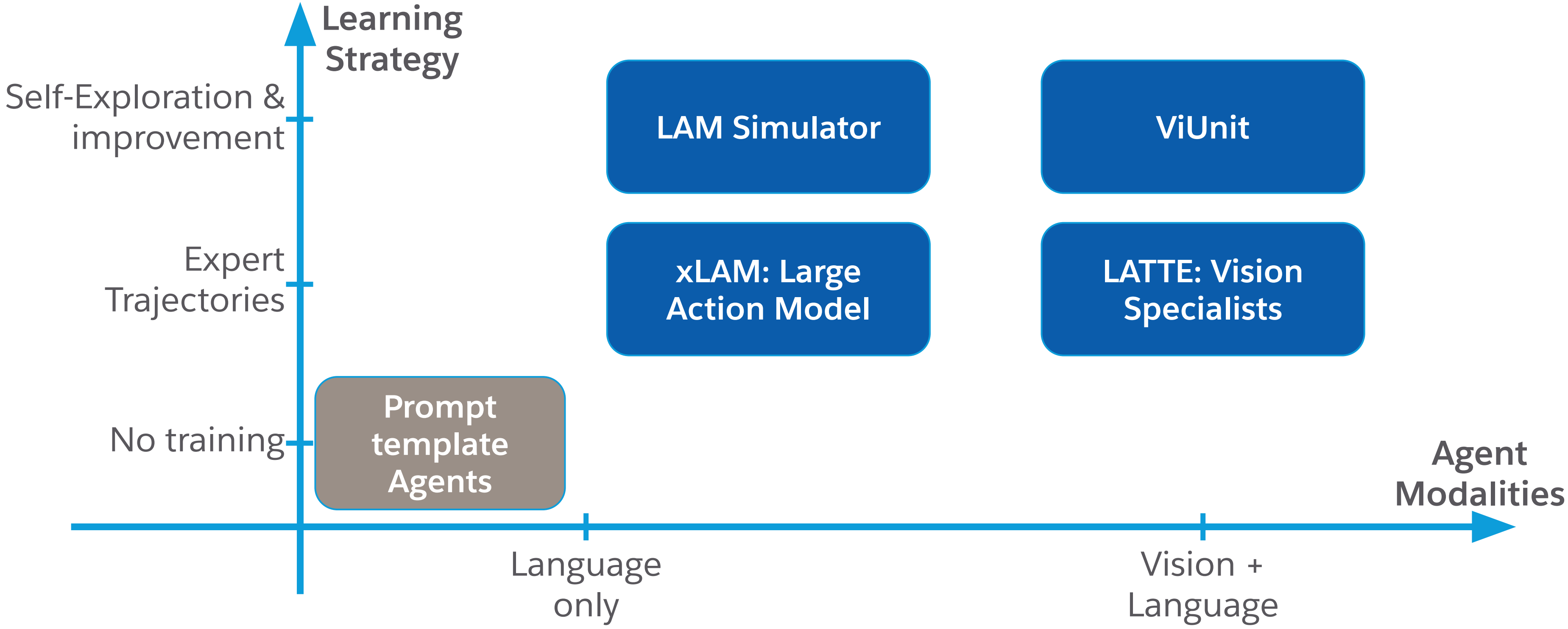


LAM Simulator: Key Results





AI Agents: From Language to Multimodality



Compositional Reasoning Remains a Challenge for VLMs

What color is the pillow on the pink bean bag?

The pillow on the pink beanbag is yellow.

GPT4o
5/31/25

The pillow on the pink beanbag is black.

InternVL3
5/31/25

There is no pillow on the pink beanbag.

Gemini 2.5 Pro
5/31/25

Image Credit: https://www.homestratosphere.com/multi-colored-living-room-ideas/

Visual Programming Emerged as a Promising Approach

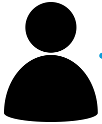
```
def execute_command(image) -> str:
    image_patch = ImagePatch(image)

    # Find the pink beanbag and crop to its region
    beanbag_patches = image_patch.find(
                                 "pink beanbag")
    if not beanbag_patches:
        return "unknown"
    beanbag_patch = beanbag_patches[0]

    # Look within the pink beanbag for a pillow
    pillow_in_beanbag_patch = beanbag_patch.find(
                                         "pillow")
    if not pillow_in_beanbag_patch:
        return "unknown"
    pillow_patch = pillow_in_beanbag_patch[0]

    return pillow_patch.simple_query(
                       "What color is the pillow?")
```

GPT4o
5/31/25



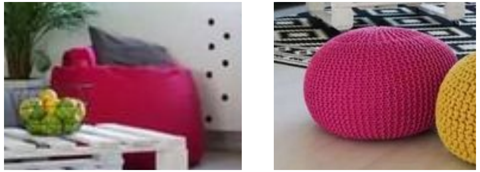
What color is the pillow on the pink bean bag?

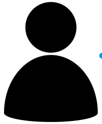
Image Credit: https://www.homestratosphere.com/multi-colored-living-room-ideas/

Visual Programming Emerged as a Promising Approach

```
# Find the pink beanbag and crop to its region
beanbag_patches = image_patch.find(
                              "pink beanbag")
```

```
beanbag_patches =
```





What color is the pillow on the pink bean bag?

Image Credit: https://www.homestratosphere.com/multi-colored-living-room-ideas/

Visual Programming Emerged as a Promising Approach

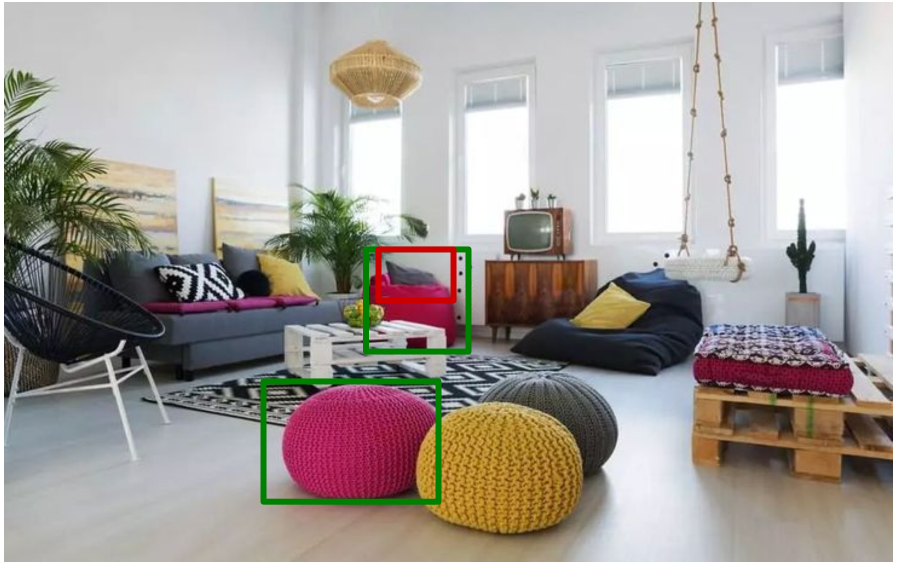
```
beanbag_patch = beanbag_patches[0]
```

```
beanbag_patch =
```

What color is the pillow on the pink bean bag?

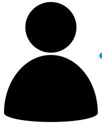
Image Credit: https://www.homestratosphere.com/multi-colored-living-room-ideas/

Visual Programming Emerged as a Promising Approach



```
pillow_in_beanbag_patch = beanbag_patch.find(
                                          "pillow")
```

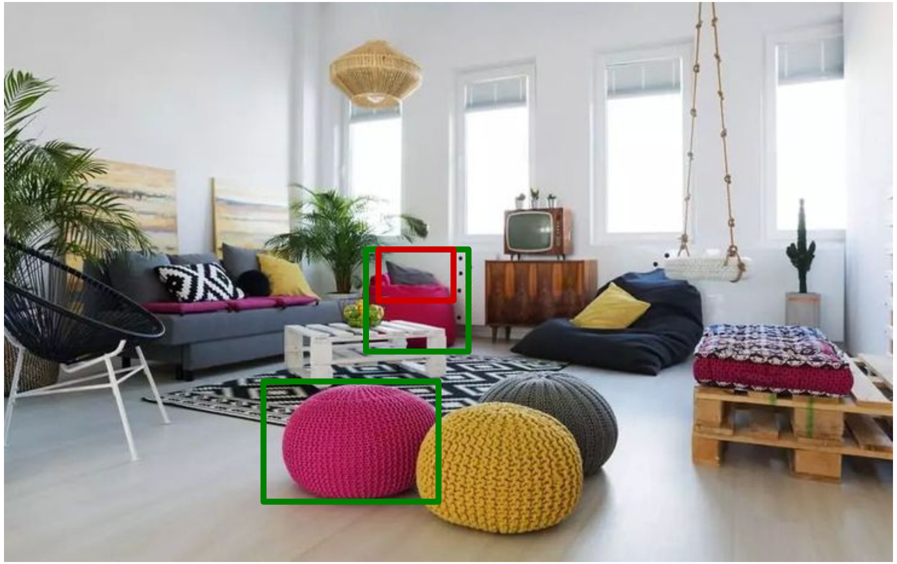
`pillow_in_beanbag_patch =` 



What color is the pillow on the pink bean bag?

Image Credit: https://www.homestratosphere.com/multi-colored-living-room-ideas/

Visual Programming Emerged as a Promising Approach



```
return pillow_patch.simple query(
                "What color is the pillow?")
```

```
.simple query(
"What color is the pillow?")
```

```
gray
```

What color is the pillow on the pink bean bag?

Image Credit: https://www.homestratosphere.com/multi-colored-living-room-ideas/

Visual Programming Emerged as a Promising Approach

```
def execute_command(image) -> str:
    image_patch = ImagePatch(image)

    # Find the pink beanbag and crop to its region
    beanbag_patches = image_patch.find(
                                "pink beanbag")
    if not beanbag_patches:
        return "unknown"
    beanbag_patch = beanbag_patches[0]

    # Look within the pink beanbag for a pillow
    pillow_in_beanbag_patch = beanbag_patch.find(
                                          "pillow")
    if not pillow_in_beanbag_patch:
        return "unknown"
    pillow_patch = pillow_in_beanbag_patch[0]

    return pillow_patch.simple_query(
                    "What color is the pillow?")
```

GPT4o

What color is the pillow on the pink bean bag?

Correct for the wrong reasons!

→ **not all pink beanbags are considered.**

Image Credit: https://www.homestratosphere.com/multi-colored-living-room-ideas/

Programs Are Often Correct for the Wrong Reasons

- 40% of correct programs are correct for the wrong reasons!
  - Manual inspection of 100 programs generated by CodeLLaMA-7B.

- Reduces **interpretability** and **generalization.**

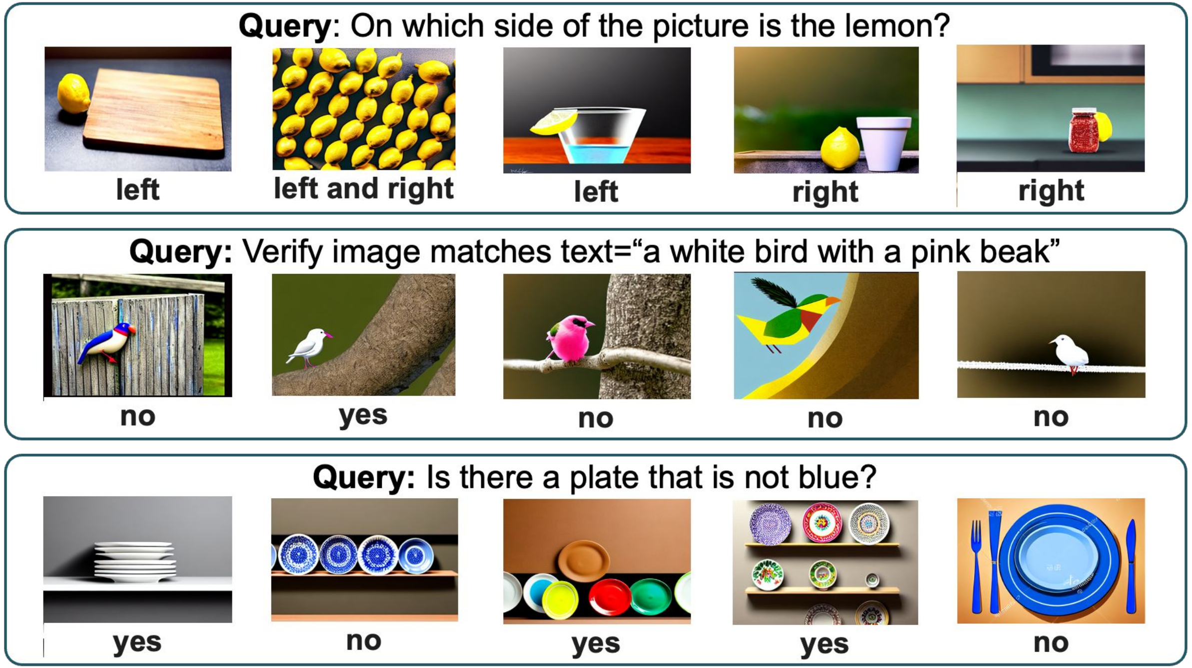
- **How can we calculate a reward signal to enable model improvement via self-exploration & RL?**

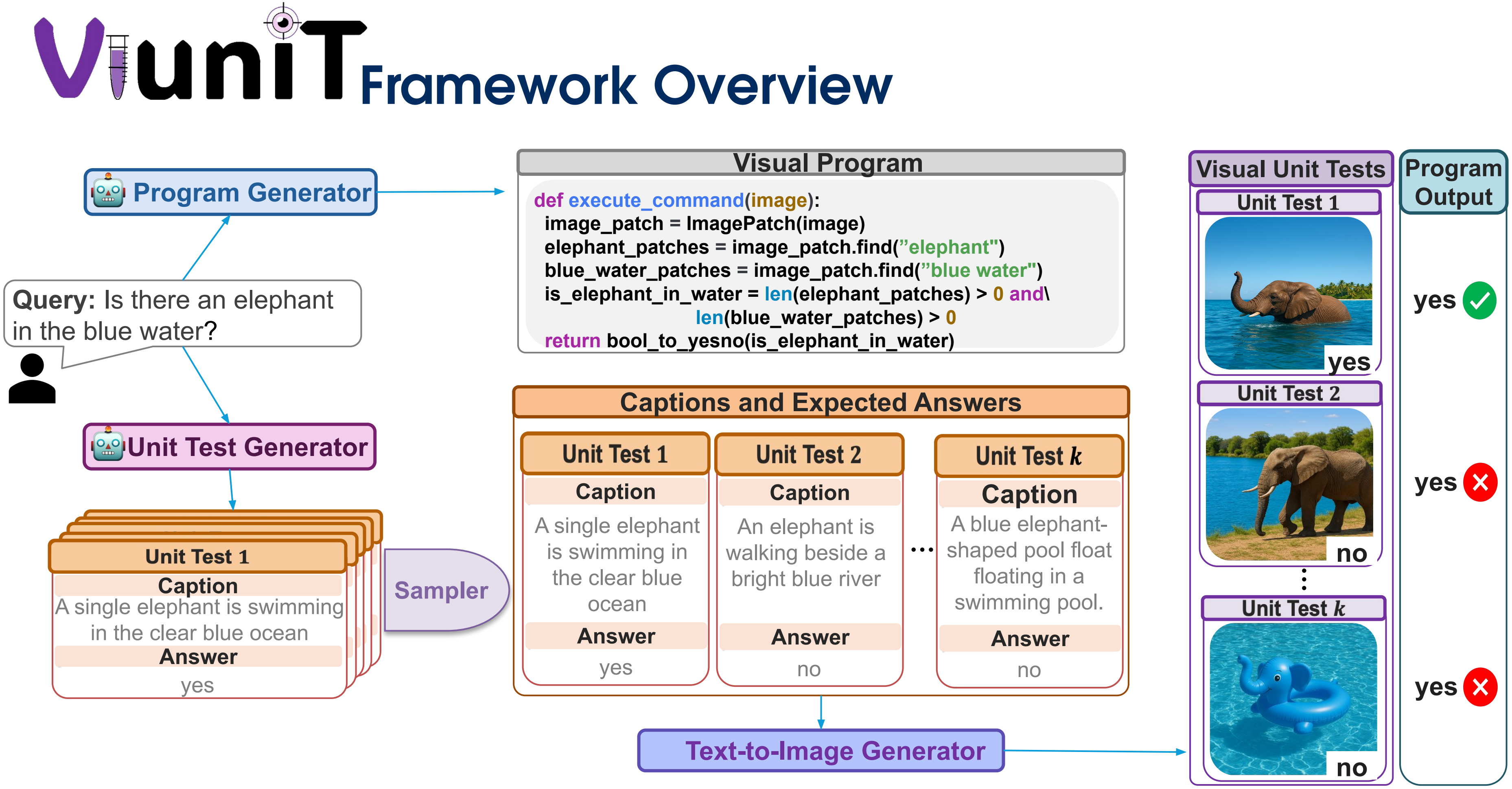
- **What if we could unit test visual programs, like human developers do?**

We introduce ViUniT a framework to improve visual programs by automatically generating unit tests.

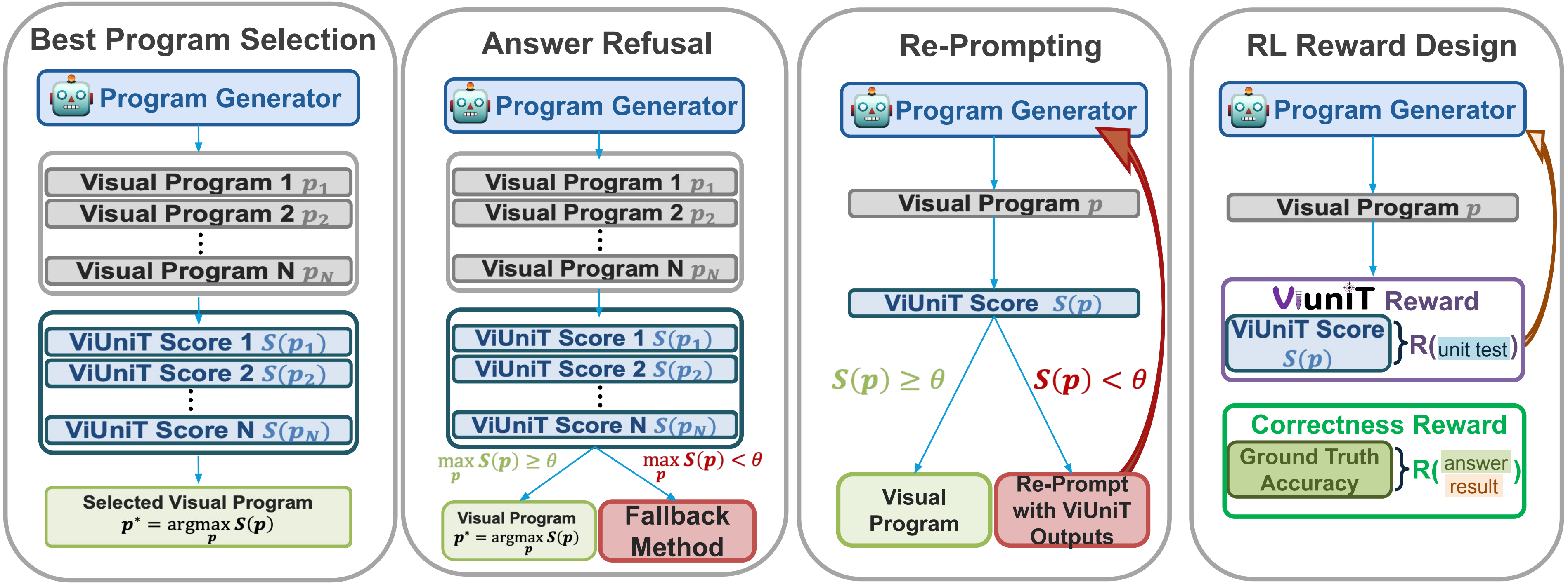
What is a Visual Unit Test?

Given a user query about an image, a unit test consists of an **image** and and **expected answer** to that query.



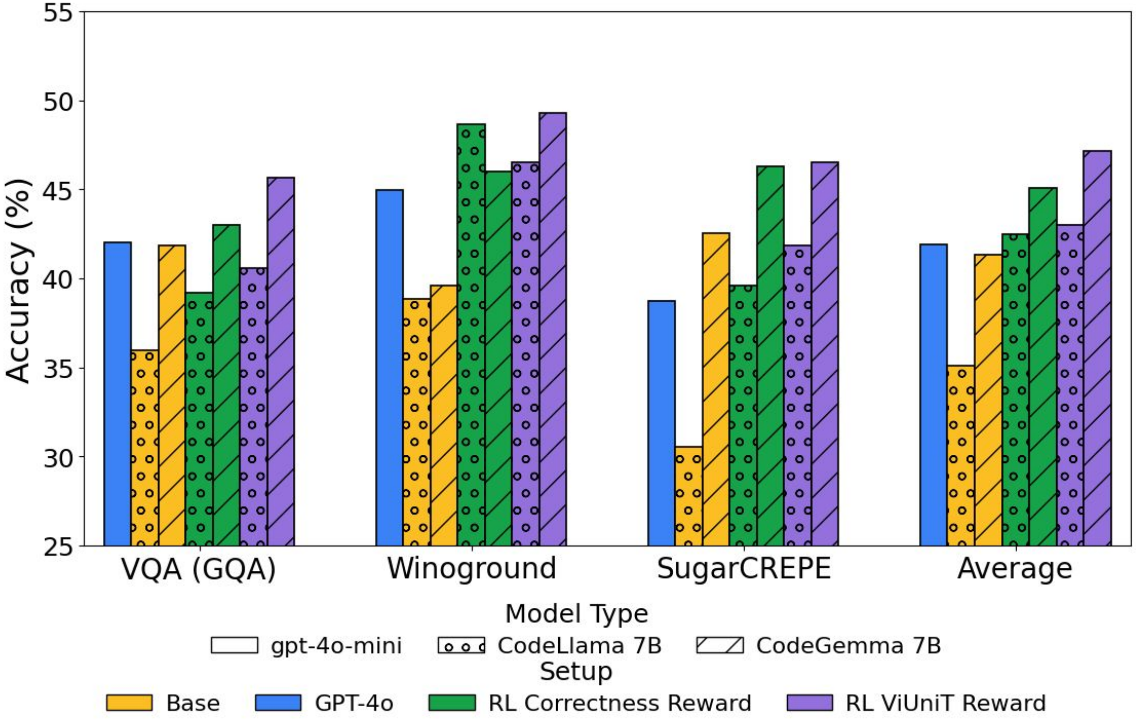

ViUniT Framework Overview
Program Generator
Query: Is there an elephant in the blue water?
Unit Test Generator
Unit Test 1
Caption
A single elephant is swimming in the clear blue ocean
Answer
yes
Sampler
Visual Program
def execute_command(image):
image_patch = ImagePatch(image)
elephant_patches = image_patch.find("elephant")
blue_water_patches = image_patch.find("blue water")
is_elephant_in_water = len(elephant_patches) > 0 and\
len(blue_water_patches) > 0
return bool_to_yesno(is_elephant_in_water)
Captions and Expected Answers
Unit Test 1
Caption
A single elephant is swimming in the clear blue ocean
Answer
yes
Unit Test 2
Caption
An elephant is walking beside a bright blue river
Answer
no
Unit Test k
Caption
A blue elephant-shaped pool float floating in a swimming pool.
Answer
no
Text-to-Image Generator
Visual Unit Tests
Unit Test 1
yes
Unit Test 2
no
Unit Test k
no
Program Output
yes
yes
yes

Use Cases of Visual Unit Tests

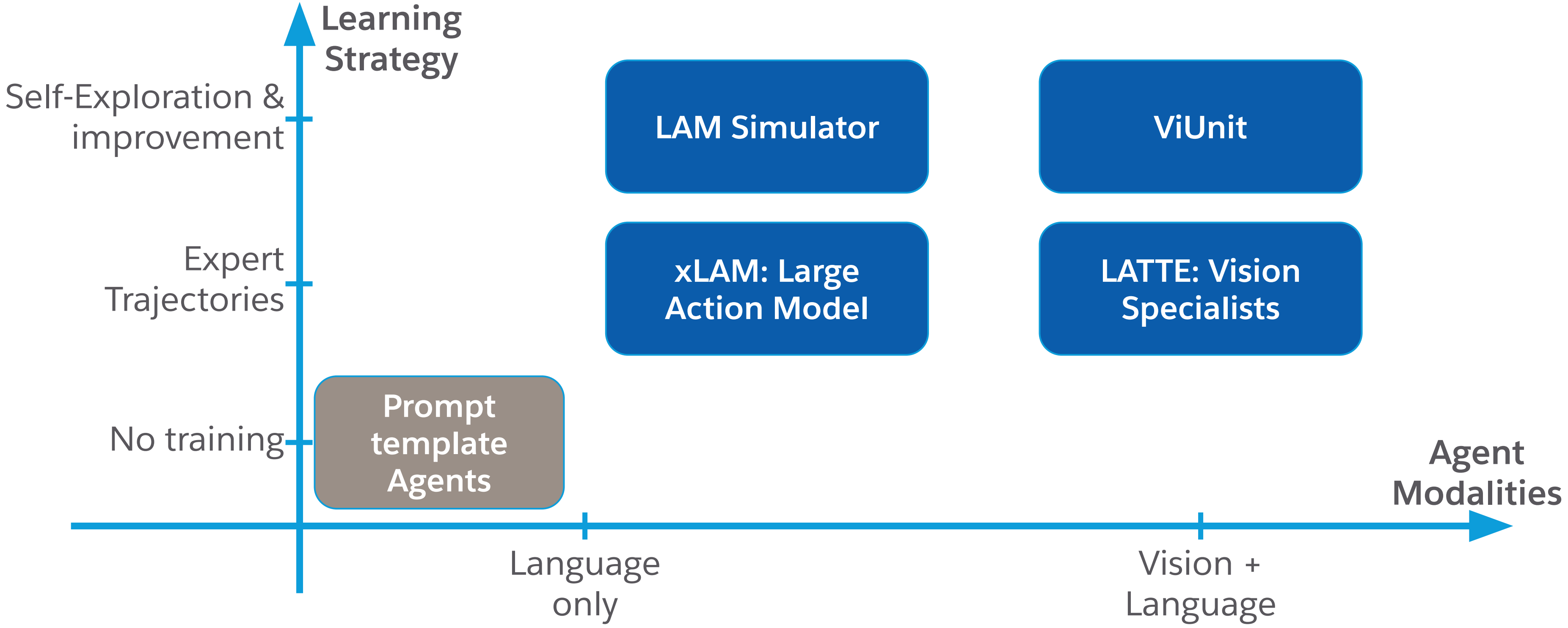


Key Result: Reinforcement Learning Reward

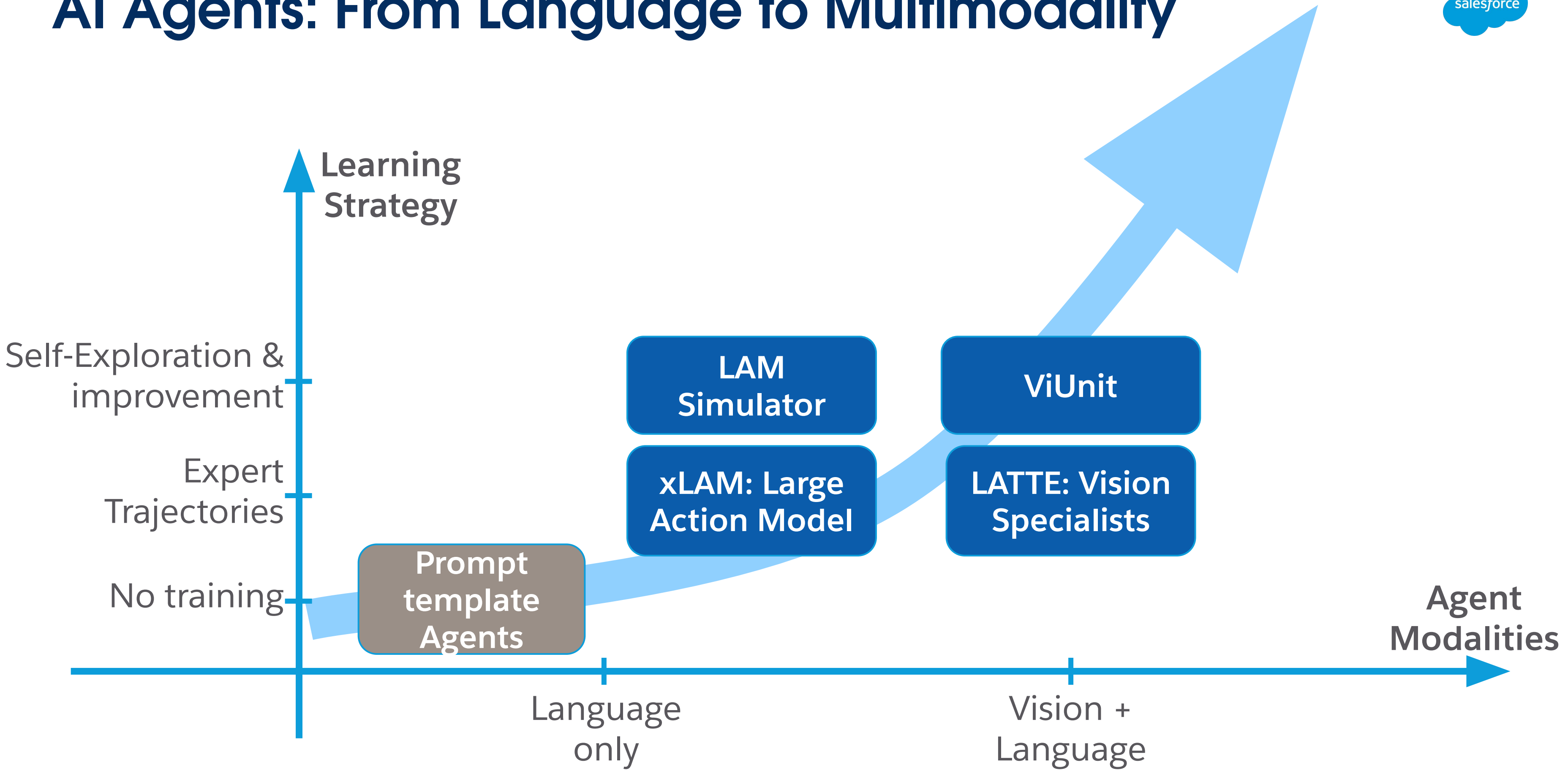
Unsupervised ViUniT Reward matches or surpasses supervised correctness reward.



AI Agents: From Language to Multimodality



AI Agents: From Language to Multimodality



References & Resources

- Zhang el al. xLAM: A Family of Large Action Models to Empower AI Agent Systems.NAACL 2025
- Ma et al. LATTE: Learning to Think with Vision Specialists. EMNLP 2025
- Hoang et al. LAM Simulator: Advancing Data Generation for Large Action Model Training via Online Exploration and Trajectory Feedback. NAACL 2025.
- Panagopoulou et al. ViUniT: Visual Unit Tests for More Robust Visual Programming. CVPR 2025
- Zhiwei Liu. “BOLAA: Benchmarking and orchestrating LLM-augmented autonomous agents”. ICLR 2024 Workshop on LLM Agents
- Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023

salesforce

Thank you