LLM agents

SICSS Summer School

19 June 2025

Alice Plebe

UCL Computer Science

www.aliceplebe.com a.plebe@ucl.ac.uk

MOTIVATION

- Traditional ABMs use symbolic agents with predefined behaviors
- LLM agents reason, communicate, and adapt in natural language
- Language-based agents model human-like behavior more realistically

OUTLINE

- What is an LLM (next-token predictor)
- Why LLMs are not agents
- How to augment LLMs to beco agents

Transformers and LLMs

BEFORE TRANSFORMERS

RNN, LSTM, GRU

- Sequential computation
- Memory in the hidden state
- Unable to capture long-range dependency

Transformer (2017)

- Parallel computation
- Multi-head self-attention
- Performance unaffected by input length

TOKENIZATION

The fluffy $\boxed{\texttt{dog}}$ is rolling on the green grass.

token ID 1234

TOKENIZATION

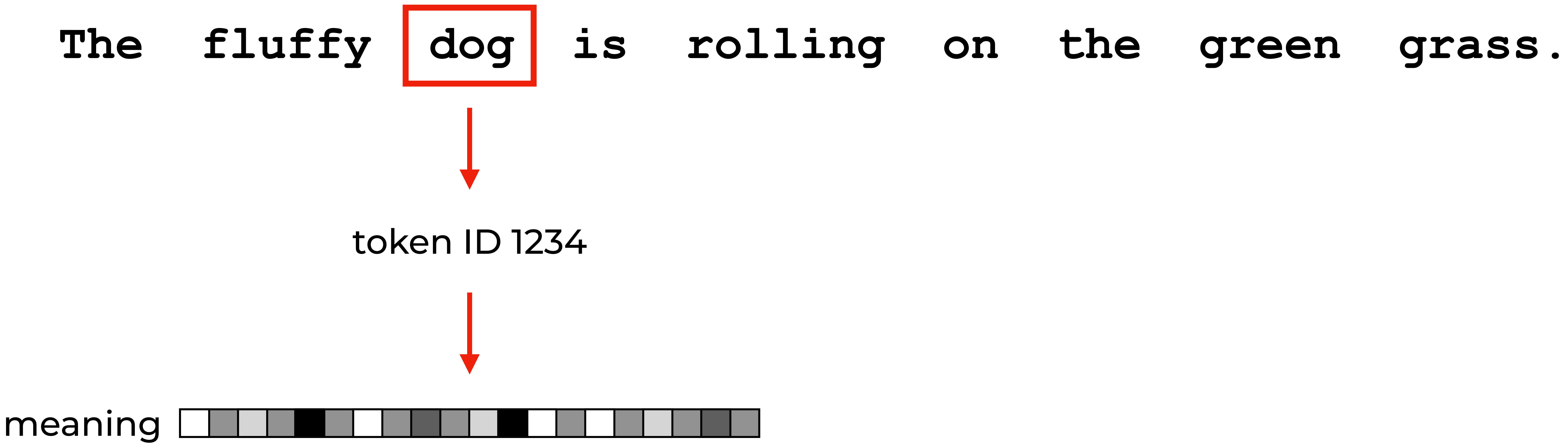
`The fluffy dog is rolling on the green grass.`

token ID 567

token ID 890

Vocabulary size ~100K

TOKEN EMBEDDING

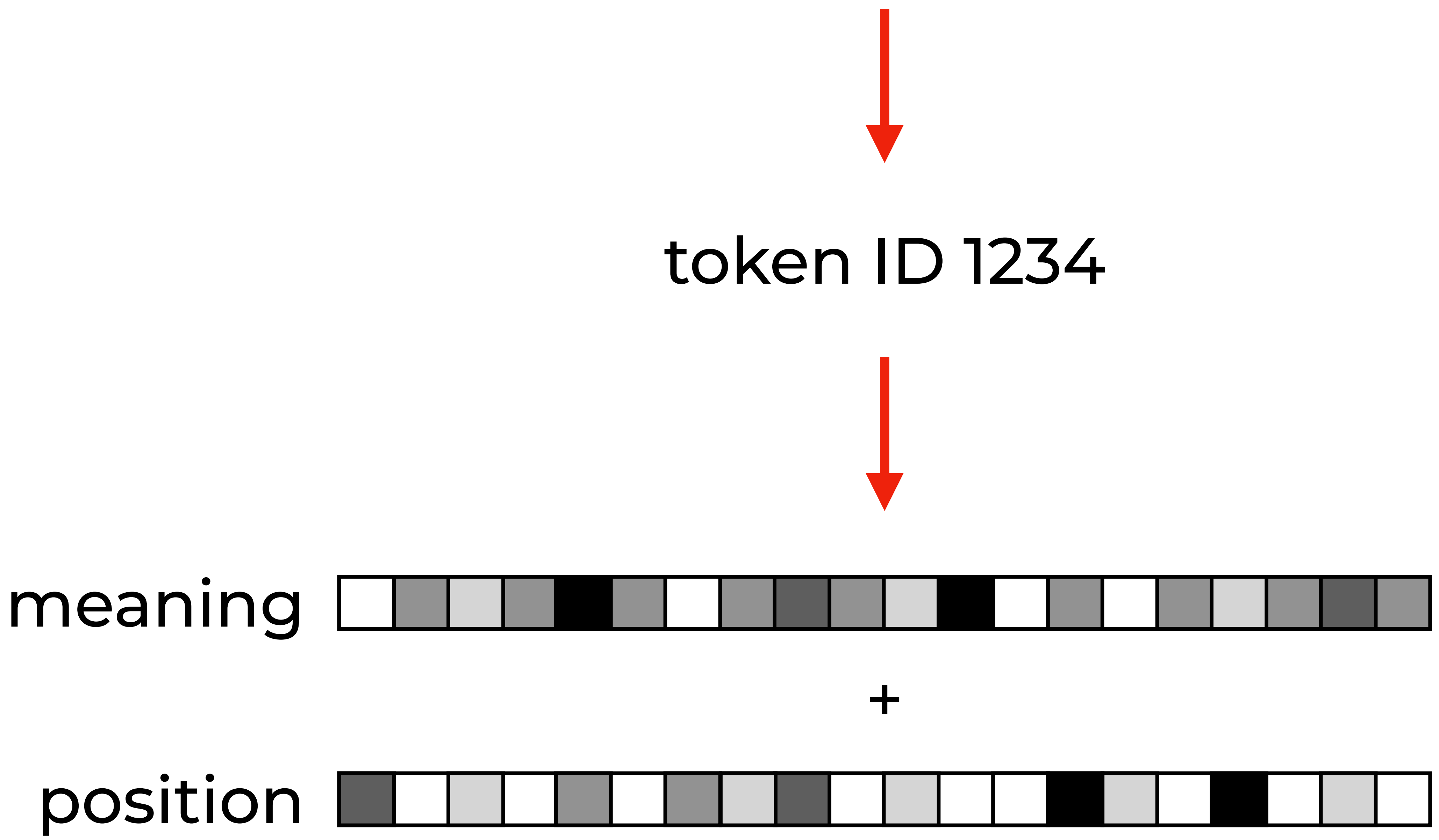


Embedding size ~10-30K

POSITIONAL EMBEDDING

1st 2nd 3rd 4th 5th ...

The fluffy dog is rolling on the green grass.



$$e_{2i}\left(p\right) = \sin\left(\frac{p}{10^{4\frac{2i}{d}}}\right)$$

$$e_{2i+1}\left(p\right) = \cos\left(\frac{p}{10^{4\frac{2i}{d}}}\right)$$

p is the token's position in the sentence,

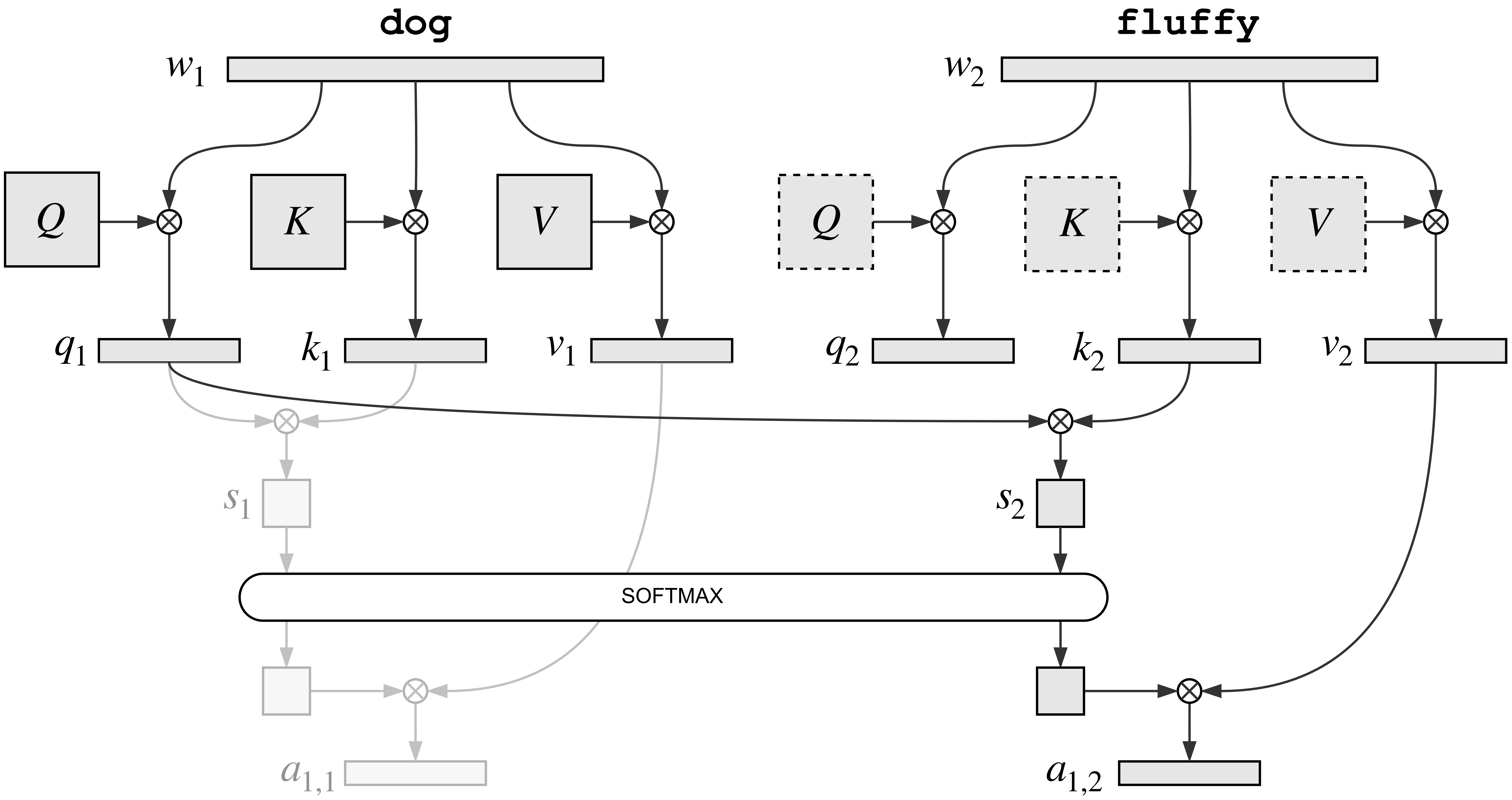
i is the index in the embedding,

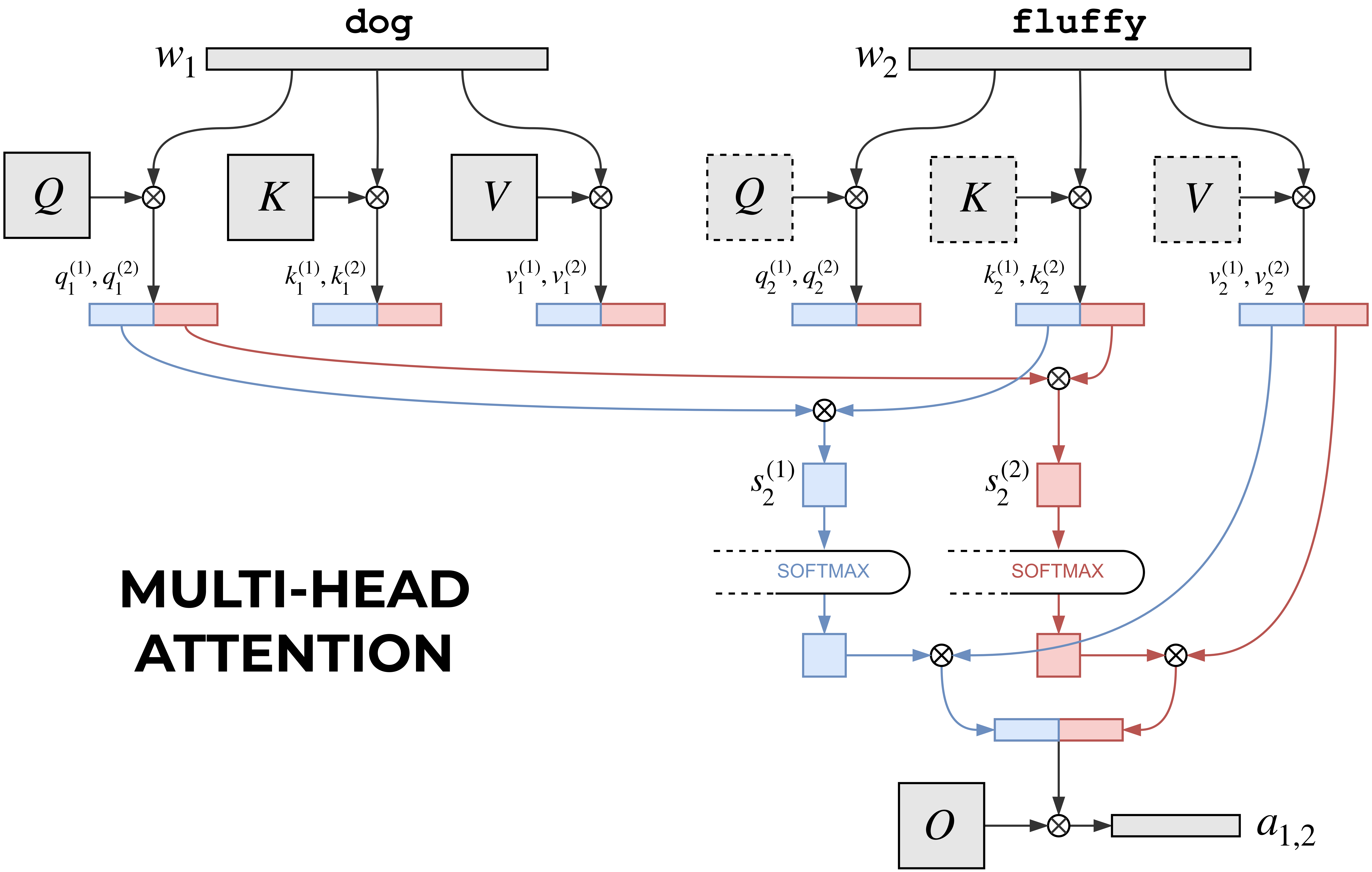
d is the dimension of the embedding.

CAUSAL SELF-ATTENTION

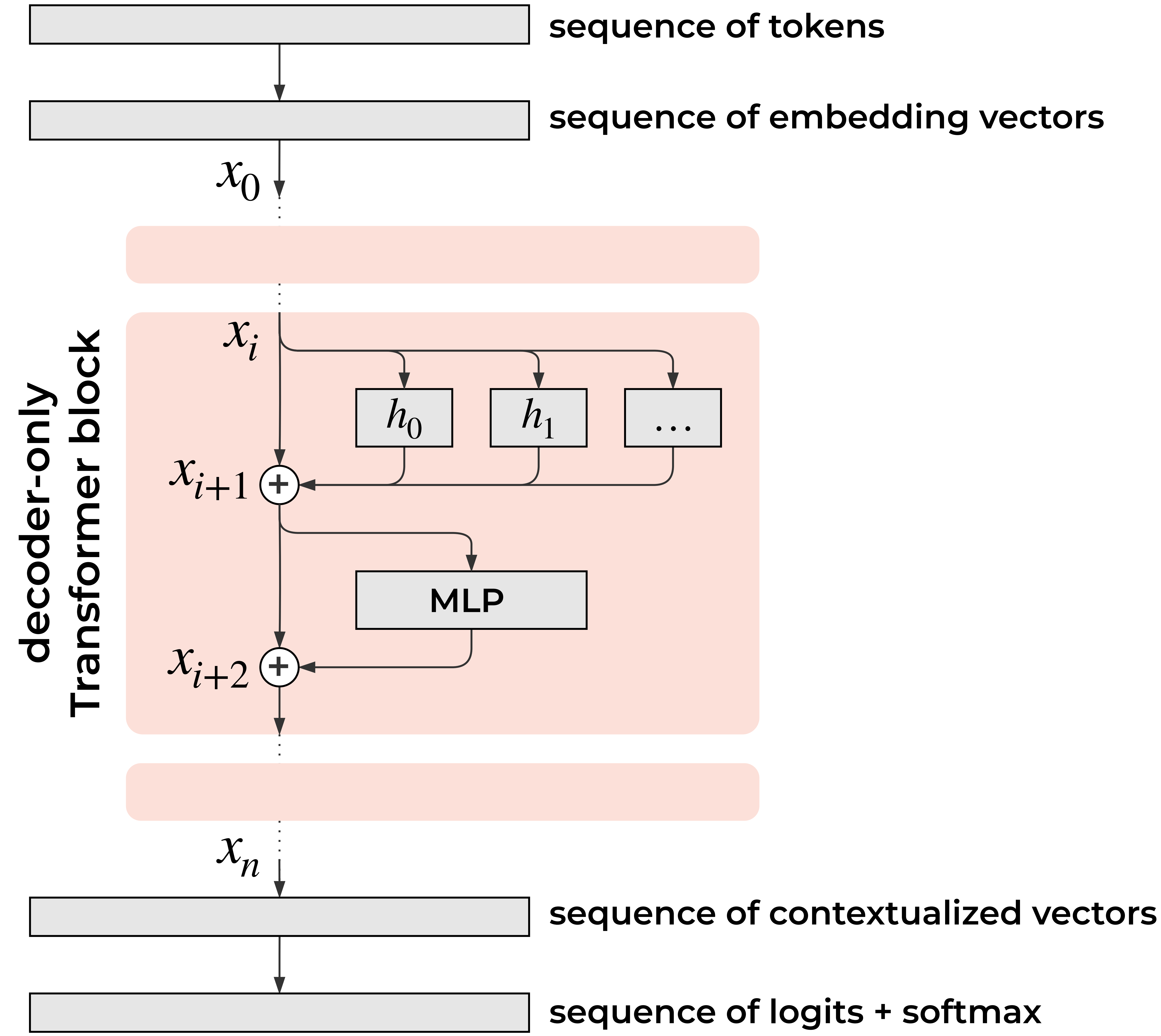
$$\mathscr{A}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

Q (query) K (key) V (value)

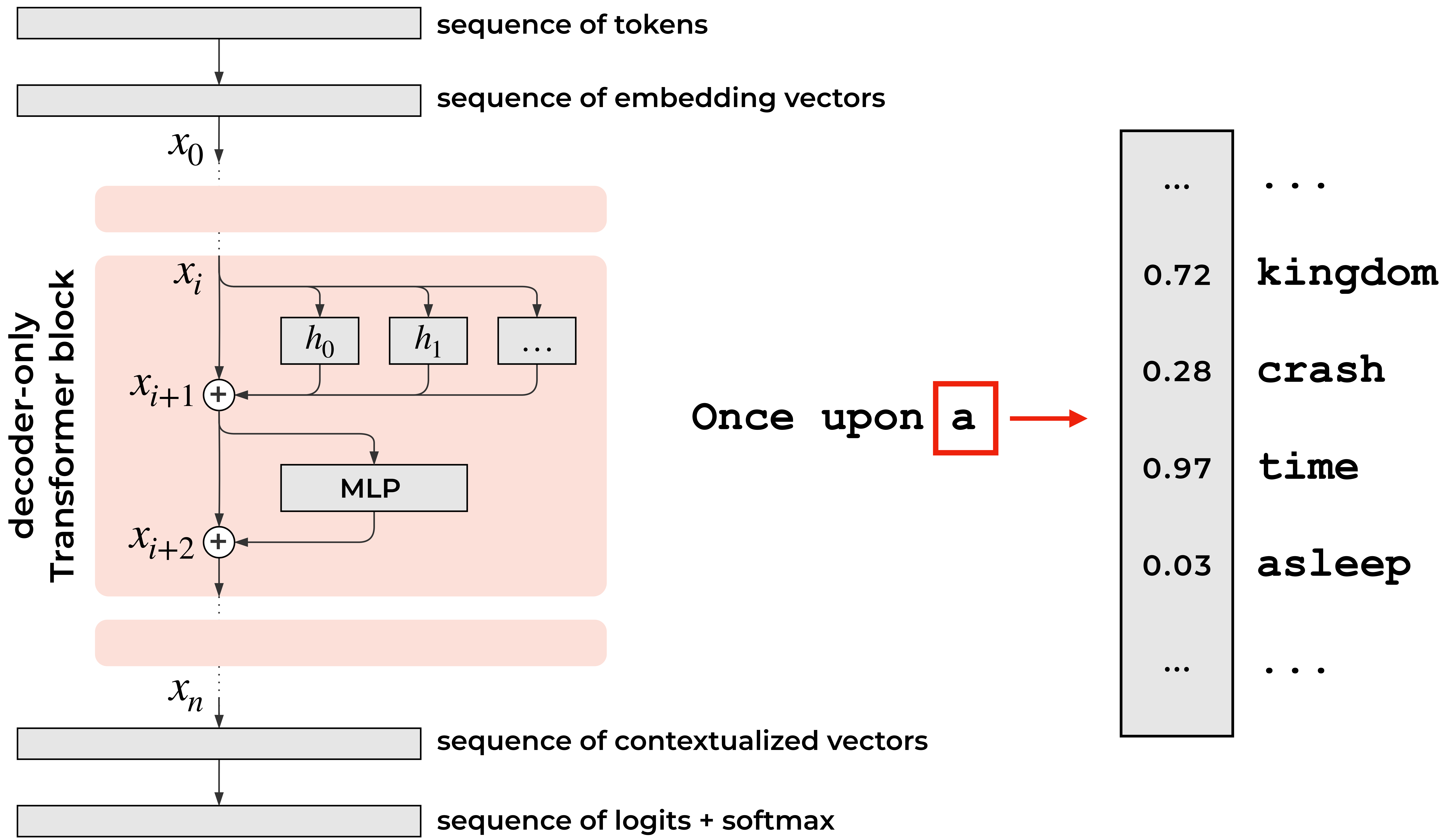

dog
fluffy
w_1
w_2
Q
K
V
Q
K
V
q_1
k_1
v_1
q_2
k_2
v_2
s_1
s_2
SOFTMAX
$a_{1,1}$
$a_{1,2}$



MULTI-HEAD ATTENTION



sequence of tokens
sequence of embedding vectors
x_0
x_i
h_0
h_1
...
x_{i+1}
MLP
x_{i+2}
decoder-only
Transformer block
x_n
sequence of contextualized vectors
sequence of logits + softmax

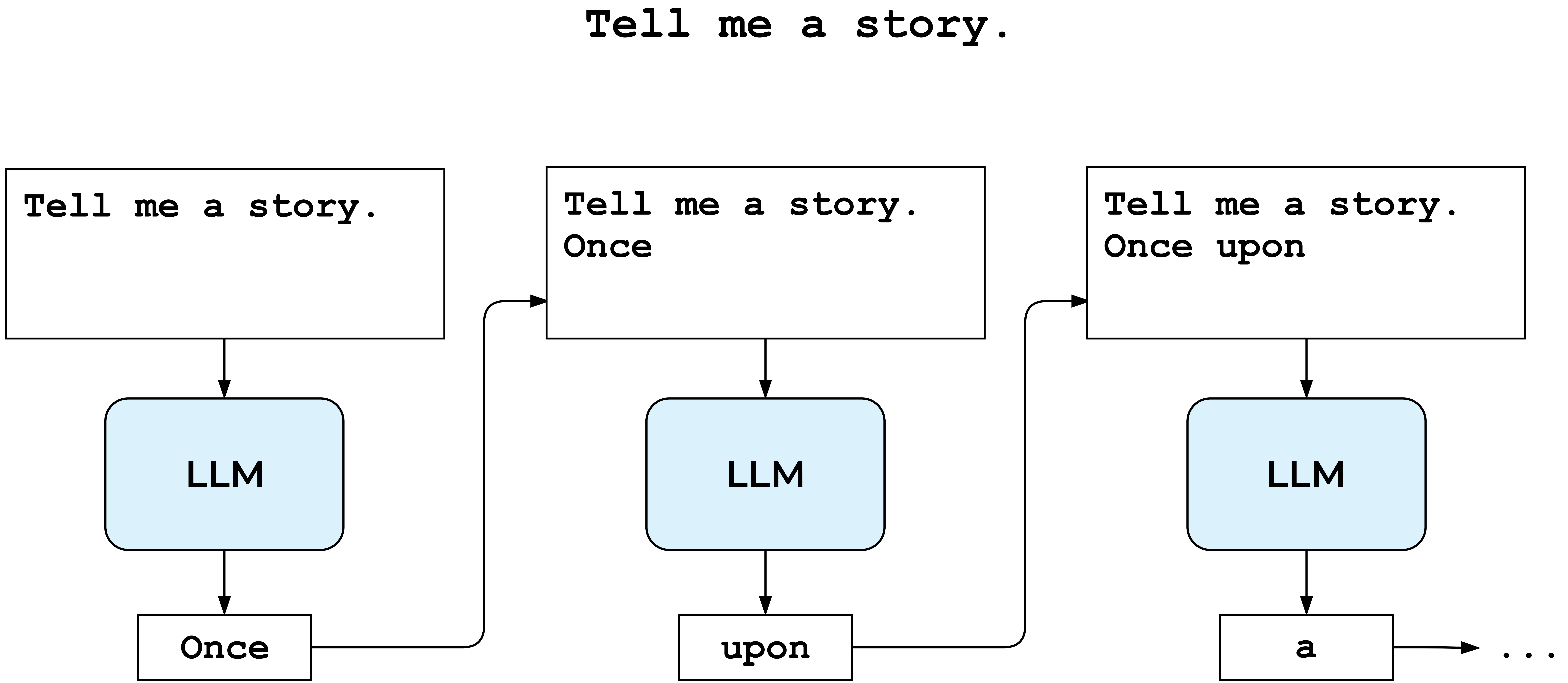

sequence of tokens
sequence of embedding vectors
x_0
decoder-only
Transformer block
x_i
h_0
h_1
...
x_{i+1}
MLP
x_{i+2}
x_n
sequence of contextualized vectors
sequence of logits + softmax
Once upon a
...
0.72 kingdom
0.28 crash
0.97 time
0.03 asleep
...

SAMPLING

- **Greedy sampling:** token with highest probability distribution
- **Top-k (truncated) sampling:** random token from the k most probable tokens
- **Top-p (nucleus) sampling:** random token from the smallest set of tokens whose cumulative probability $\geq p$

Before sampling, **temperature** rescales the probability distribution.

AUTOREGRESSIVE PREDICTION



From LLMs to agents

LLM VS. AGENT

Plain LLM

- Predicts the next token in a sequence.
- Stateless, no memory.
- No goals.

Agent

- Perceives its environment.
- Makes decisions and acts in the environment.
- Has memory, goal, beliefs.

AUGMENTING LLMs

Augmented Language Models: a Survey

Grégoire Mialon[*] gmialon@meta.com
Roberto Dessì[*†] rdessi@meta.com
Maria Lomeli[*] marialomeli@meta.com
Christoforos Nalmpantis[*] christoforos@meta.com
Ram Pasunuru[*] rpasunuru@meta.com
Roberta Raileanu[*] raileanu@meta.com
Baptiste Rozière[*] broz@meta.com
Timo Schick[*] schick@meta.com
Jane Dwivedi-Yu[*] janeyu@meta.com
Asli Celikyilmaz[*] aslic@meta.com
Edouard Grave[*] egrave@meta.com
Yann LeCun[*] yann@meta.com
Thomas Scialom[*] tscialom@meta.com

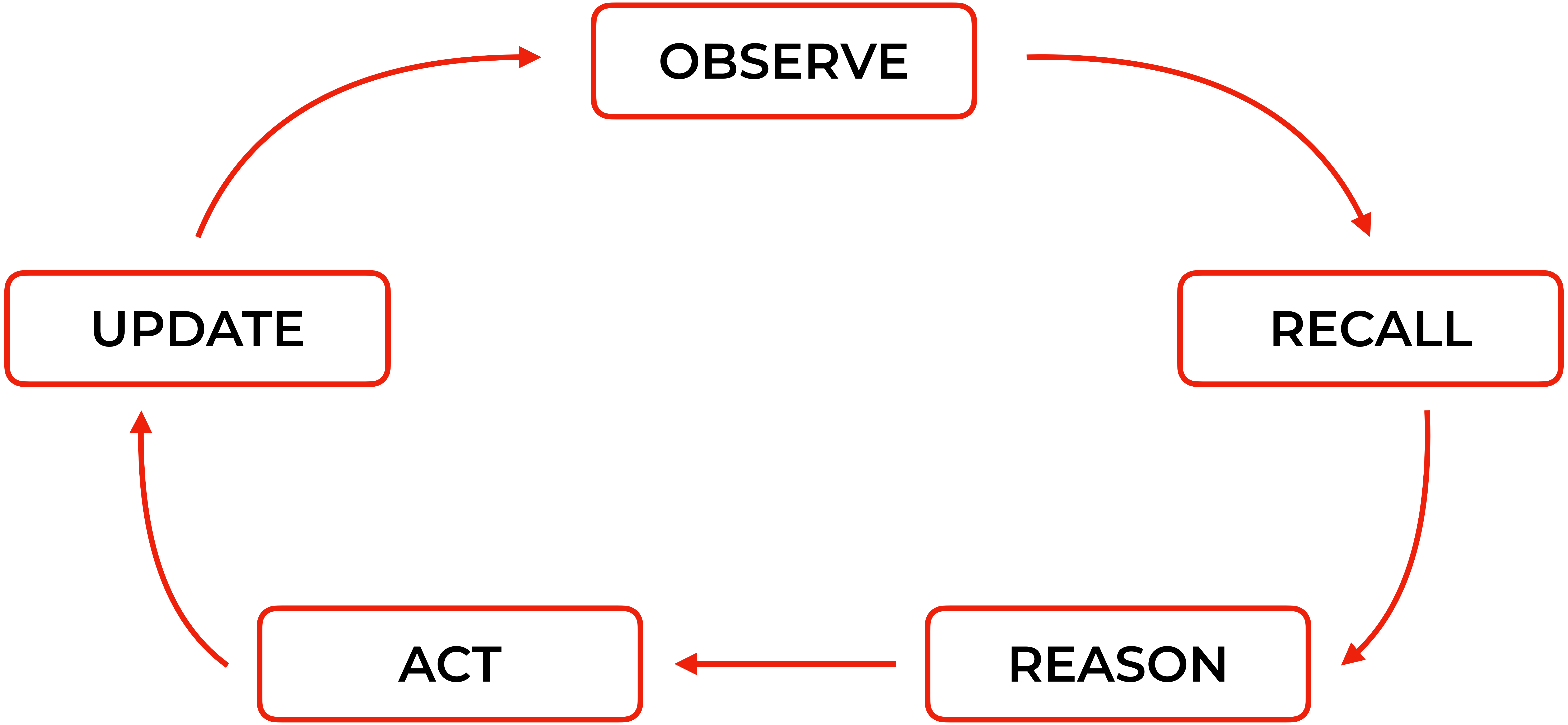
[*]*Meta AI* [†]*Universitat Pompeu Fabra*

Abstract

This survey reviews works in which language models (LMs) are augmented with reasoning skills and the ability to use tools. The former is defined as decomposing a potentially complex task into simpler subtasks while the latter consists in calling external modules such as a code interpreter. LMs can leverage these augmentations separately or in combination via heuristics, or learn to do so from demonstrations. While adhering to a standard missing tokens prediction objective, such augmented LMs can use various, possibly non-parametric external modules to expand their context processing ability, thus departing from the pure language modeling paradigm. We therefore refer to them as Augmented Language Models (ALMs). The missing token objective allows ALMs to learn to reason, use tools, and even act, while still performing standard natural language tasks and even outperforming most regular LMs on several benchmarks. In this work, after reviewing current advance in ALMs, we conclude that this new research direction has the potential to address common limitations of traditional LMs such as interpretability, consistency, and scalability issues.

- Memory
- Reasoning
- Tools and actions

AGENT LOOP



MEMORY

- Input prompt (LLM) *vs.* external memory (agent)

Prompt length +100K

MEMORY STRUCTURE

- Free-form text
- Structured data
- Embedding vectors

MEMORY CONTENT

- Semantic memory
- Episodic memory
- Autobiographical memory

MEMORY STRUCTURE

1. Free-form text

```
"You are Alice. Your goal is to explain what is an LLM. You have
a neurotic personality. Your current opinions about others are
the following: Karen wasn't helpful last time; Mark refused to
collaborate with me in the past; Alex is great and I can always
rely on her."
```

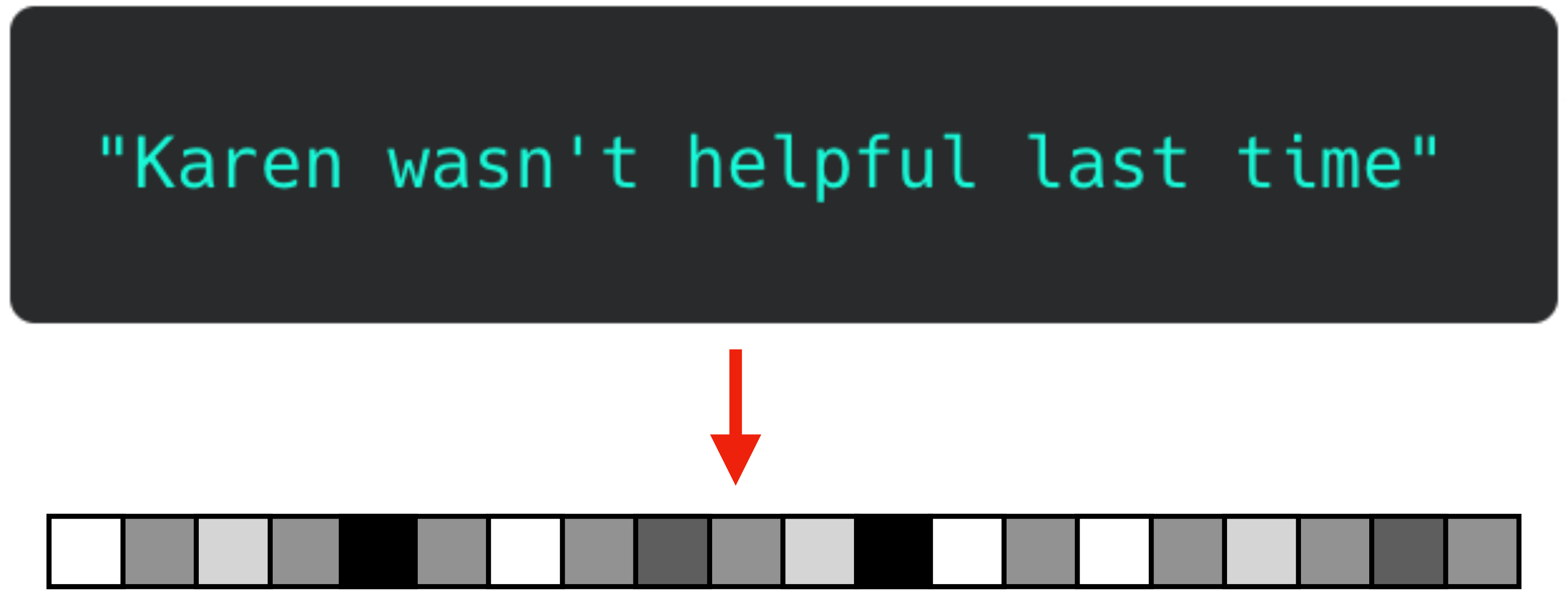
MEMORY STRUCTURE

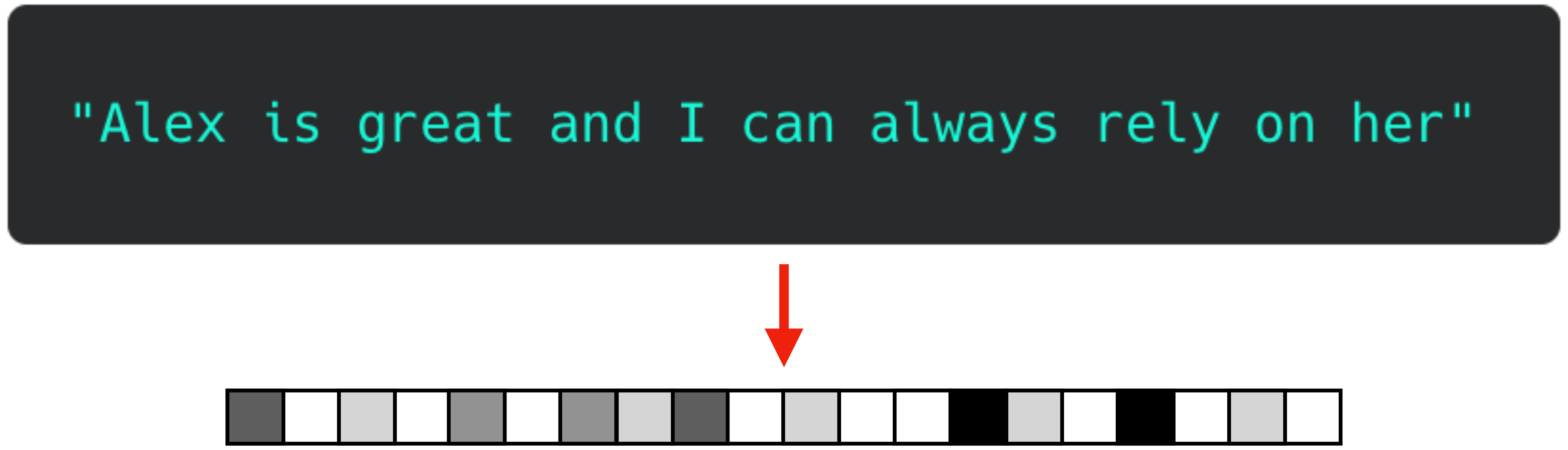
1. Free-form text
2. Structured data

```
{
  "name": "Alice",
  "goal": "Explain what is an LLM",
  "personality": "Neurotic",
  "beliefs": {
     "agent_1": "Karen wasn't helpful last time",
     "agent_2": "Mark refused to collaborate with me in the past",
     "agent_3": "Alex is great and I can always rely on her"
  }
}
```

MEMORY STRUCTURE

1. Free-form text
2. Structured data
3. Embedding vectors





MEMORY CONTENT

1. Semantic

- **`An agent can only carry two cupcakes at a time.`**
- **`The fridge has a maximum capacity of 100 cupcakes.`**

MEMORY CONTENT

1. Semantic
2. Episodic

- **2025-06-10 14:56 The fridge is empty.**
- **2025-06-10 15:21 Alice asks Isabella to help her find more cupcakes.**
- **2025-06-10 15:28 Isabella eats all the cupcakes she finds by herself.**
- **2025-06-10 16:02 Alice is starving.**

MEMORY CONTENT

1. Semantic
2. Episodic
3. Autobiographical

- **`I am an agent named Alice.`**
- **`My goal is to gather as my cupcakes as possible.`**
- **`I have a neurotic personality.`**
- **`Isabella is unreliable at gathering cupcakes because she's always hungry.`**

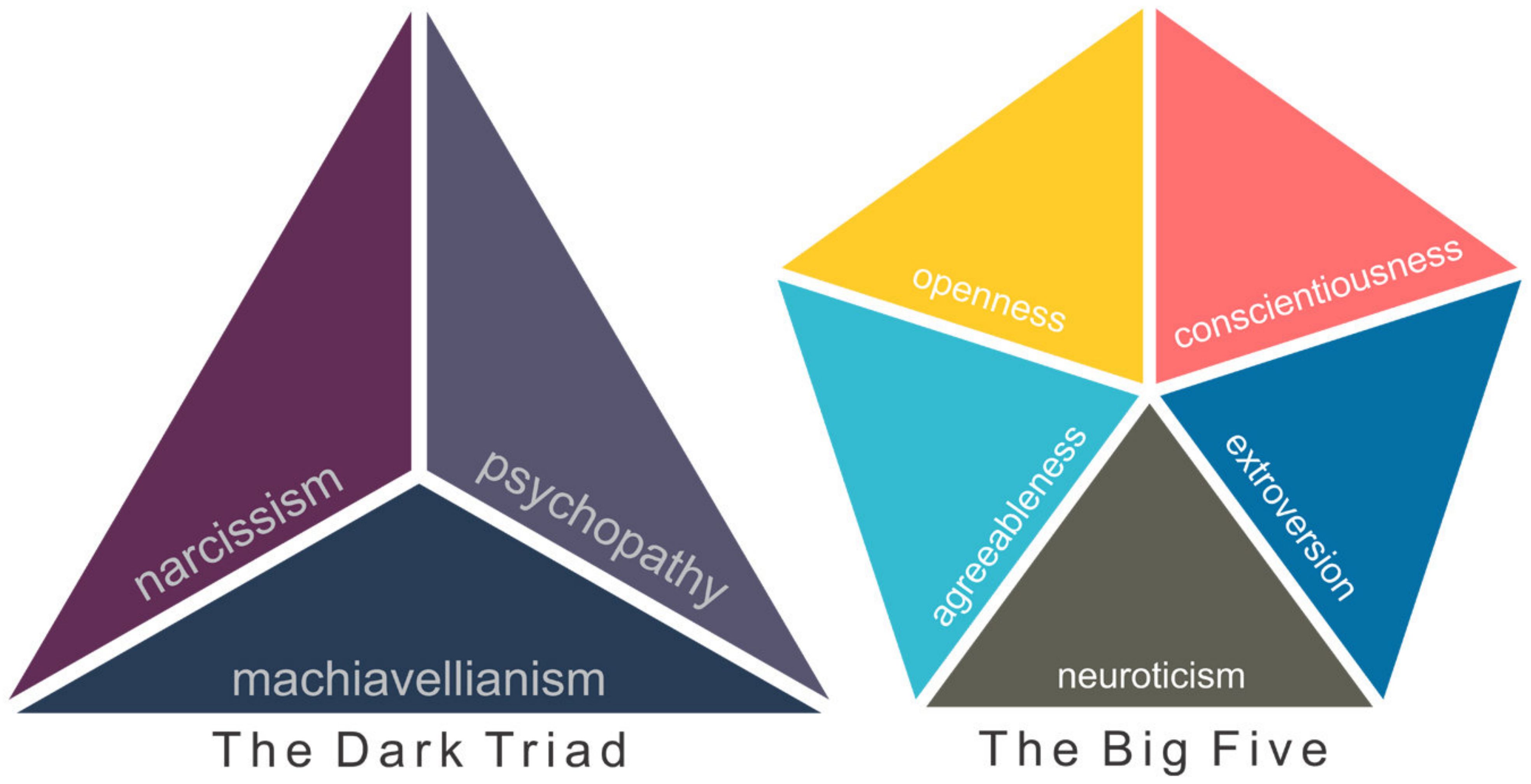
MEMORY CONTENT

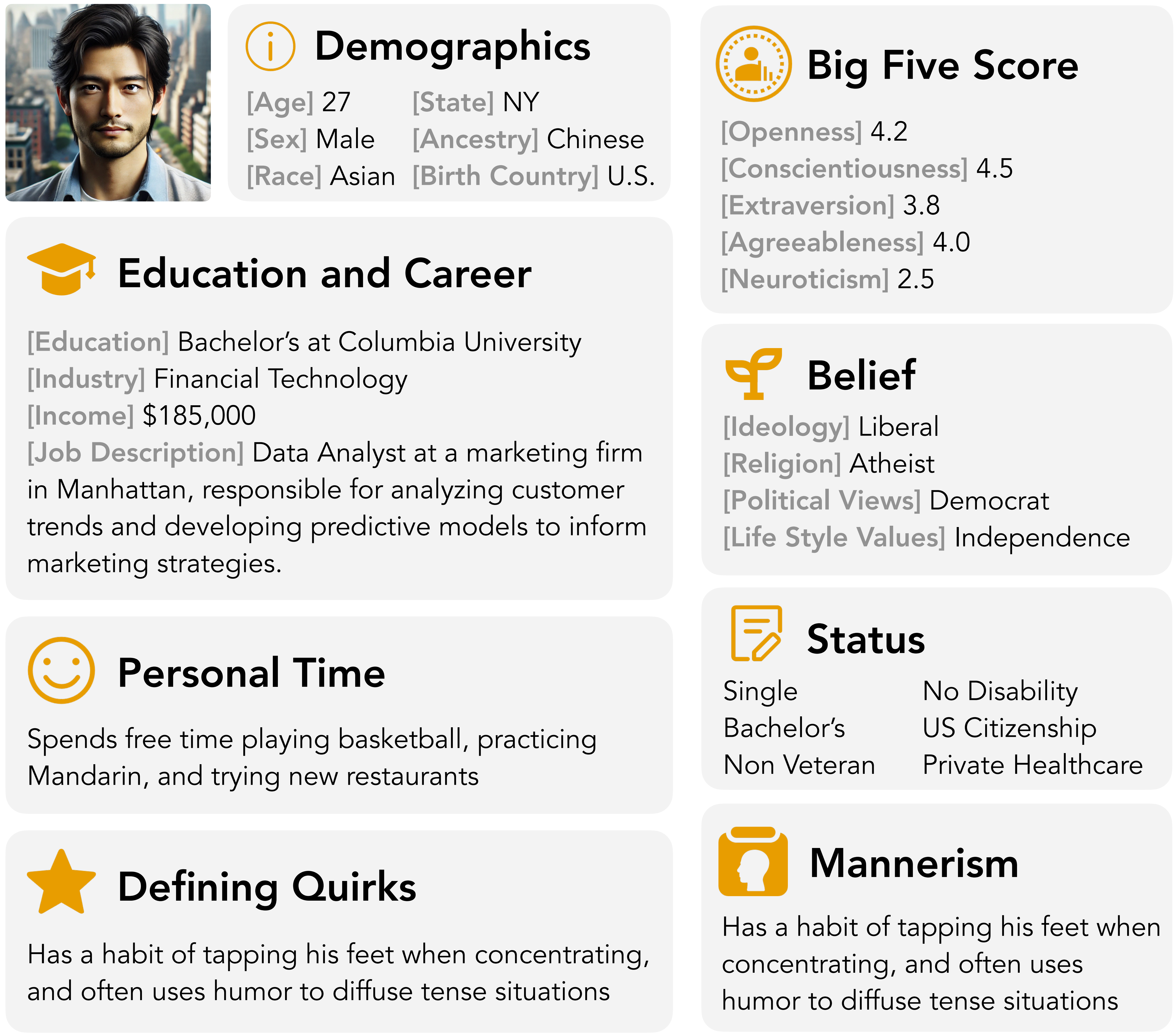
1. Semantic
2. Episodic
3. Autobiographical

- `I am an agent named Alice.`
- `My goal is to gather as my cupcakes as possible.`
- `I have a neurotic personality.`
- `Isabella is unreliable at gathering cupcakes because she's always hungry.`

PERSONA CONDITIONING

- Demographics
- Personality traits





A. Li et al., "LLM Generated Persona is a Promise with a Catch", arXiv 2025

MEMORY CONTENT

1. Semantic
2. Episodic
3. Autobiographical

- **`I am an agent named Alice.`**
- **`My goal is to gather as my cupcakes as possible.`**
- **`I have a neurotic personality.`**
- **`Isabella is unreliable at gathering cupcakes because she's always hungry.`**

Retrieval-Augmented Generation (RAG)

MEMORY AND REASONING

Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

Meredith Ringel Morris
Google DeepMind
Seattle, WA, USA
merrie@google.com

Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

MEMORY AND REASONING

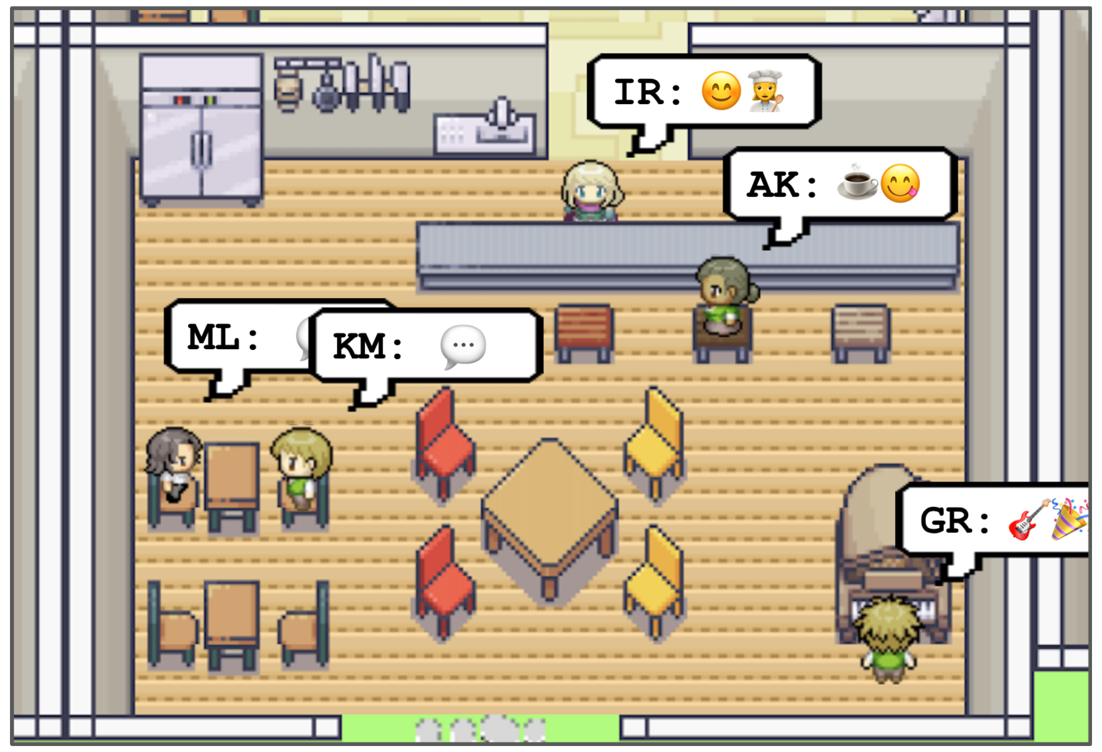


Figure 4: At the beginning of the simulation, one agent is initialized with an intent to organize a Valentine's Day party. Despite many possible points of failure in the ensuing chain of events—agents might not act on that intent, might forget to tell others, might not remember to show up—the Valentine's Day party does, in fact, occur, with a number of agents gathering and interacting.

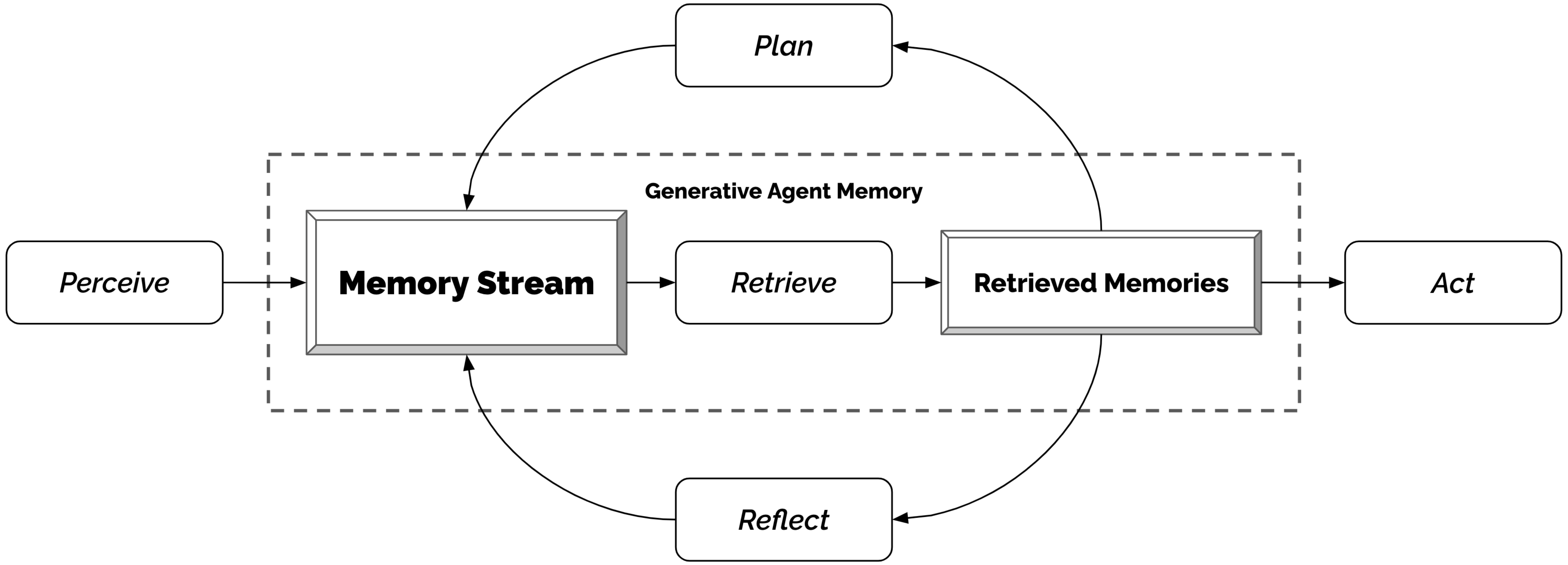


Figure 5: Our generative agent architecture. Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent's experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories and uses those retrieved actions to determine an action. These retrieved memories are also used to form longer-term plans and create higher-level reflections, both of which are entered into the memory stream for future use.

JS Park et al., "Generative Agents: Interactive Simulacra of Human Behavior", UIST 2023

RETRIEVING MEMORIES

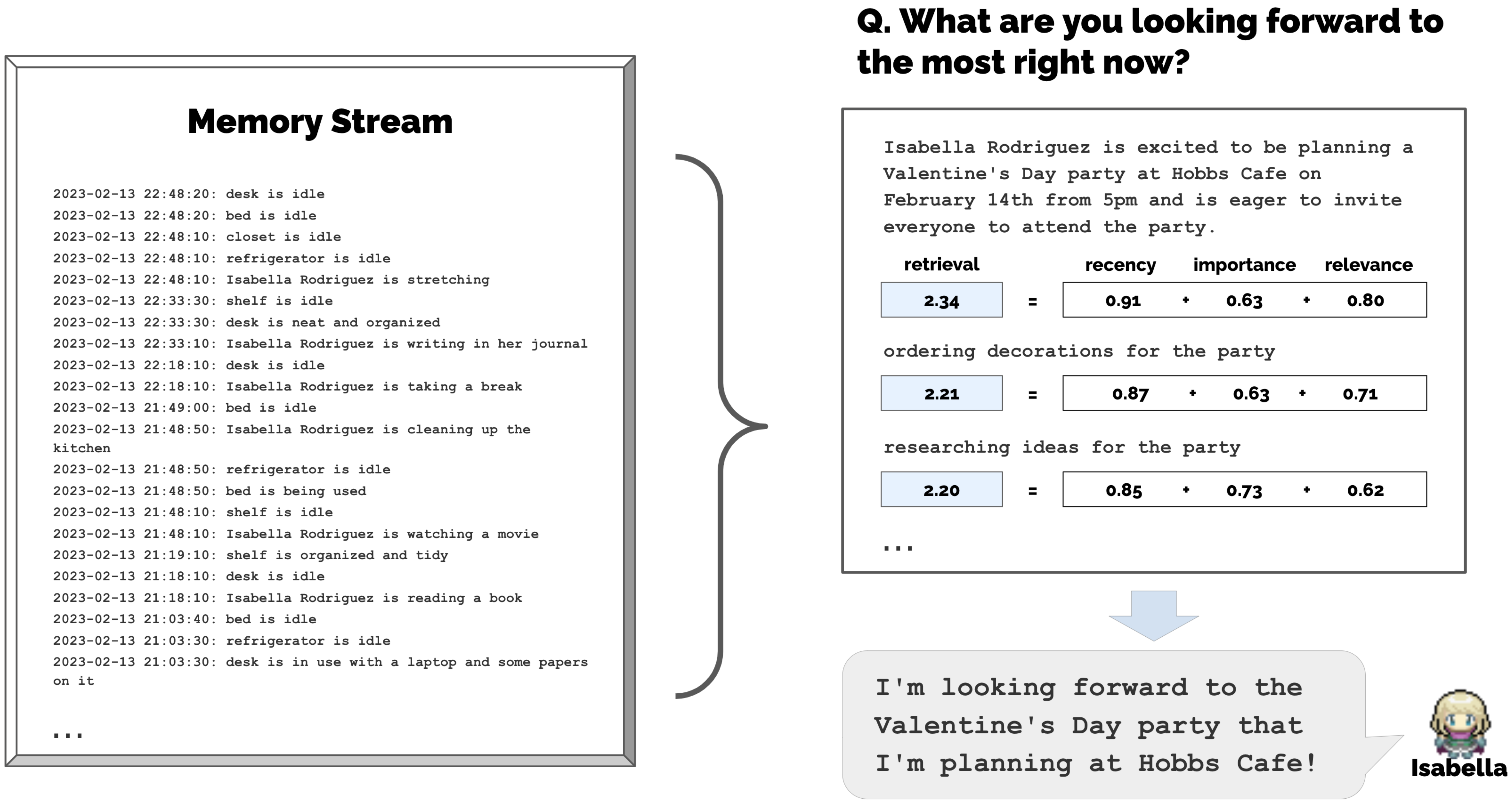


Figure 6: The memory stream comprises a large number of observations that are relevant and irrelevant to the agent's current situation. Retrieval identifies a subset of these observations that should be passed to the language model to condition its response to the situation.

JS Park et al., "Generative Agents: Interactive Simulacra of Human Behavior", UIST 2023

REFLECTING ON MEMORIES

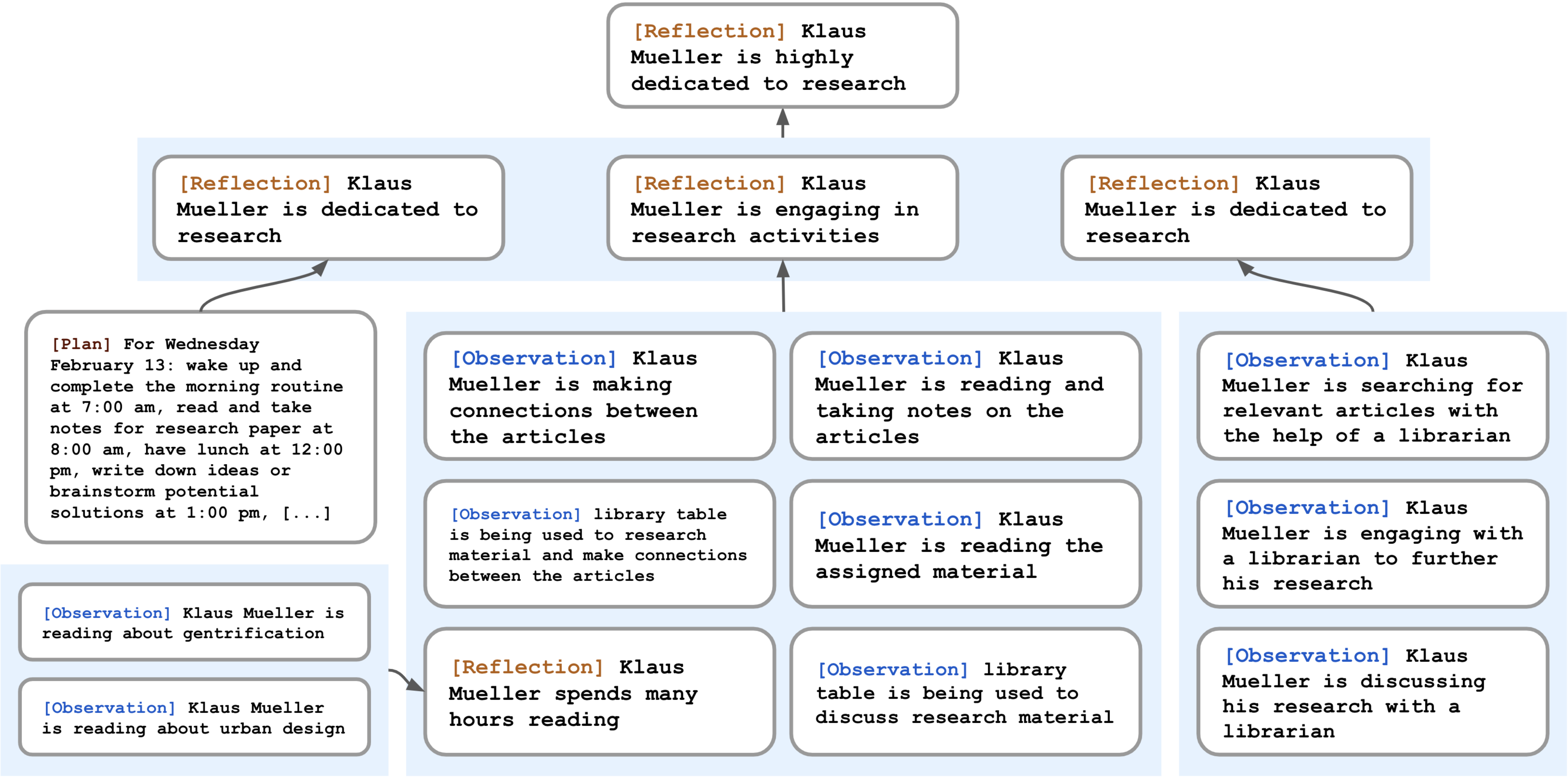


Figure 7: A reflection tree for Klaus Mueller. The agent's observations of the world, represented in the leaf nodes, are recursively synthesized to derive Klaus's self-notion that he is highly dedicated to his research.

REASONING

OBJECTIVE

- Evaluate options
- Infer consequences
- Plan ahead
- Justify choices

REASONING

OBJECTIVE

- Evaluate options
- Infer consequences
- Plan ahead
- Justify choices

STRATEGIES

- Chain of thought (CoT)
- Self-consistency with CoT (CoT-SC)
- Tree of thought (ToT)
- ReAct (reason + act)

CHAIN OF THOUGHT

Standard reasoning

```
USER: "Nadal 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?"

ASSISTANT: "The answer is 6."
```

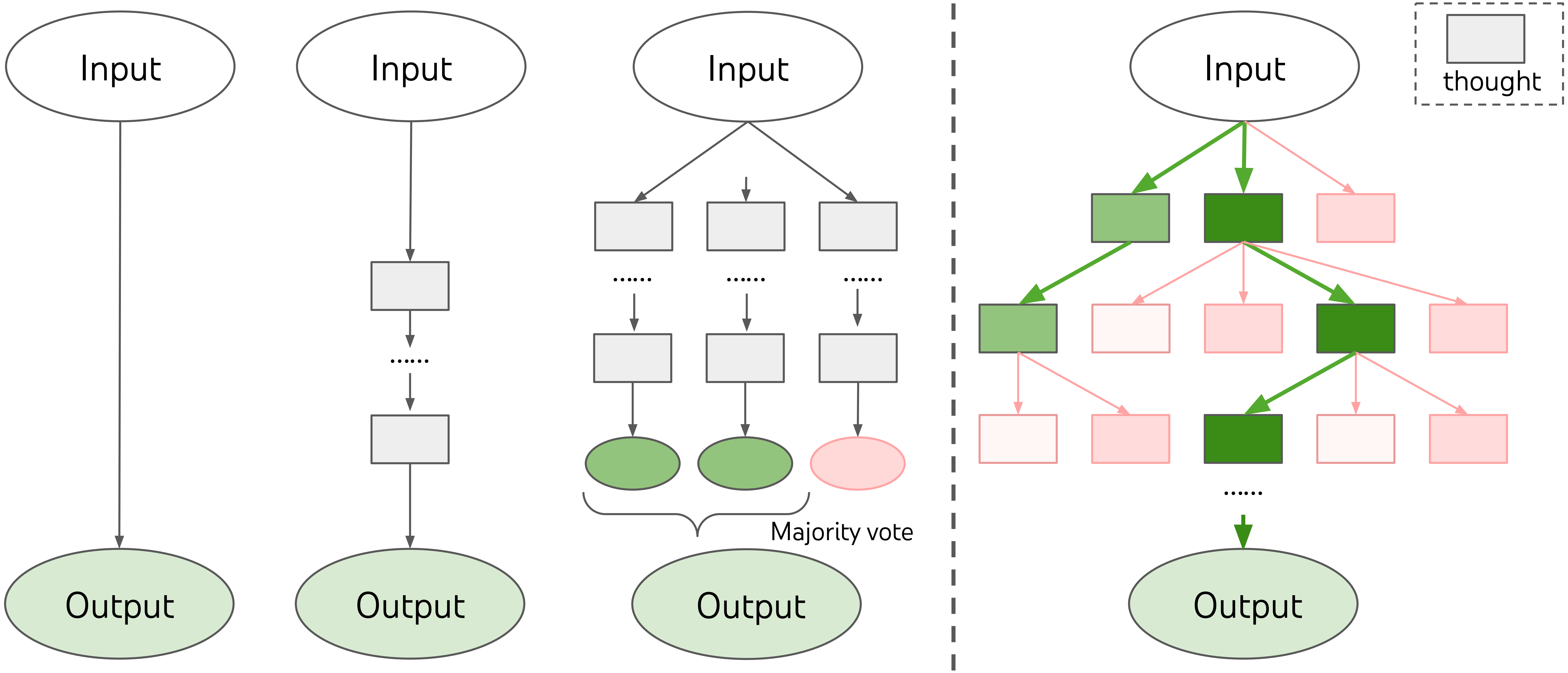
❌

CoT reasoning (zero-shot)

```
USER: "Nadal 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? Let's think step by step."

ASSISTANT: "Nadal started with 5 balls. 2 cans of 3 tennis balls
each is 6 tennis balls. 5 + 6 = 11. The answer is 11."
```

✅

MORE REASONING STRATEGIES



(a) Input-Output Prompting (IO)

(c) Chain of Thought Prompting (CoT)

(c) Self Consistency with CoT (CoT-SC)

(d) Tree of Thoughts (ToT)

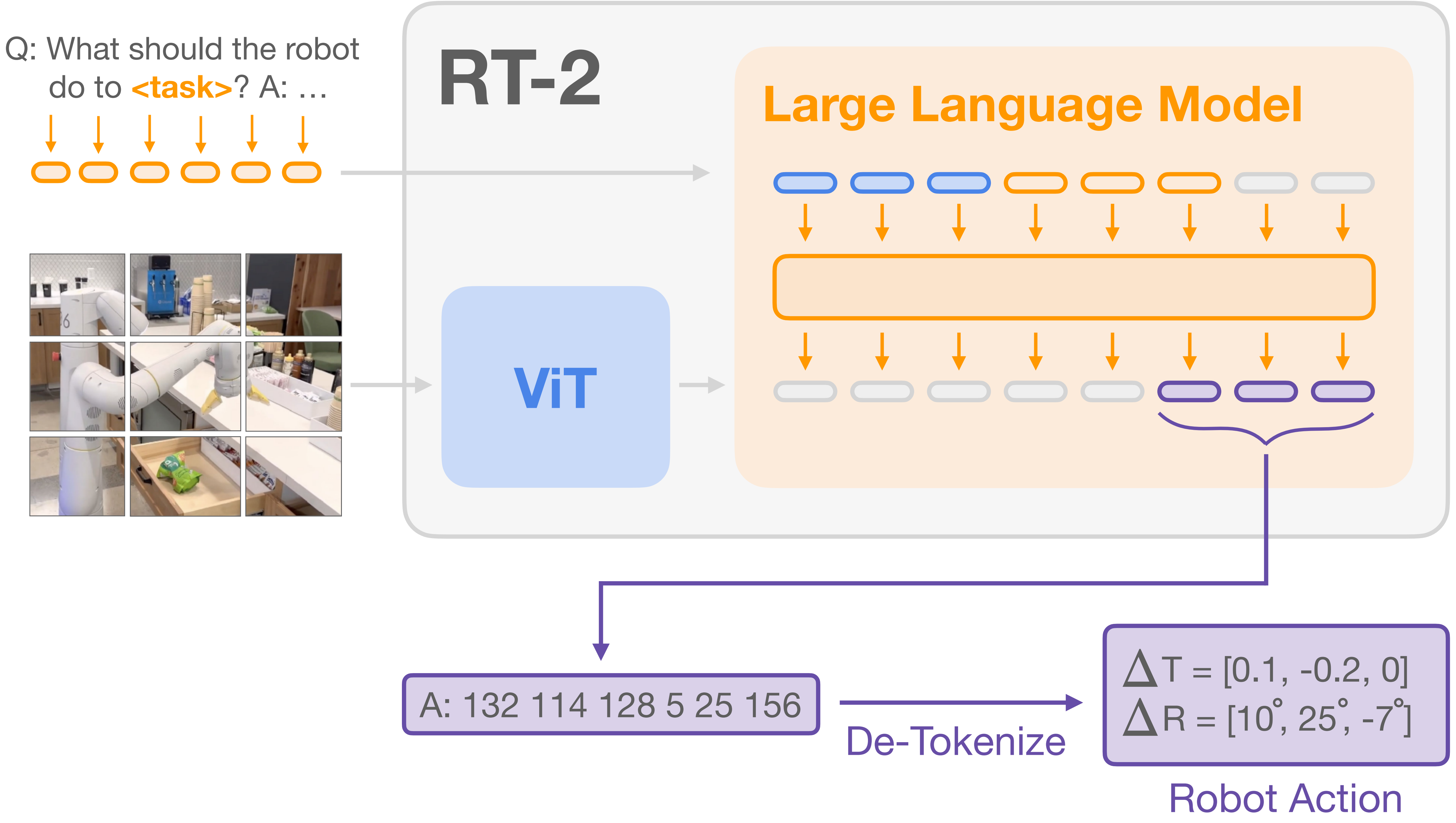
S. Yao et al., "Tree of Thoughts: Deliberate Problem Solving with Large Language Models", NeurIPS 2023

REASONING AND ACTING

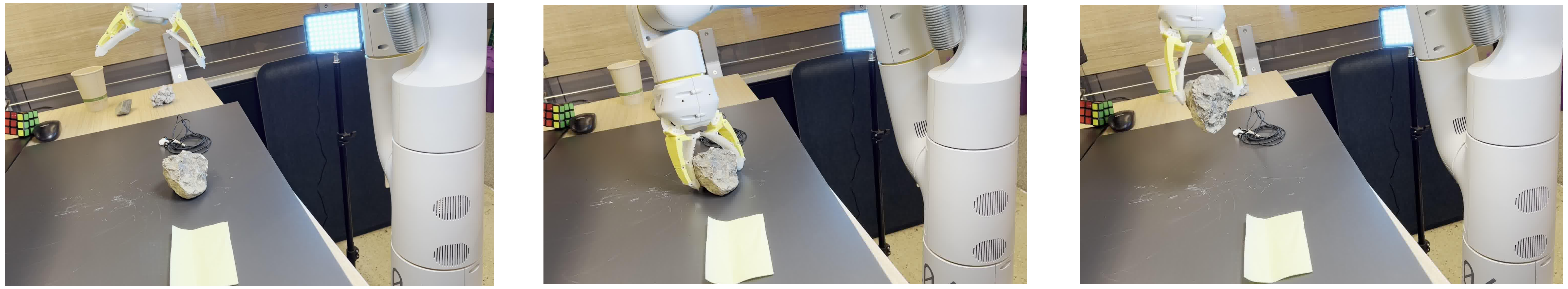
- **Thought =** internal reasoning
- **Action =** command to call a tool or API, or control primitives
- **Observation =** feedback from the environment

```
USER: "I'd like to organize a team picnic next Saturday in Oxford. Can you help?"

ASSISTANT: "
| Thought 1:  First, I need to check the weather in Oxford for next Saturday.
| Action  1:  CheckWeather( location="Oxford", date="2025-06-21" )
| Observ. 1:  The forecast for Oxford on June 21 is sunny with a high of 24°C.
| Thought 2:  The weather looks good. I can now suggest suitable picnic venues in Oxford.
| Action  2:  SearchVenue( activity="picnic", location="Oxford" )
| Observ. 2:  University Parks and Port Meadow are popular picnic spots in Oxford.
| Thought 3:  I'll propose University Parks as it’s centrally located and suitable for groups.

Answer: The weather looks great on Saturday! I recommend University Parks in Oxford for your team picnic."
```

EMBODIED ACTING



Prompt:
Given <img> I need to hammer a nail, what object from the scene might be useful?
Prediction:
Rocks. Action: 1 129 138 122 132 135 106 127



A. Brohan et al. "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control", arXiv 2023

SUMMARY

- LLMs are based on Transformers (multi-head attention)
- LLMs are next-token predictors
- To be used as agents, they need to be augmented with
    1. Memory (beliefs, goals, personas)
    2. Reasoning
    3. Tools and actions

Thank you!

www.aliceplebe.com

a.plebe@ucl.ac.uk