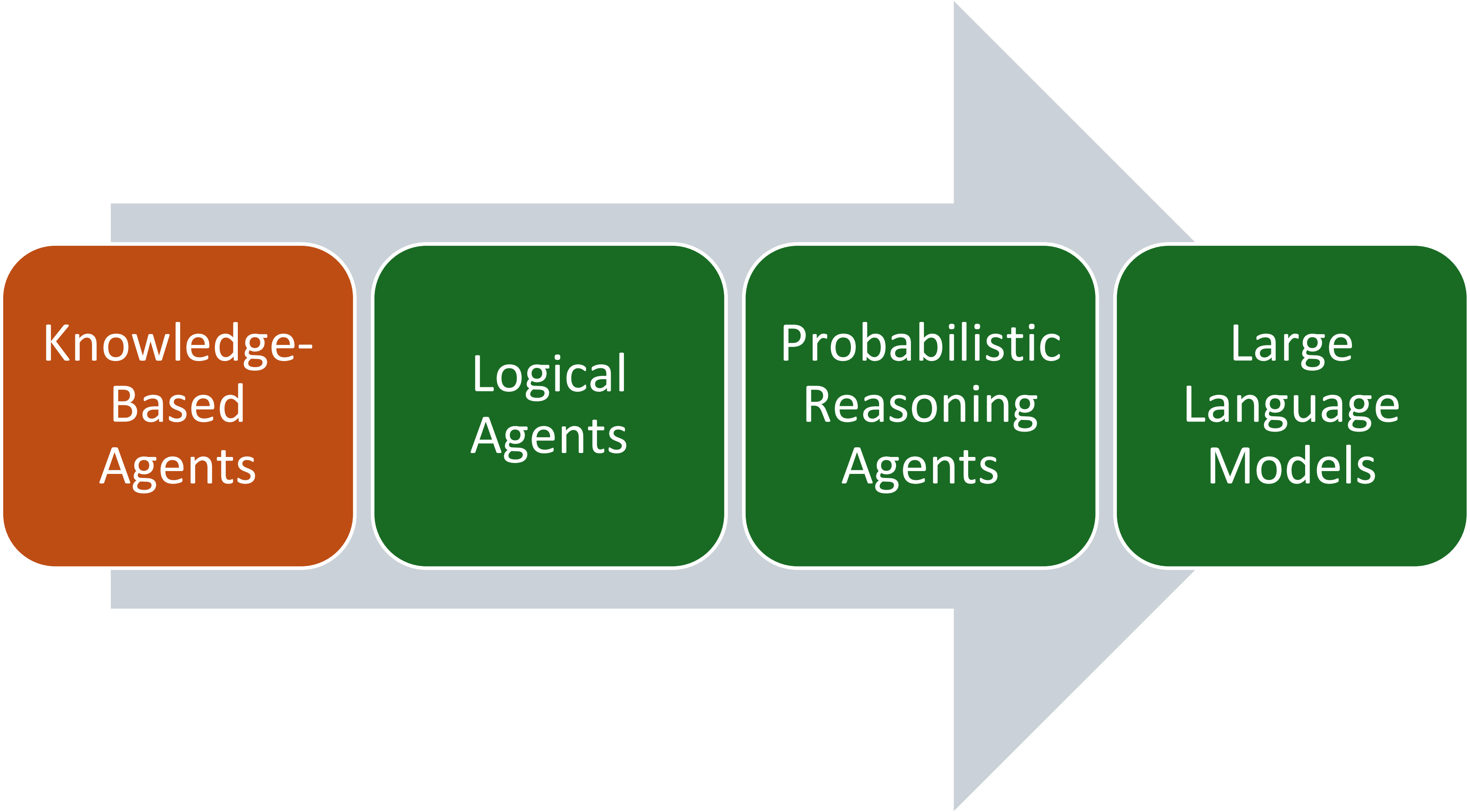
# Knowledge-Based Agents

# Outline



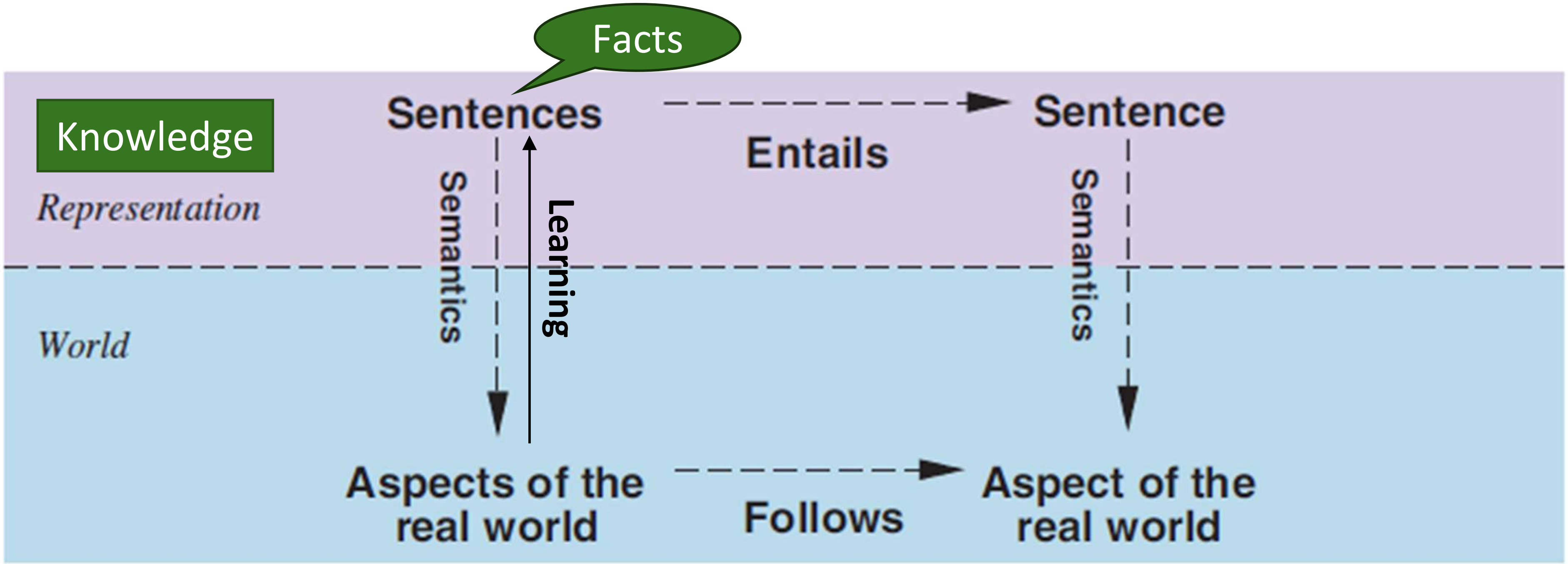
Knowledge-Based Agents

Logical Agents

Probabilistic Reasoning Agents

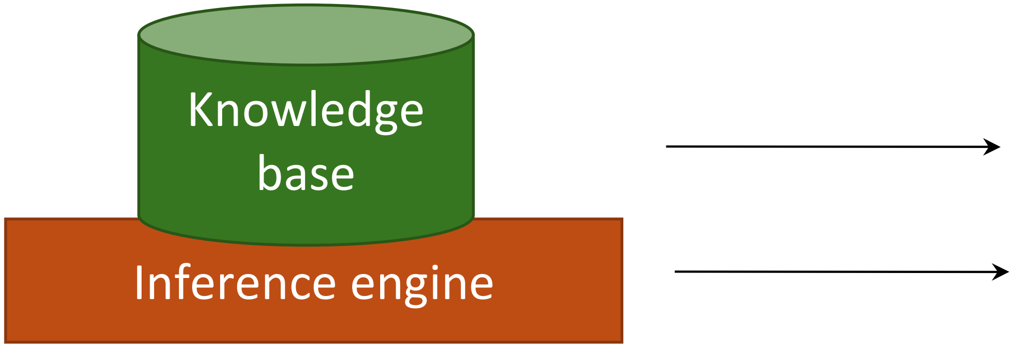
Large Language Models

# Reality vs. Knowledge Representation



- **Facts:** Sentences we know to be true.
- **Possible worlds**: all worlds/models which are consistent with the facts we know (compare with belief state).
- **Learning** new facts reduces the number of possible worlds.
- **Entailment:** A new sentence logically follows from what we already know.

# Knowledge-Based Agents



Knowledge base

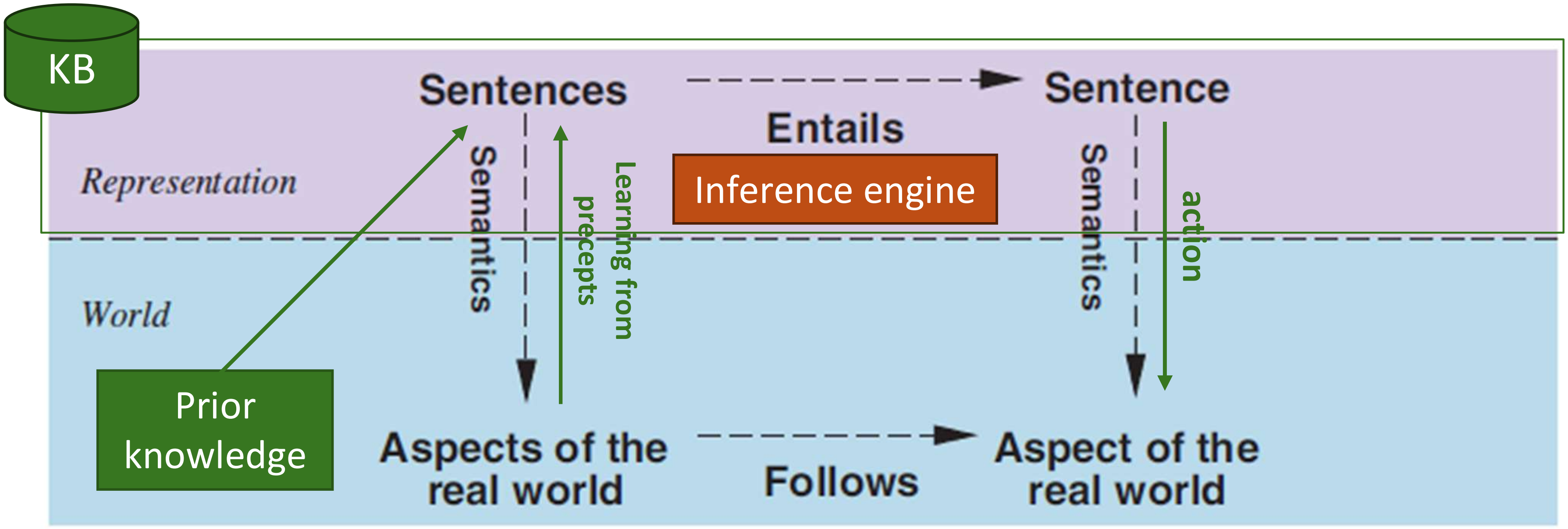
Inference engine

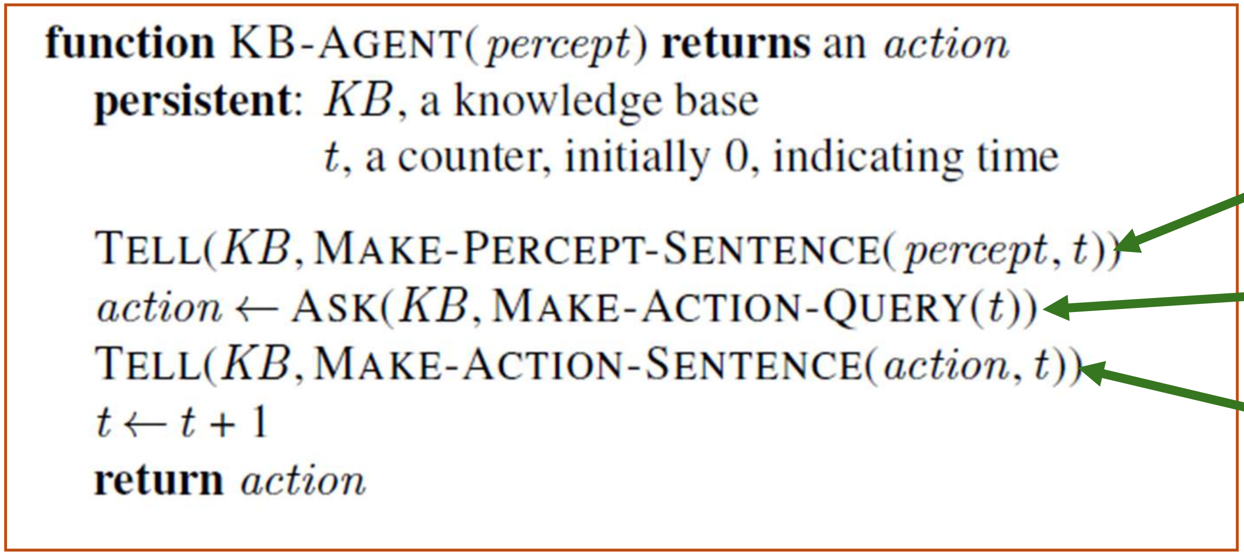
Domain-specific content

Domain-independent algorithms that find new sentences using entailment.

- Knowledge base (KB) = **set of facts.** E.g., set of **sentences** in a **formal language** that are known to be true.
- **Declarative** approach to building an agent: Define what it needs to know in its KB.
- **Separation** between data (knowledge) and program (inference).
- Actions are based on knowledge (sentences + inferred sentences) + an **objective function**. E.g., the agent knows the effects of 5 possible actions and chooses the action with the largest utility.

# Generic Knowledge-based Agent





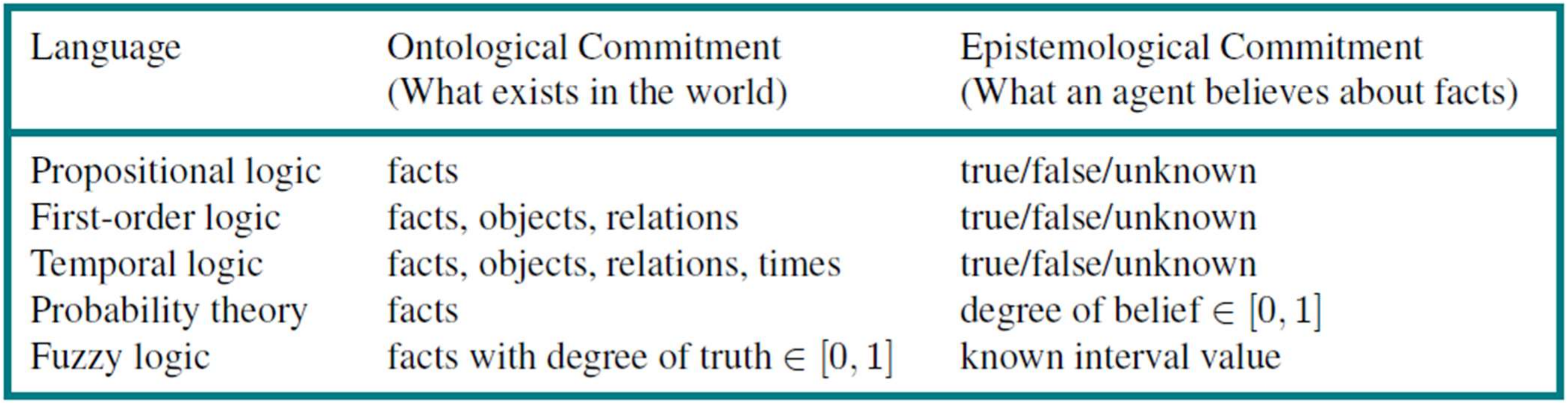
```
function KB-AGENT(percept) returns an action
  persistent: KB, a knowledge base
              t, a counter, initially 0, indicating time

  TELL(KB, MAKE-PERCEPT-SENTENCE(percept, t))
  action ← ASK(KB, MAKE-ACTION-QUERY(t))
  TELL(KB, MAKE-ACTION-SENTENCE(action, t))
  t ← t + 1
  return action
```

Memorize percept at time t

Ask for logical action given an objective

Record action taken at time t

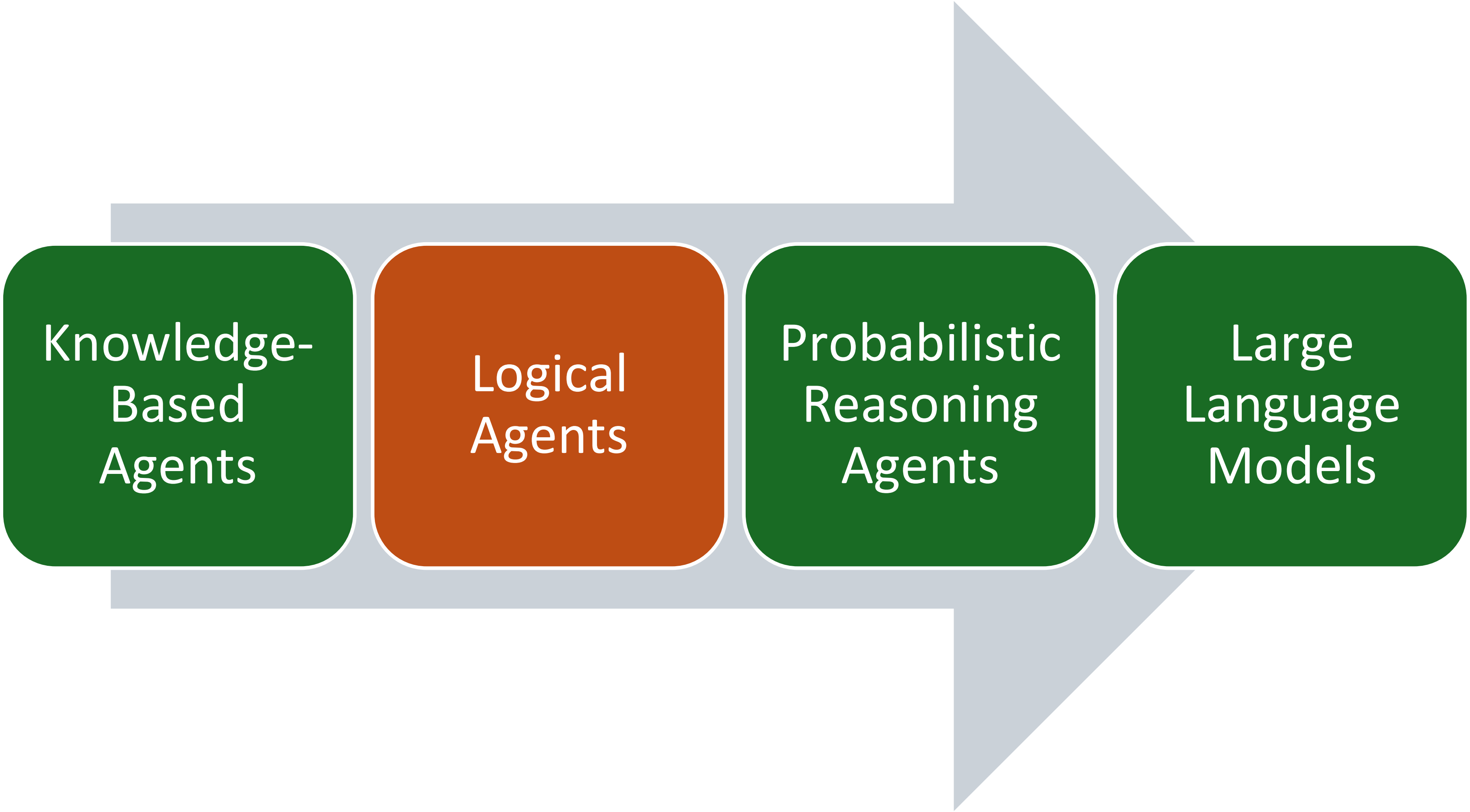
# Different Languages to Represent Knowledge



| Language | Ontological Commitment<br>(What exists in the world) | Epistemological Commitment<br>(What an agent believes about facts) |
|---|---|---|
| Propositional logic | facts | true/false/unknown |
| First-order logic | facts, objects, relations | true/false/unknown |
| Temporal logic | facts, objects, relations, times | true/false/unknown |
| Probability theory | facts | degree of belief $\in [0, 1]$ |
| Fuzzy logic | facts with degree of truth $\in [0, 1]$ | known interval value |

+ Natural Language word patterns representing
facts, objects, relations, … ???

# Outline



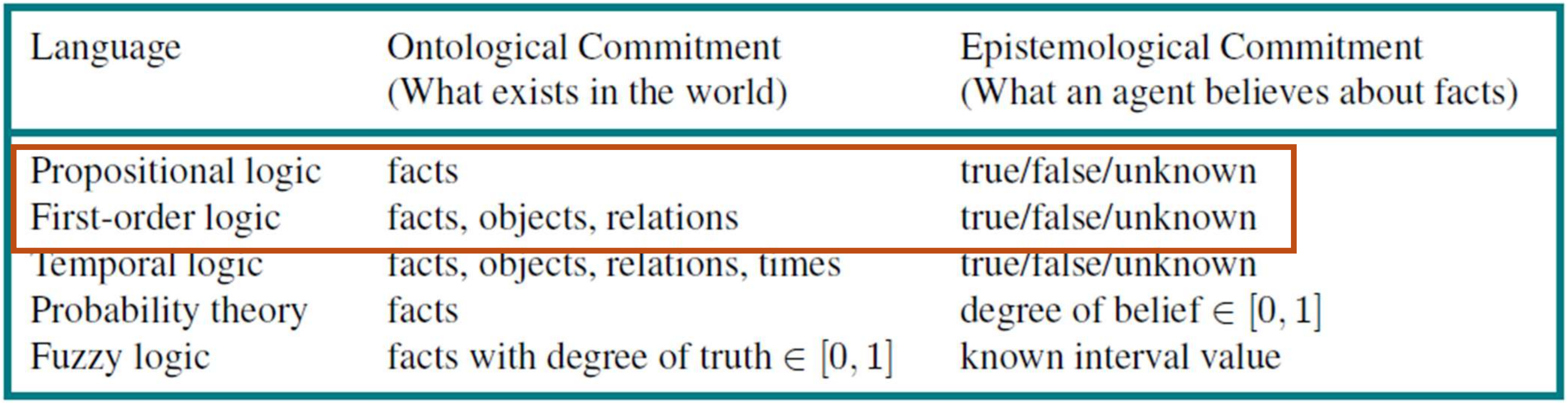
Knowledge-Based Agents

Logical Agents

Probabilistic Reasoning Agents

Large Language Models

# Logical Agents



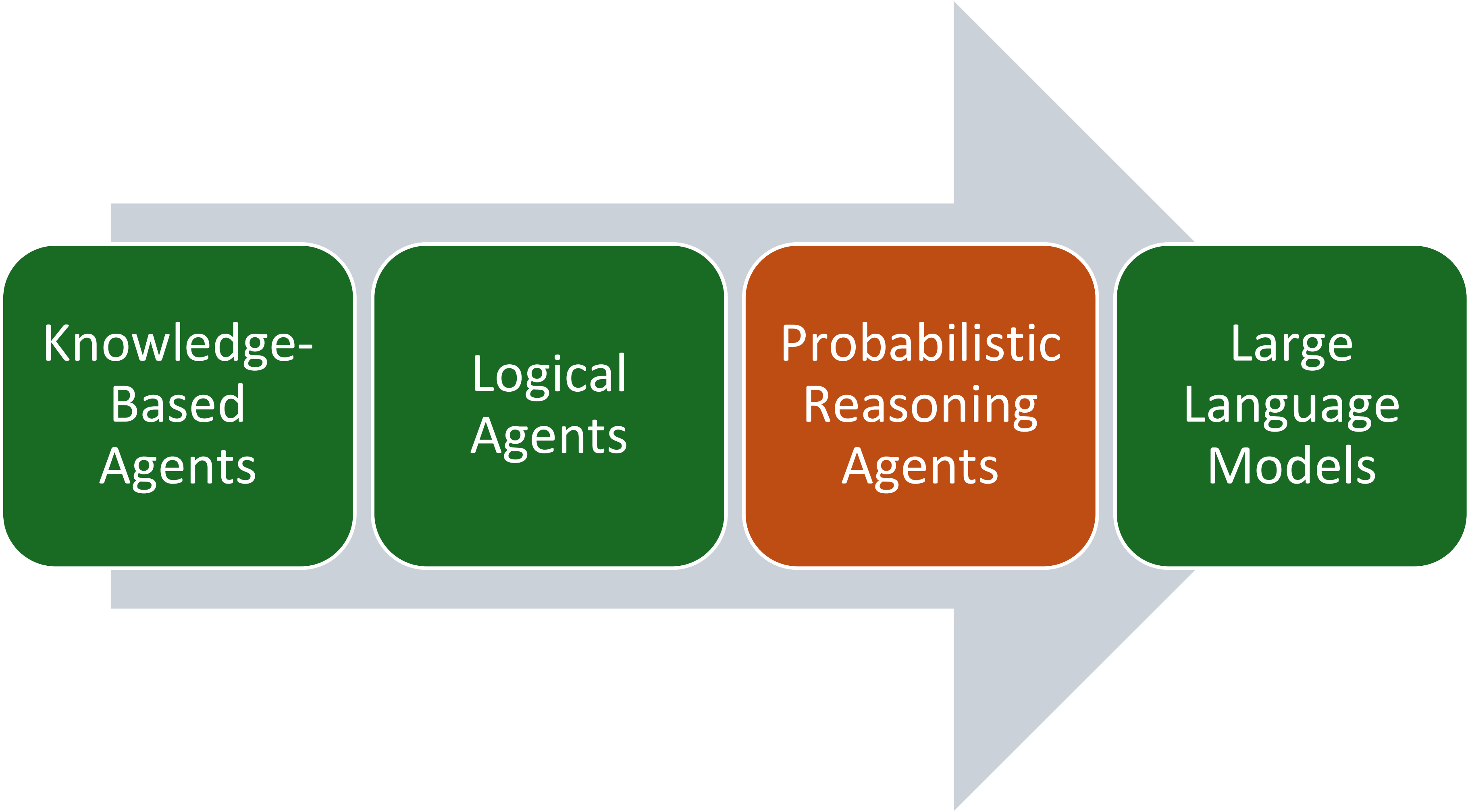
| Language | Ontological Commitment (What exists in the world) | Epistemological Commitment (What an agent believes about facts) |
|---|---|---|
| Propositional logic | facts | true/false/unknown |
| First-order logic | facts, objects, relations | true/false/unknown |
| Temporal logic | facts, objects, relations, times | true/false/unknown |
| Probability theory | facts | degree of belief $\in [0, 1]$ |
| Fuzzy logic | facts with degree of truth $\in [0, 1]$ | known interval value |

- Facts are logical sentences that are known to be true.
- Inference: Generate new sentences that are entailed by all known sentences.
- Implementation: Typically using Prolog
  - Declarative logic programing language.
  - Runs queries over the program (= the knowledge base)

Issues:

- Inference is computationally very expensive.
- Logic cannot deal with uncertainty.

# Outline



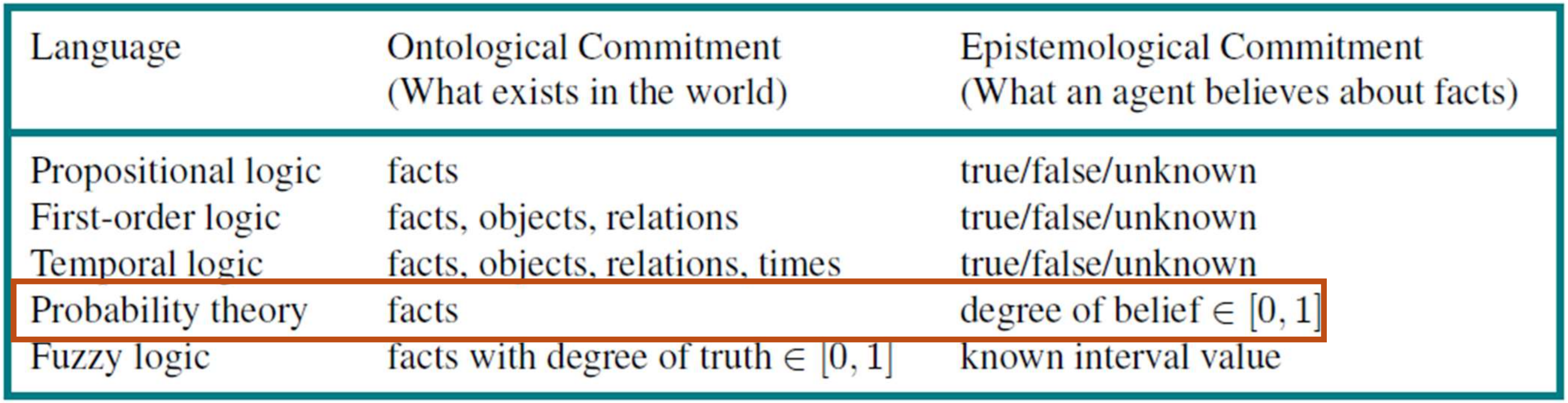
Knowledge-Based Agents

Logical Agents

Probabilistic Reasoning Agents

Large Language Models

# Probabilistic Reasoning

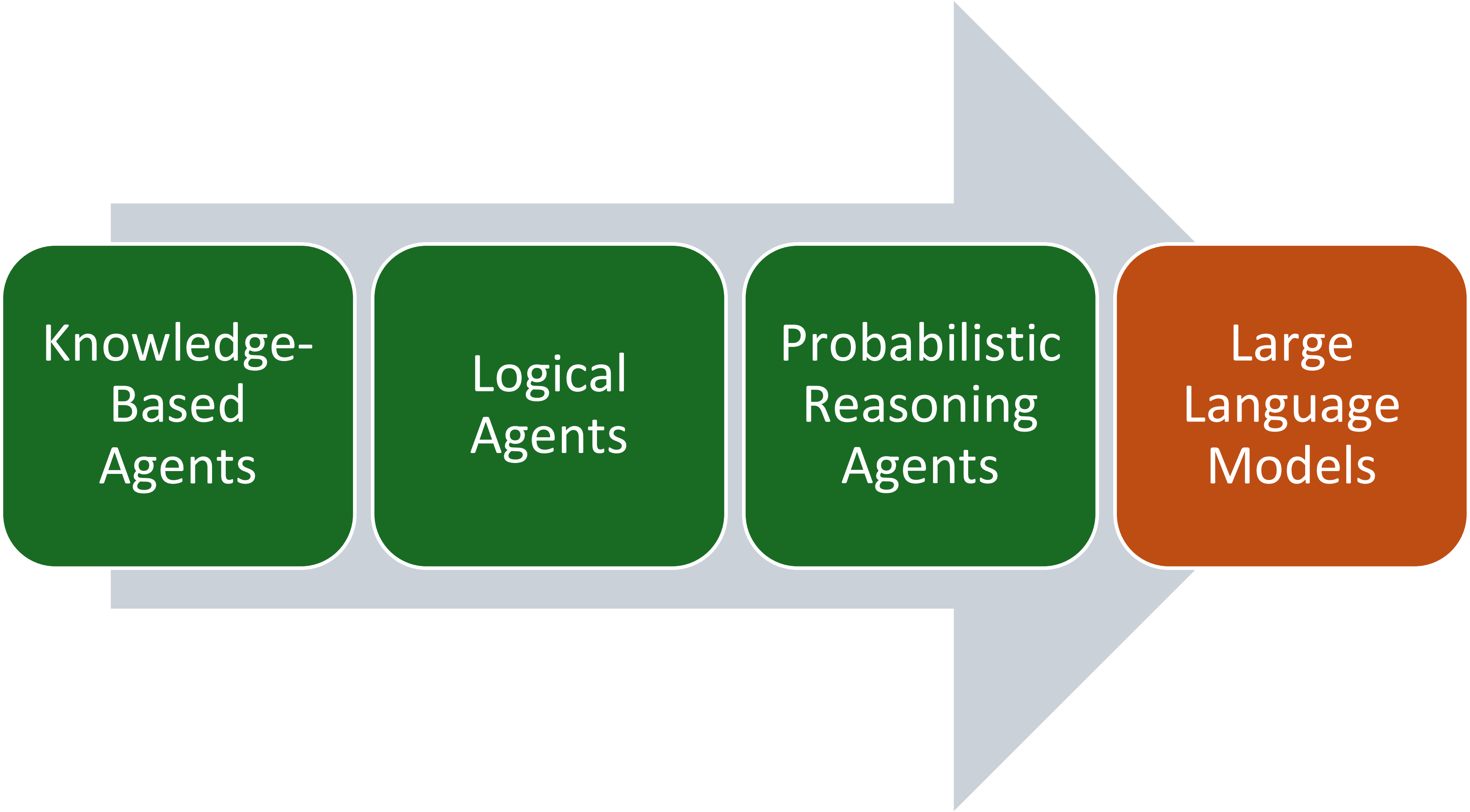


| Language | Ontological Commitment<br>(What exists in the world) | Epistemological Commitment<br>(What an agent believes about facts) |
|---|---|---|
| Propositional logic | facts | true/false/unknown |
| First-order logic | facts, objects, relations | true/false/unknown |
| Temporal logic | facts, objects, relations, times | true/false/unknown |
| Probability theory | facts | degree of belief $\in [0, 1]$ |
| Fuzzy logic | facts with degree of truth $\in [0, 1]$ | known interval value |

- Replaces true/false with a probability.
- This is the basis for
  - Probabilistic reasoning under uncertainty
  - Decision theory
  - Machine Learning

We will talk about these topics a lot

# Outline



Knowledge-Based Agents

Logical Agents

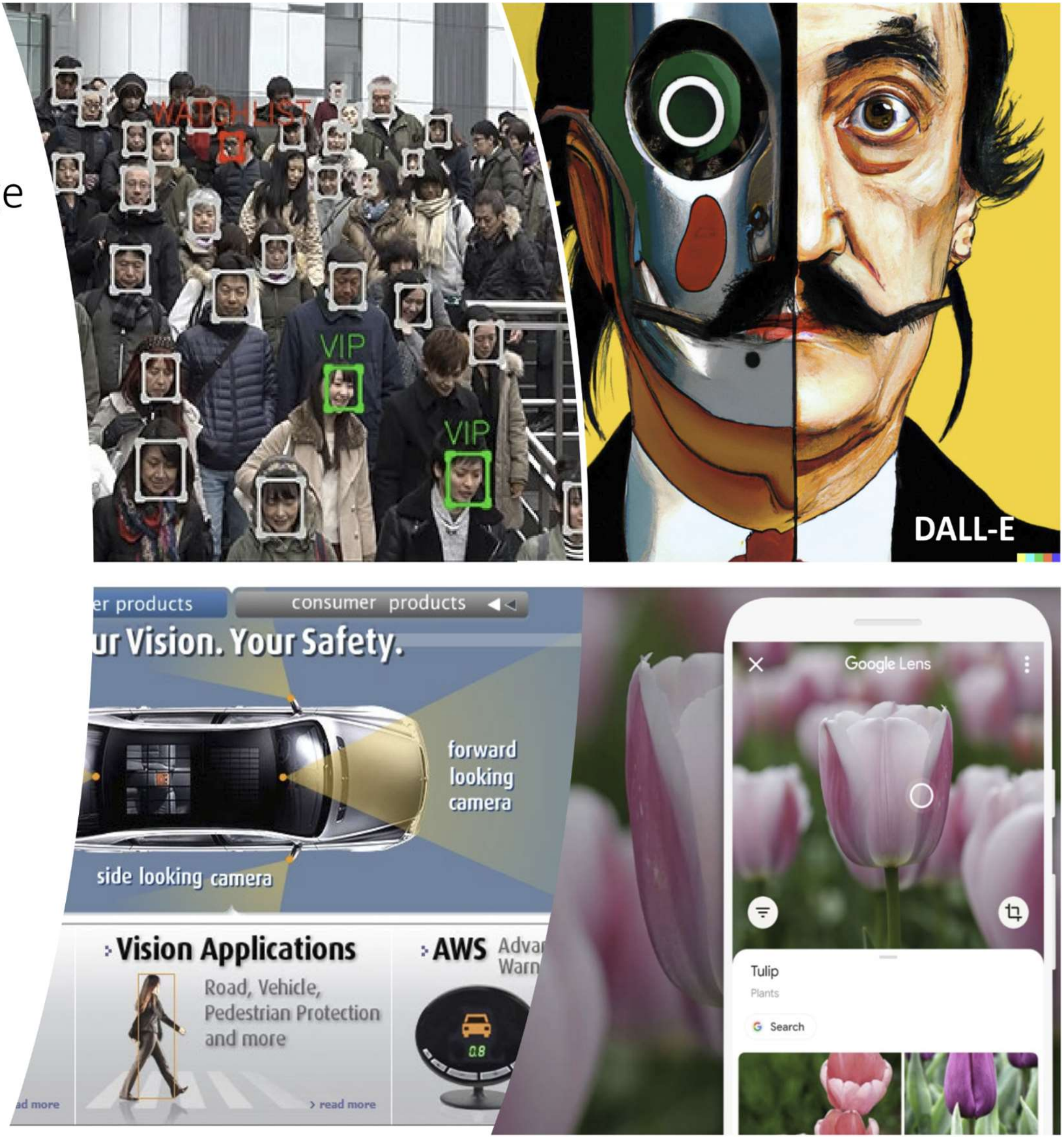
Probabilistic Reasoning Agents

Large Language Models

# Vision and Image Processing

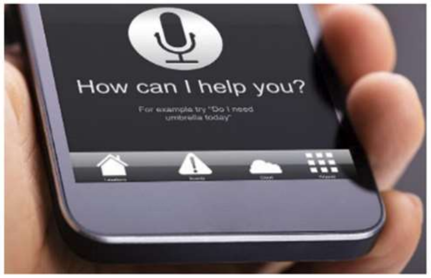
- **OCR**: read license plates, handwriting recognition (e.g., mail sorting).
- **Face detection**: now standard for smart phone cameras.
- **Vehicle safety systems**
- **Visual search**
- **Image generation**

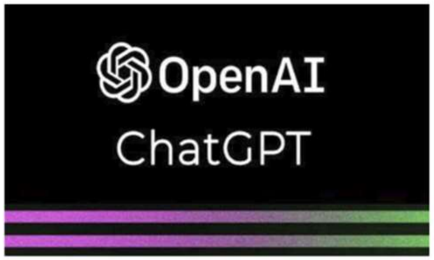
**All these technologies operate now at superhuman performance.**



# Natural Language Processing

- Text-to-speech
- Speech-to-text to detect voice commands
- Machine translation
- Text generation (Q/A systems) using Large Language Models

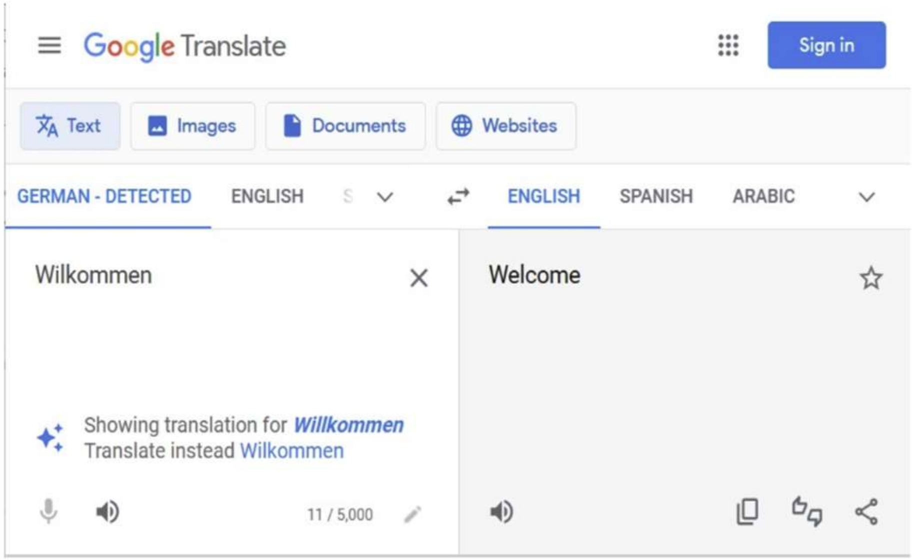




**These technologies operate now with close to or even superhuman performance.**

**Humans use language to reason. Does that mean AI that can create good language can reason?**

**Language understanding is still elusive!**



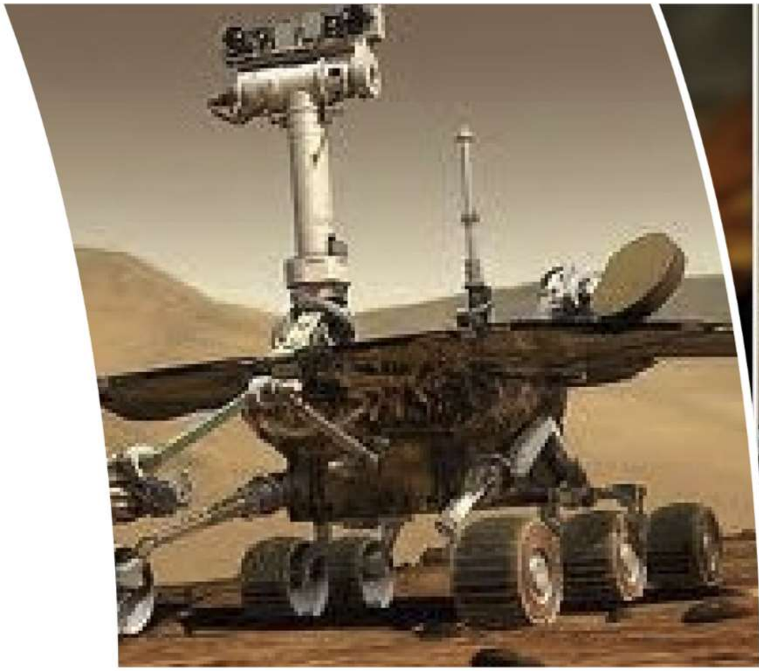
Wilkommen

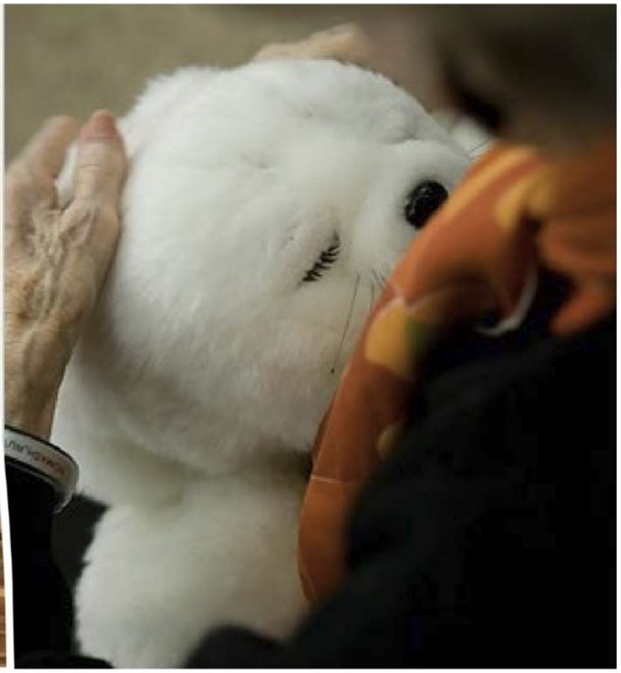
Welcome

Showing translation for *Willkommen*
Translate instead Wilkommen

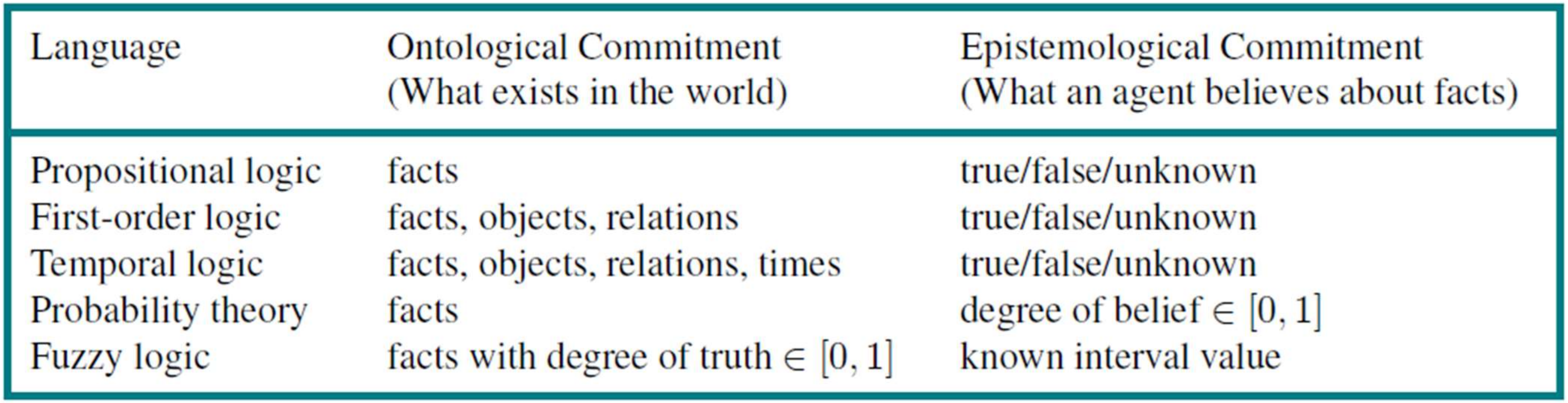
# Robotics

- Mars rovers
- Autonomous vehicles
    - DARPA Grand Challenge
    - Google self-driving cars
- Autonomous helicopters and drones
- Robot soccer
    - RoboCup
- Personal robotics
    - Humanoid robots
    - Robotic pets
    - Personal assistants?





# LLMs - Large Language Models

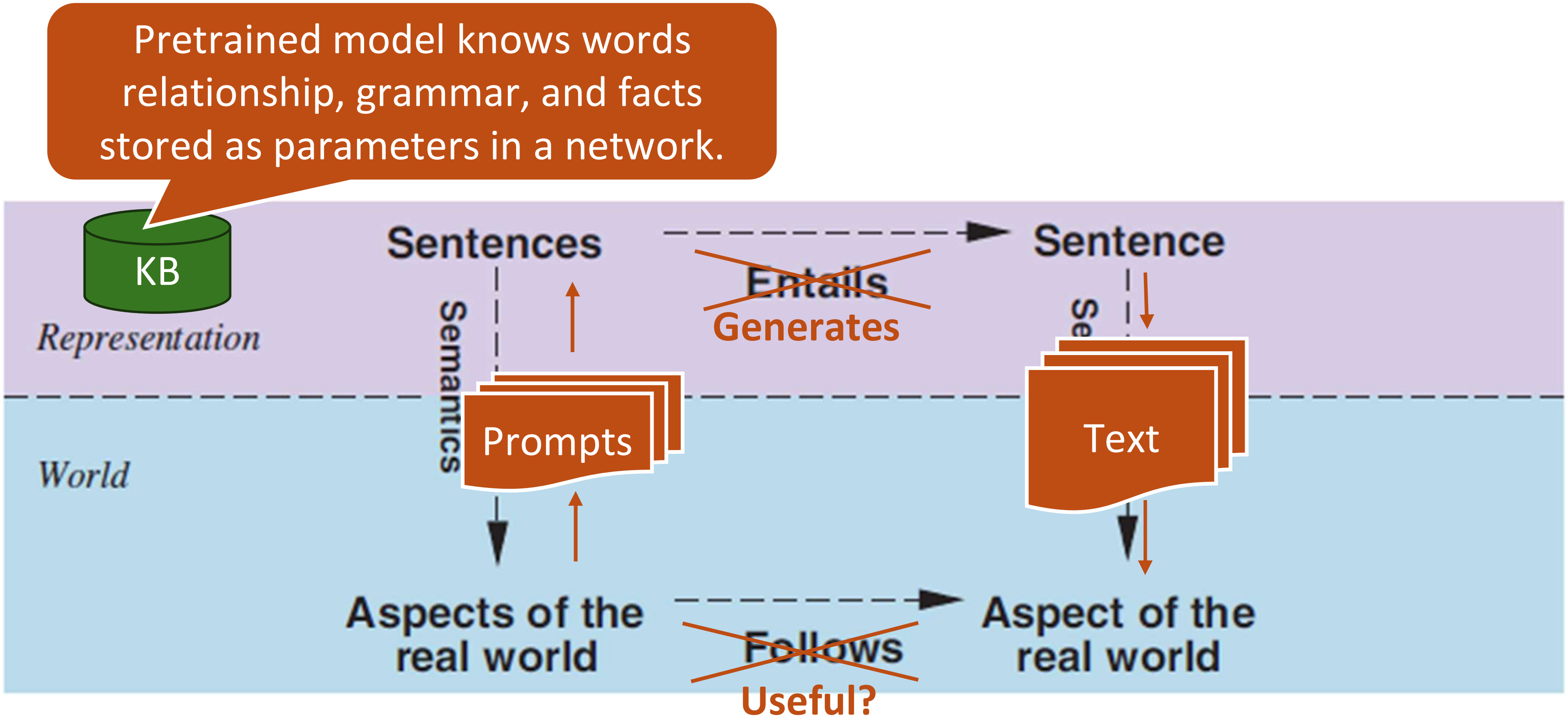


| Language | Ontological Commitment<br>(What exists in the world) | Epistemological Commitment<br>(What an agent believes about facts) |
|---|---|---|
| Propositional logic | facts | true/false/unknown |
| First-order logic | facts, objects, relations | true/false/unknown |
| Temporal logic | facts, objects, relations, times | true/false/unknown |
| Probability theory | facts | degree of belief $\in [0, 1]$ |
| Fuzzy logic | facts with degree of truth $\in [0, 1]$ | known interval value |

| | | |
|---|---|---|
| + Natural Language | word patterns representing<br>facts, objects, relations, … | ??? |

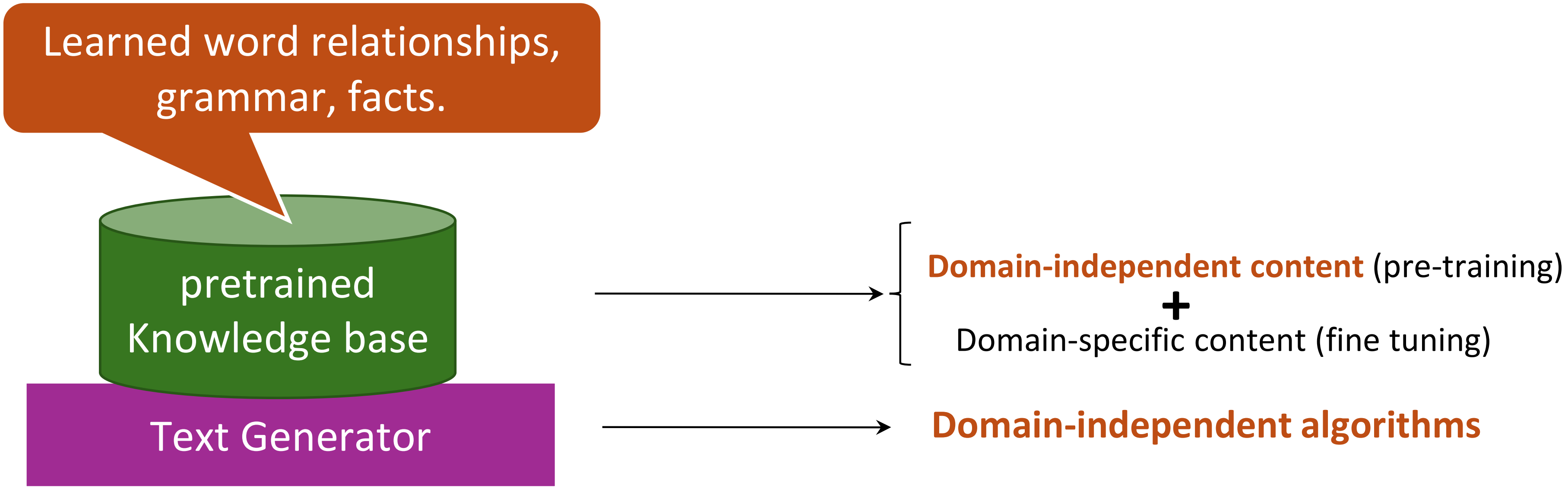
- Store knowledge as parameters in a deep neural networks.

# Using Natural Language for Knowledge Representation



- The user formulates a question about the real world as a natural language prompt (a sequence of tokens).
- The LLM generates text using a model representing its knowledge base.
- The text (hopefully) is useful in the real world. The **objective function** is not clear. Maybe it is implied in the prompt?

# LLM as a Knowledge-Based Agents



Learned word relationships, grammar, facts.

pretrained Knowledge base

Text Generator

**Domain-independent content** (pre-training)

**+**

Domain-specific content (fine tuning)

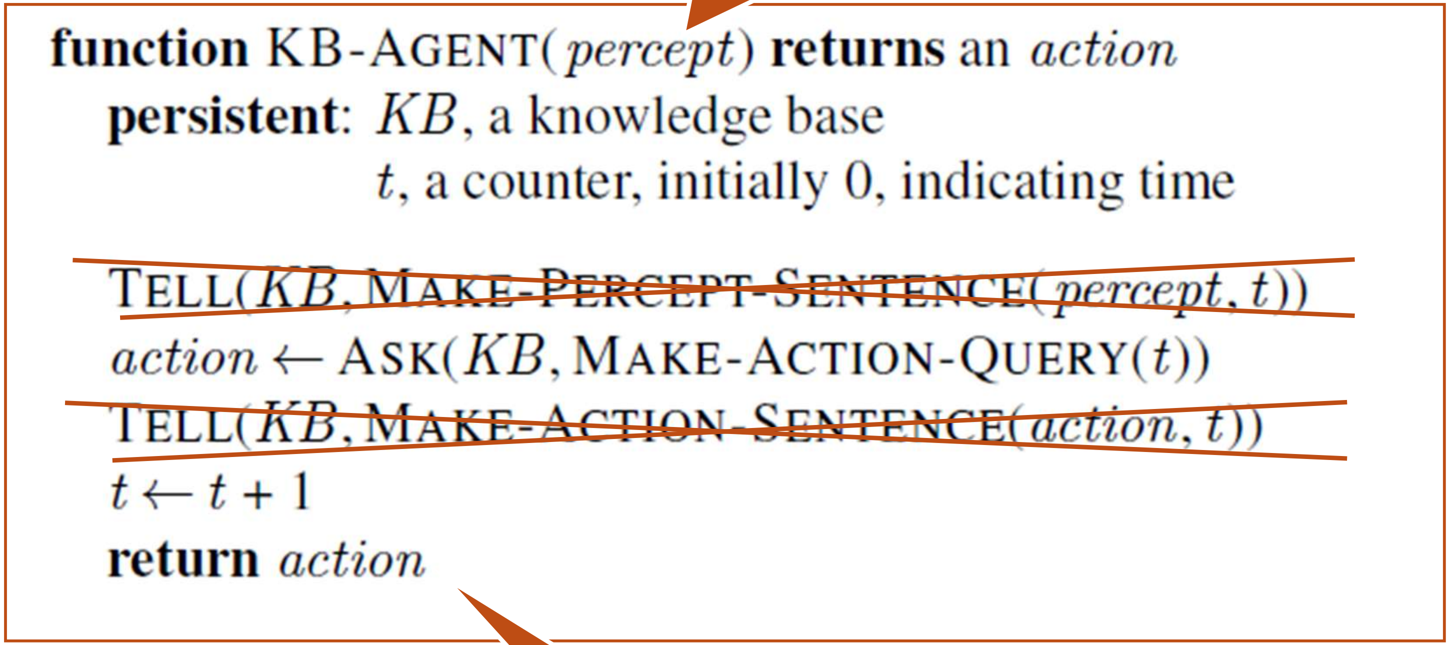
**Domain-independent algorithms**

Current text generators are:

- Pretrained decoder-only transformer models (e.g., GPT stands for Generative Pre-trained Transformer). The knowledge base is not updated during interactions.
- Tokens are created autoregressively. One token is generated at a time based on all the previous tokens using the transformer attention mechanism.

# LLM as a Generic Knowledge-based Agent

Prompt + already generated tokens

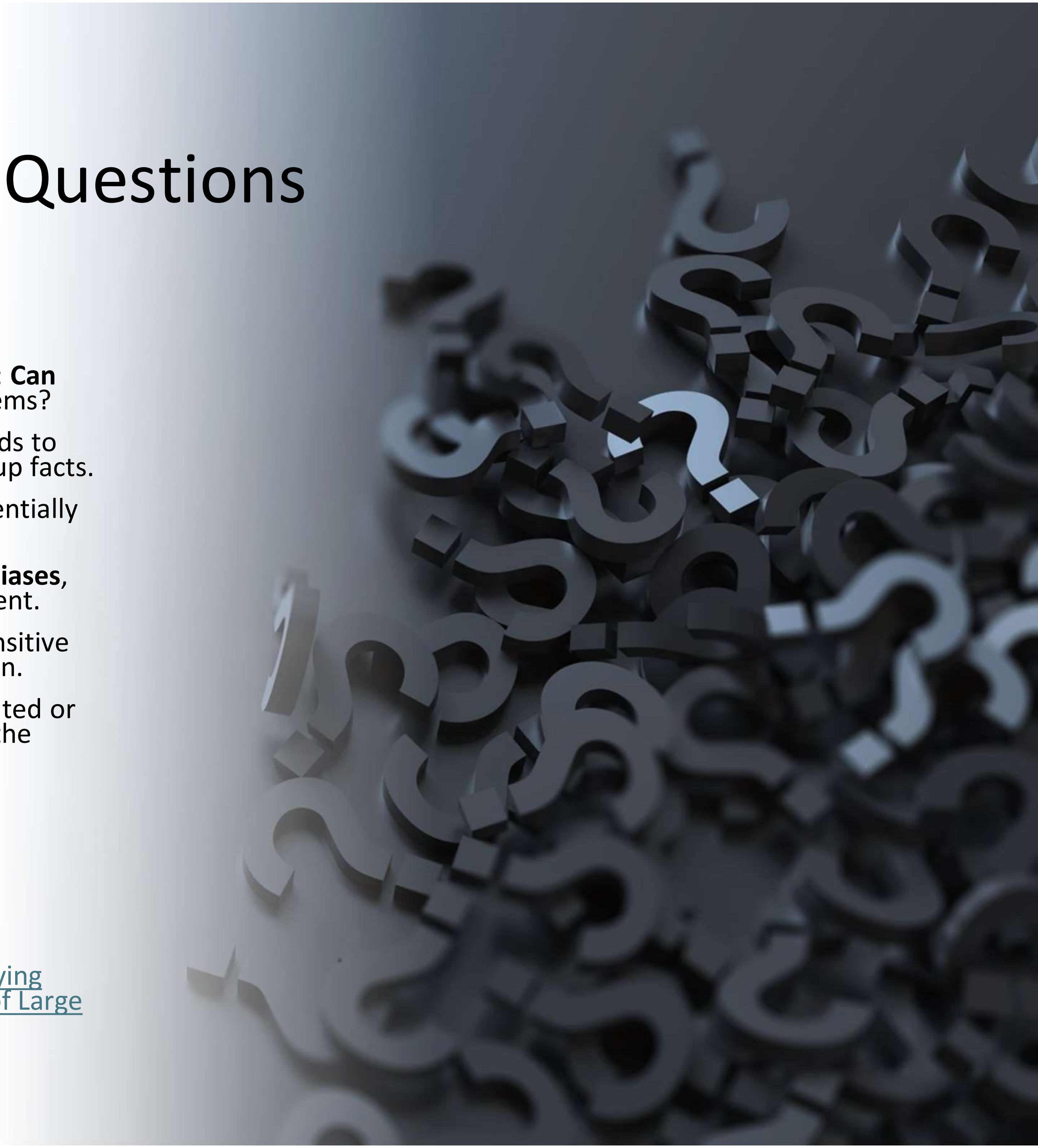


```
function KB-AGENT(percept) returns an action
    persistent: KB, a knowledge base
                $t$, a counter, initially 0, indicating time

    ~~TELL(KB, MAKE-PERCEPT-SENTENCE(percept, $t$))~~
    action ← ASK(KB, MAKE-ACTION-QUERY($t$))
    ~~TELL(KB, MAKE-ACTION-SENTENCE(action, $t$))~~
    $t \leftarrow t + 1$
    return action
```

Next token

- A chatbot repeatedly calls the agent function till the agent function returns the ‘end’ token.

# Many Open Questions about LLMs



- Correlation is not causation: **Can LLMs reason** to solve problems?
- Generative stochasticity leads to **hallucinations**: LLM makes up facts.
- Autoregression is an exponentially **diverging** diffusion process.
- The training data contains **biases**, nonsense and harmful content.
- **Security**: LLM can reveal sensitive information it was trained on.
- **Rights-laundering**: Copyrighted or licensed material can be in the training data.
- Leaky data makes it hard to evaluate true **reasoning performance**.

Reading: [2307.04821] Amplifying Limitations, Harms and Risks of Large Language Models (arxiv.org)

# Conclusion

- The **clear separation between knowledge and inference engine** is very useful.
- **Pure logic** is often not flexible enough. The fullest realization of knowledge-based agents using logic was in the field of expert systems or knowledge-based systems in the 1970s and 1980s.
- **Pretrained Large Language Models** are an interesting new application of knowledge-based agents based on natural language.
- Next, we will talk about **probability theory** which is the standard language to reason under uncertainty and forms the basis of machine learning.