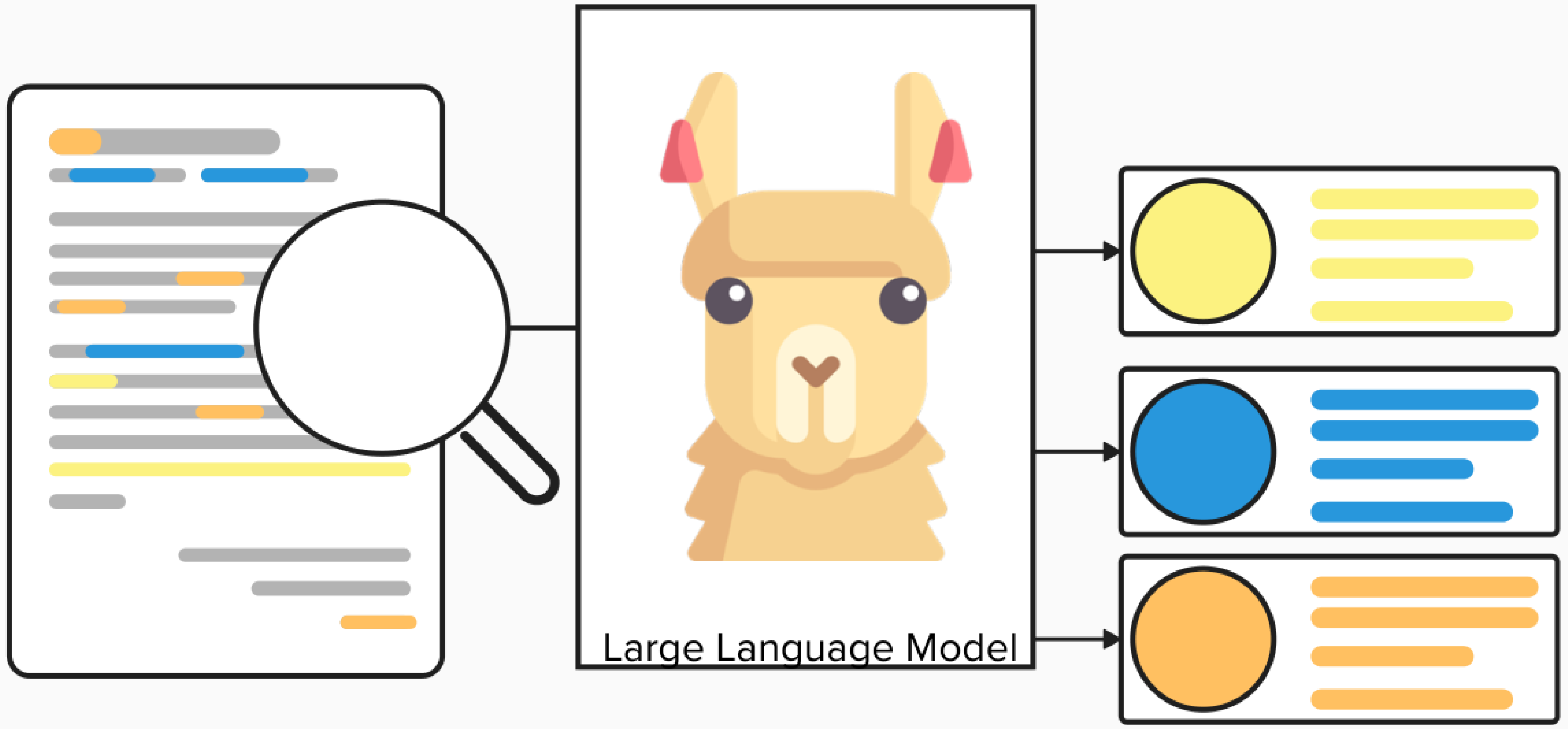
# Generative Large Language Models



## Lesson Overview

**Key Topics Covered:**

1. **LLM Theory**
   - From Discriminative to Generative Models
     - Understanding the shift from BERT to GPT-style architectures
     - Key differences in purpose and training objectives
   - Model Parameters Deep Dive
     - Understanding model size and resource requirements
     - Generation parameters (temperature, top-k, top-p)

2. **LLM Ecosystem**
   - Closed Models vs Open Models
   - Comparison of proprietary and open-source options

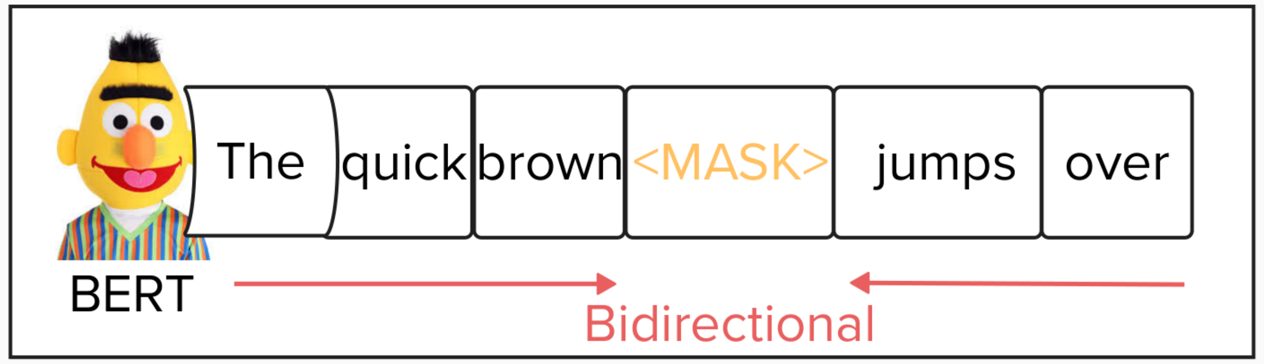
3. **Context and Prompting**
   - Understanding Context Windows
   - Context window limitations and implications
   - In-Context Learning

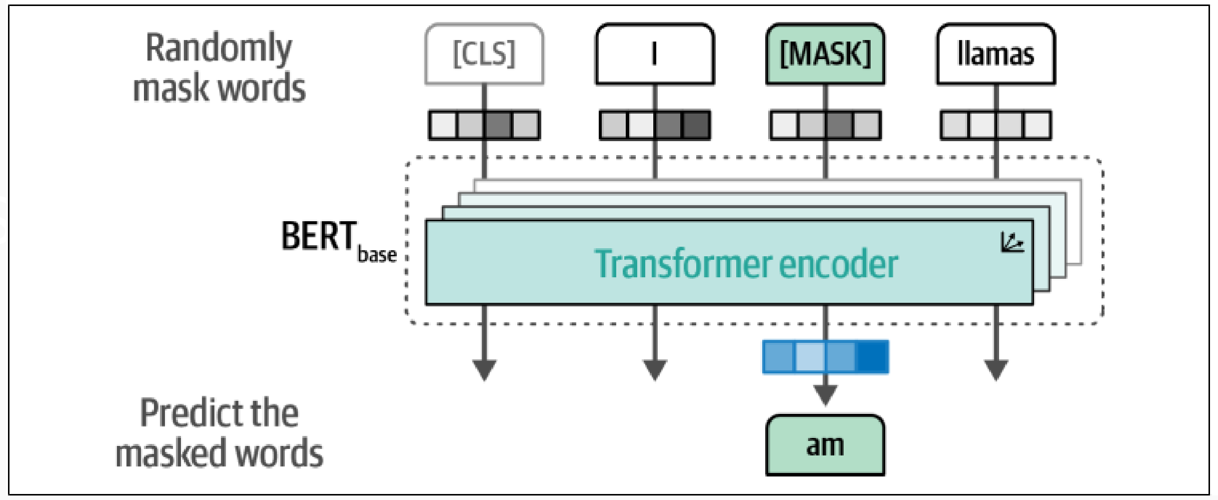
4. **Practical Workshop**
   - Prompting with  LLMs
   - In-Context Learning
   - Sentiment Analysis and Comparisons
   - Using Tools
   - And more

# 1. From Discriminative to Generative Models

**Discrimitave Models (e.g., BERT)**

- **Purpose:** Classify or predict specific labels
- **Training Objective:** Masked world prediction
- **Context Direction**: Bidirectional
- Can’t generate new text

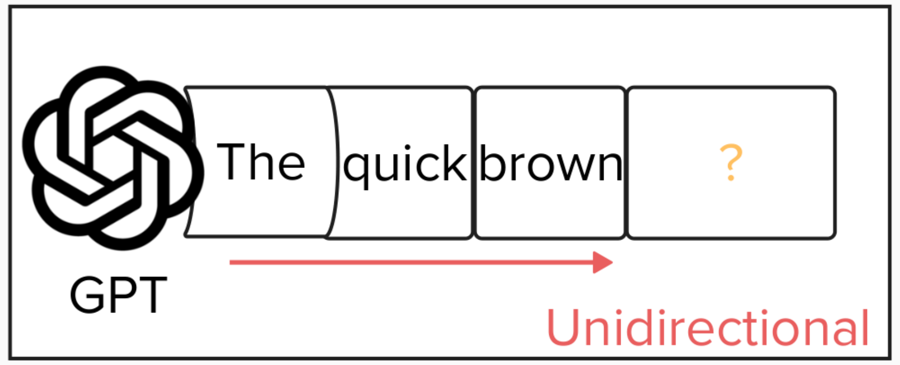


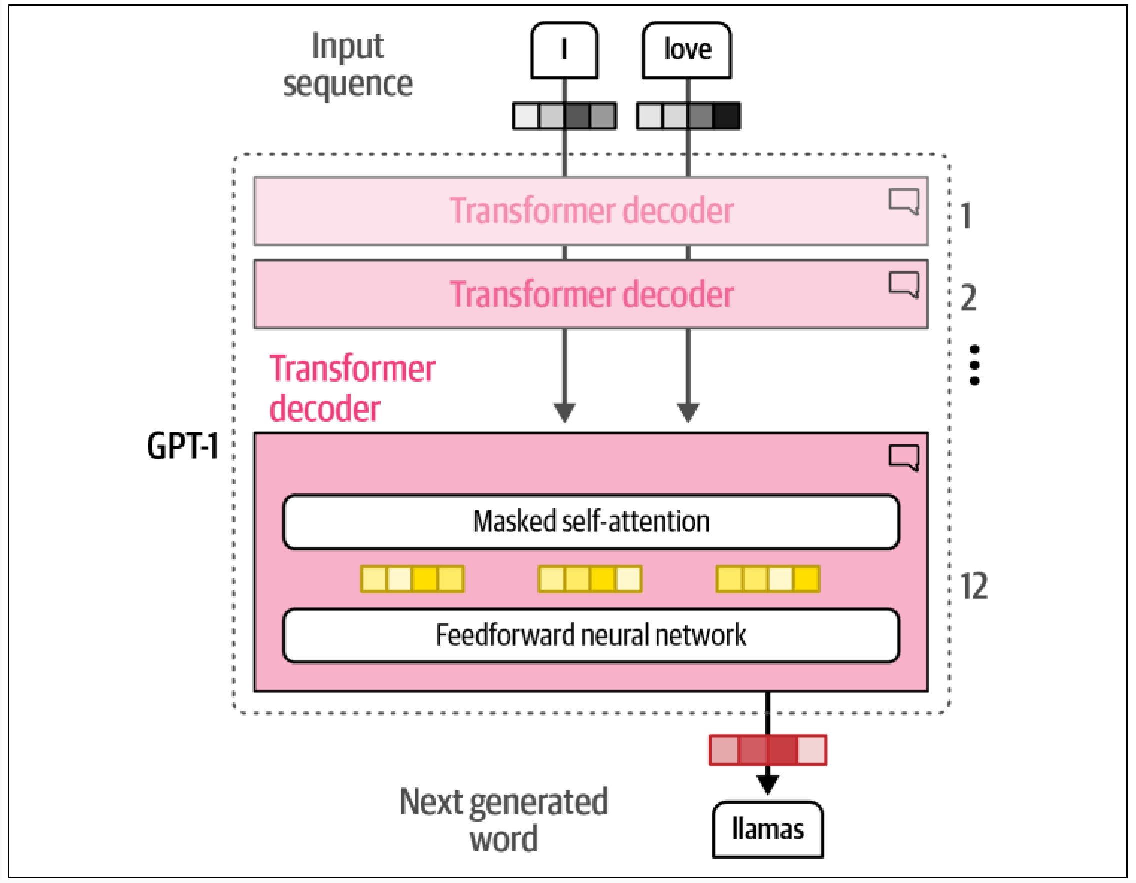



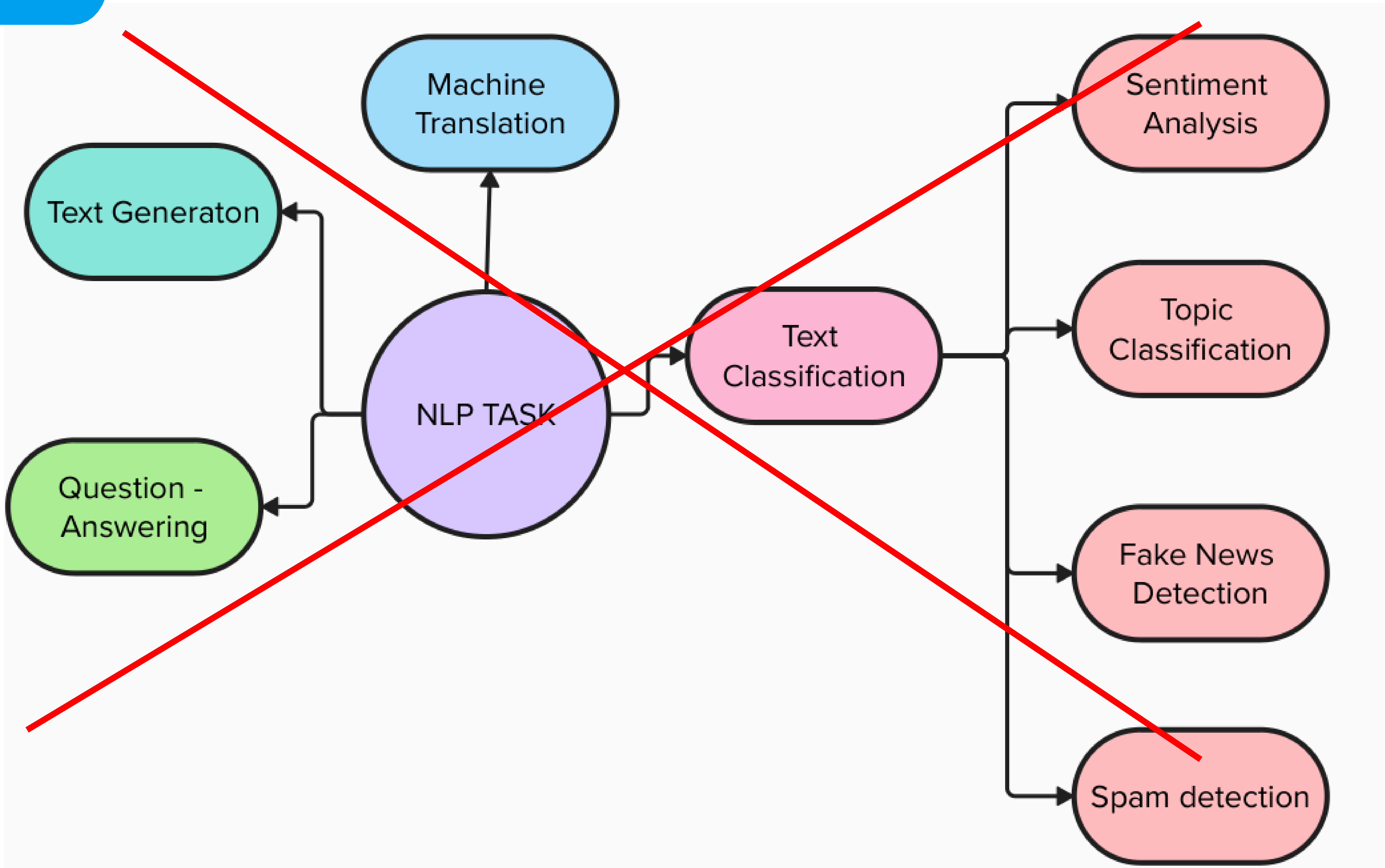

*From Hands-On Large Language Models: Language Understanding and Generation.*

**Generative Models (e.g., GPT)**

- **Purpose:** Generate new content
- **Training Objective:** Next world Prediction
- **Context Direction**: Undirectional - From left to right



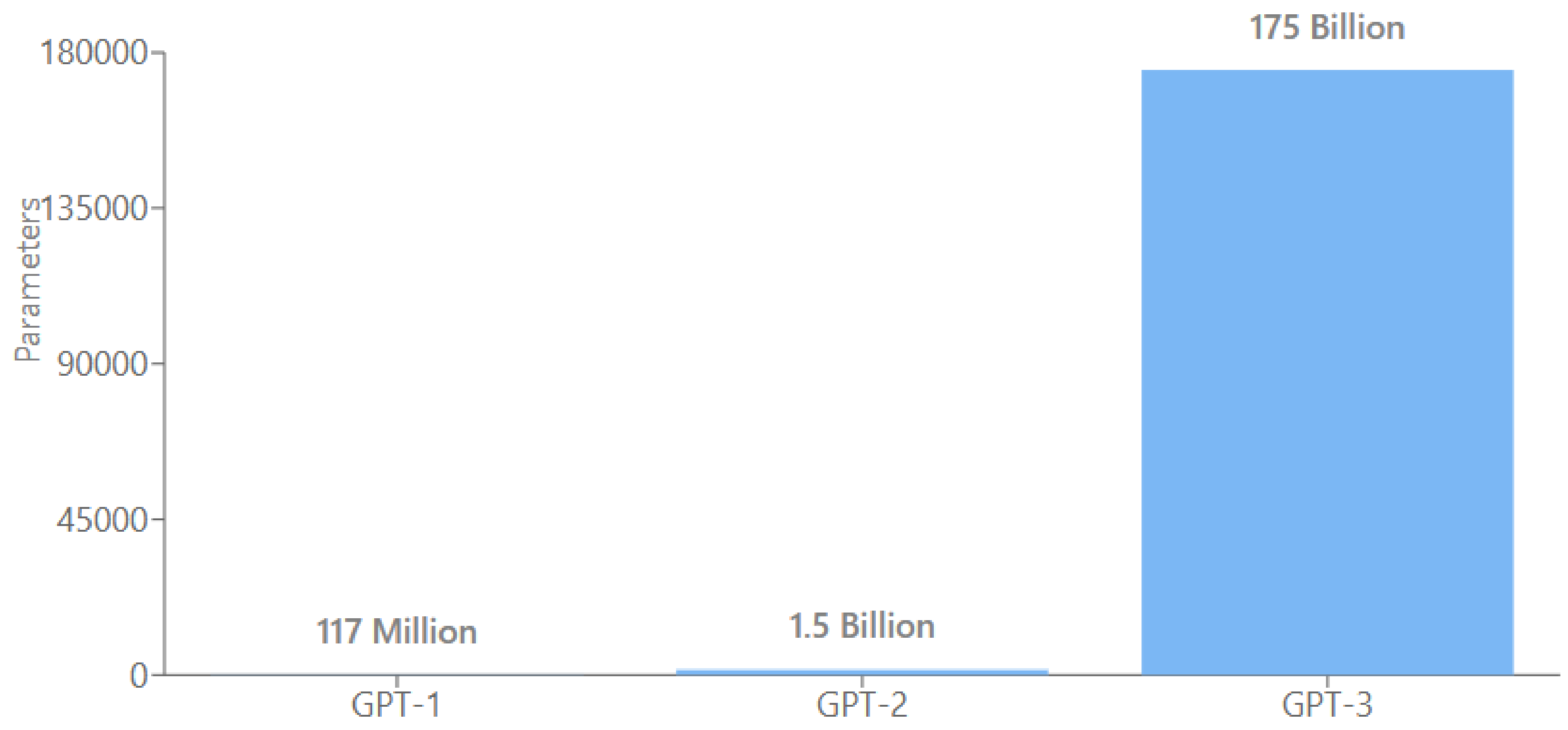



With generative large language models this is no longer valid.

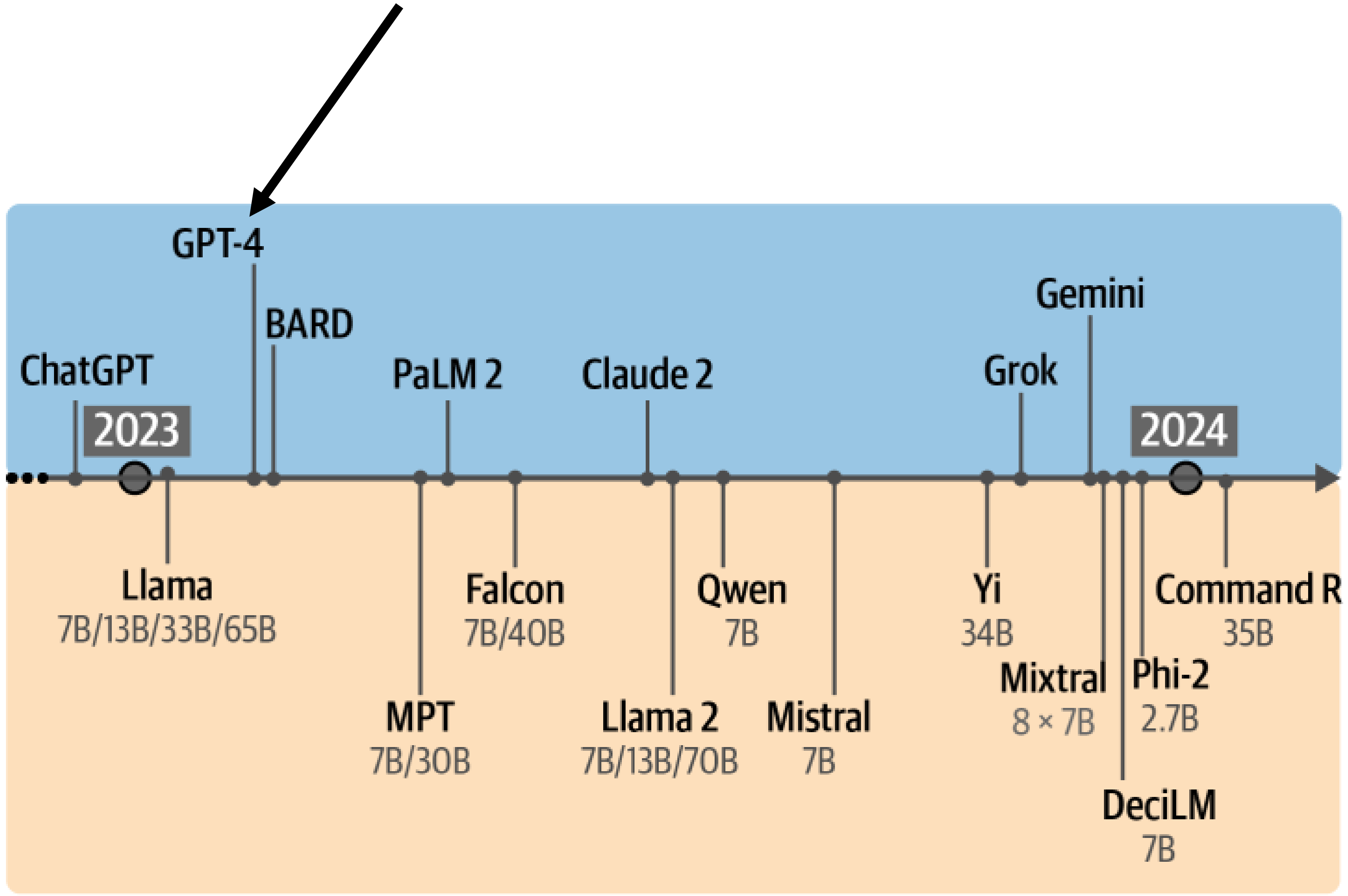
# 1.2 Understanding Parameters: Model parameters

**Definition:** The weights learned during the training
**Examples:** GPT-3 reached 175B parameters
**Memory Requirements (estimation):**

1B parameters ≈ 4GB
70B parameters ≈ 280GB in
....



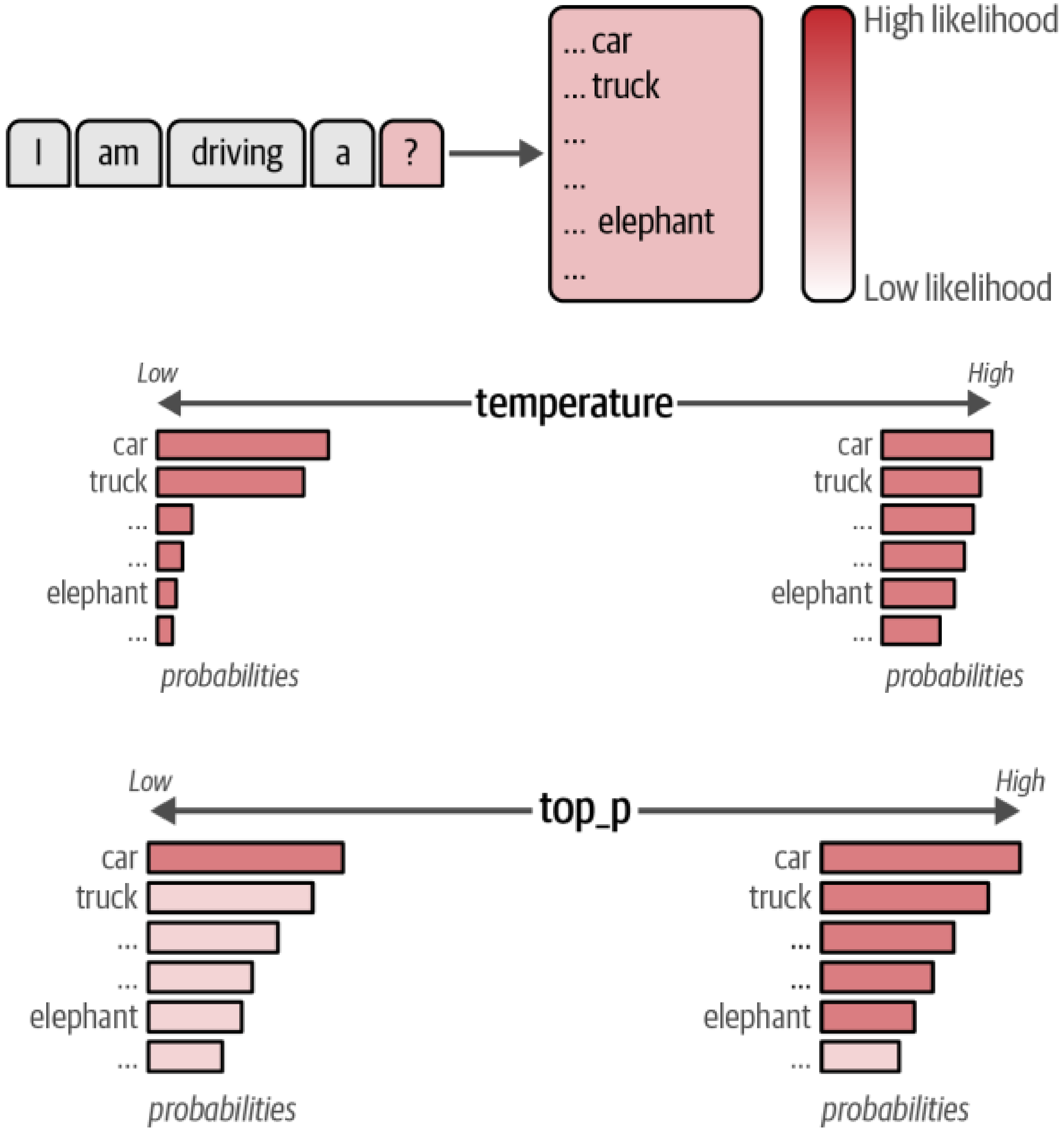

Training with 25,000 A100 GPUs over a period of 90 to 100 days



## Controlling LLM Output

| Parameter | Range | Description | Effects |
|---|---|---|---|
| **Temperature** | 0.0 - 1.0 | Controls randomness in token selection | Low: Deterministic, repetitive<br>High: More diverse, less reliable |
| **Top-k** | 1 - 100 | Limits selection to k most likely tokens | Low: Conservative outputs<br>High: More diverse |
| **Top-p** | 0.0 - 1.0 | Cumulative probability threshold for token selection | Low: Stable, conservative<br>High: More unexpected tokens |
| **Max Tokens** | Model dependent | Maximum length of generated response | Too low: Truncated output<br>Too high: Wasted compute |
| **Presence Penalty** | -2.0 to 2.0 | Penalizes tokens already used | Low: May repeat themes High: More diverse but might lose focus |
| **Frequency Penalty** | -2.0 to 2.0 | Penalizes tokens based on frequency | Low: Natural language<br>High: More unique vocabulary |

## Controlling LLM Output



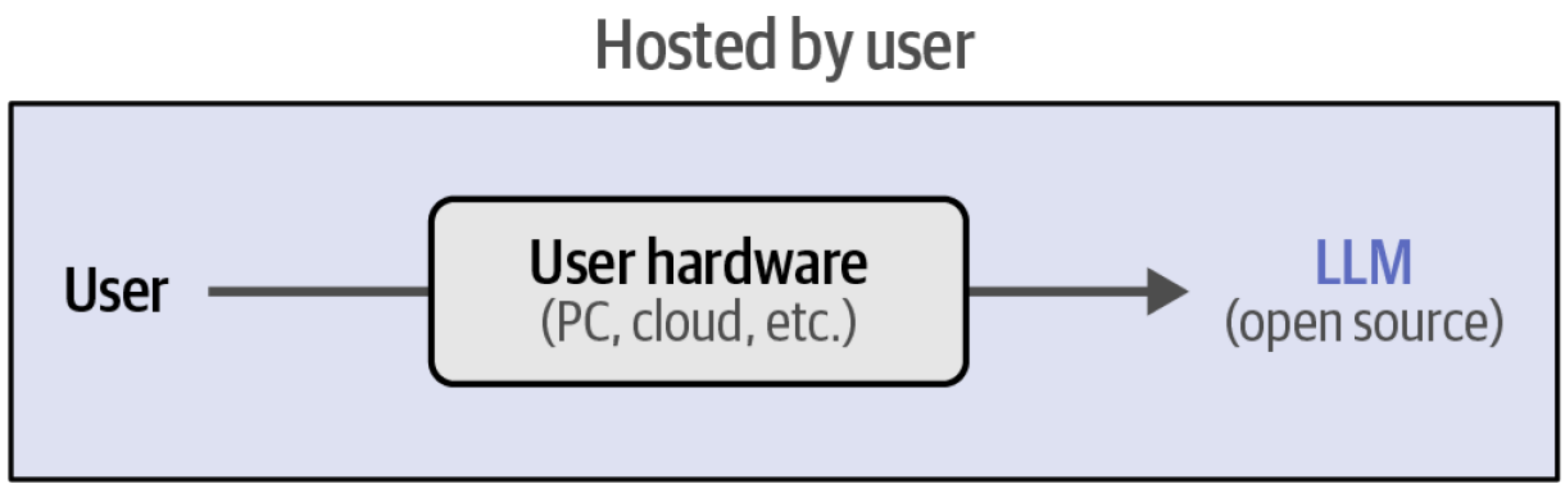

| Example Use Case | Temperature | Top_p | Description |
|---|---|---|---|
| Poetry Writing | High | High | Maximum creativity for generating unique verses and metaphors. Produces diverse poetic expressions with unexpected word combinations and original imagery. |
| API Documentation | Low | Low | Highly precise and consistent technical writing. Generates standardized documentation with exact terminology and predictable formatting. |
| Blog Writing | High | Low | Creative content generation with controlled vocabulary. Creates engaging articles while maintaining consistent tone and subject focus. |
| Language Translation | Low | High | Accurate translation with flexible word choice. Maintains original meaning while exploring different ways to express concepts in the target language. |

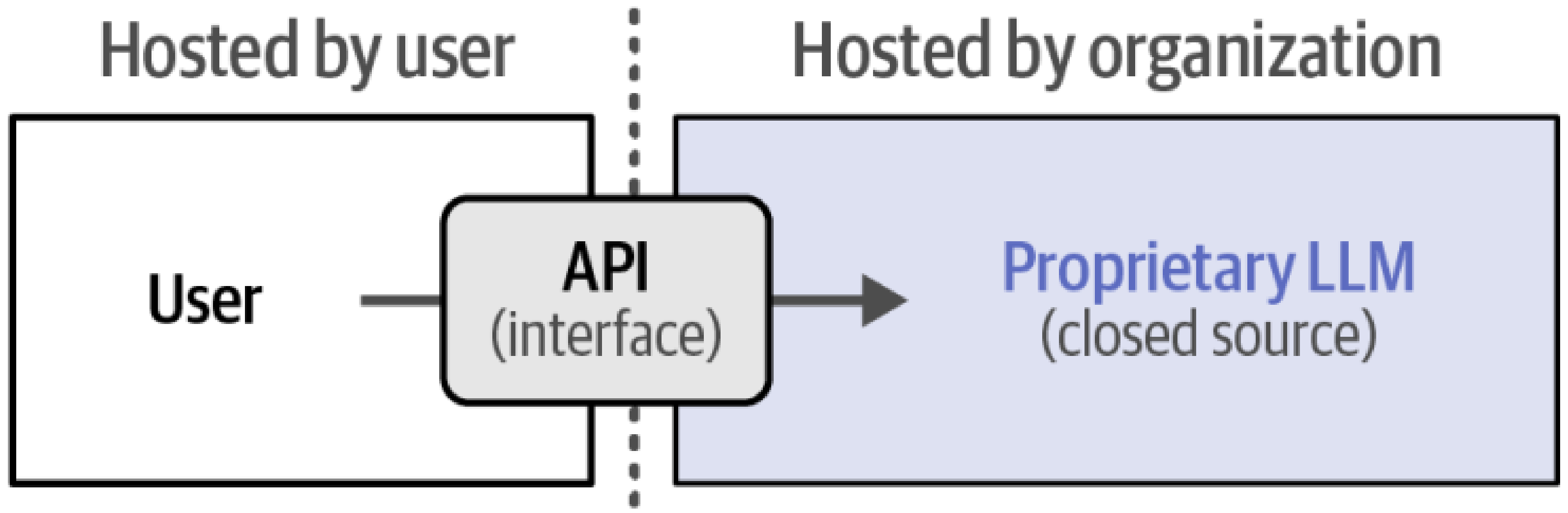
**Open Models**

- **Model weights publicly downloadable**
- **Architecture fully documented**
- **Clear licensing terms**
- **User can host it with his own hardware**
- **Companies made the inferencing available through APIs**
  - Llama
  - Mistral
  - Gemma
  - Etc
- **Key Aspects:**
  - **Can be fine-tuned**
  - **Local deployment possible**
  - **Full control**

**Closed Models (Proprietary)**

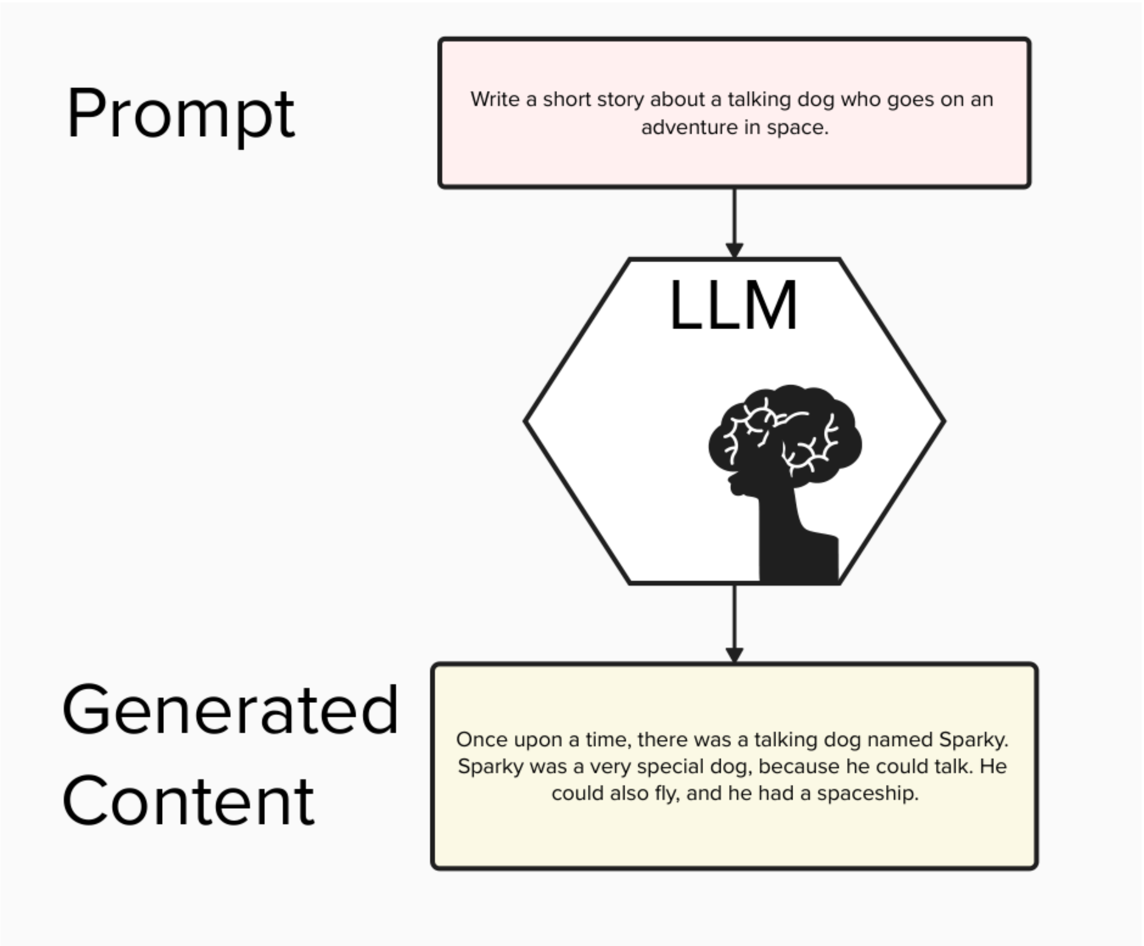
- **Model weights & architecture not accessible**
- **Available only through APIs**
- **Training data & methods are not public**
- **Examples:**
  - GPT
  - Claude
  - Gemini
  - Etc
- **Key Aspects:**
  - User-based pricing
  - No direct model control
  - Provider handls security and updates

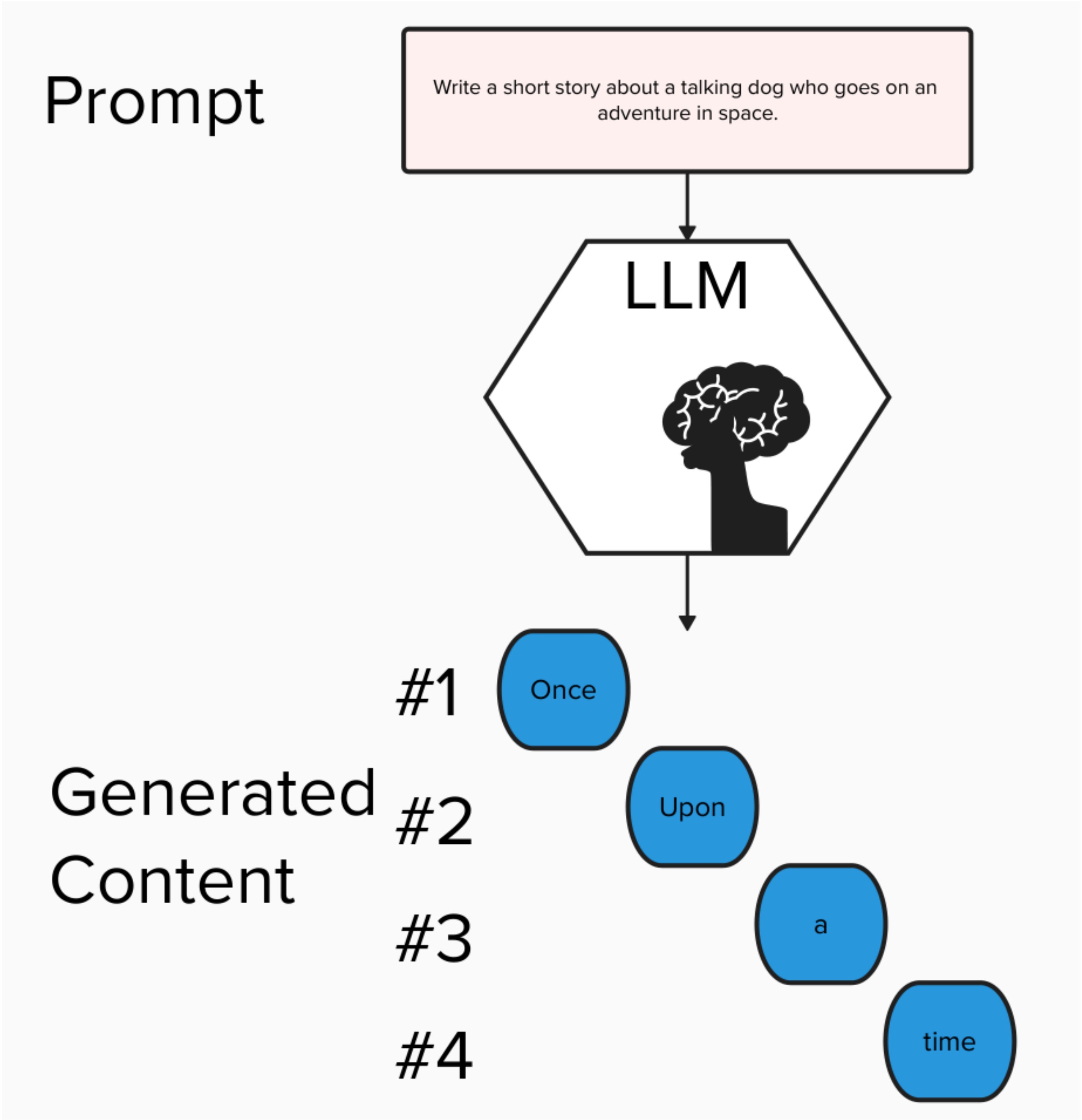


## How does the model generate?

## How does the model generate?



## 3. Context – What is and what is its purpose?


From TRECCANI.it

**Linguistic Context**

- textual elements surrounding a word or phrase within discourse.
- The meaning of a word can change based on the sentences that precede or follow it

**Extralinguistic Context**

- This includes the physical, temporal, social, and cultural circumstances in which communication occurs
- Factors such as location, time, relationships between participants, and cultural norms play a crucial role in interpreting messages

### What about LLMs?

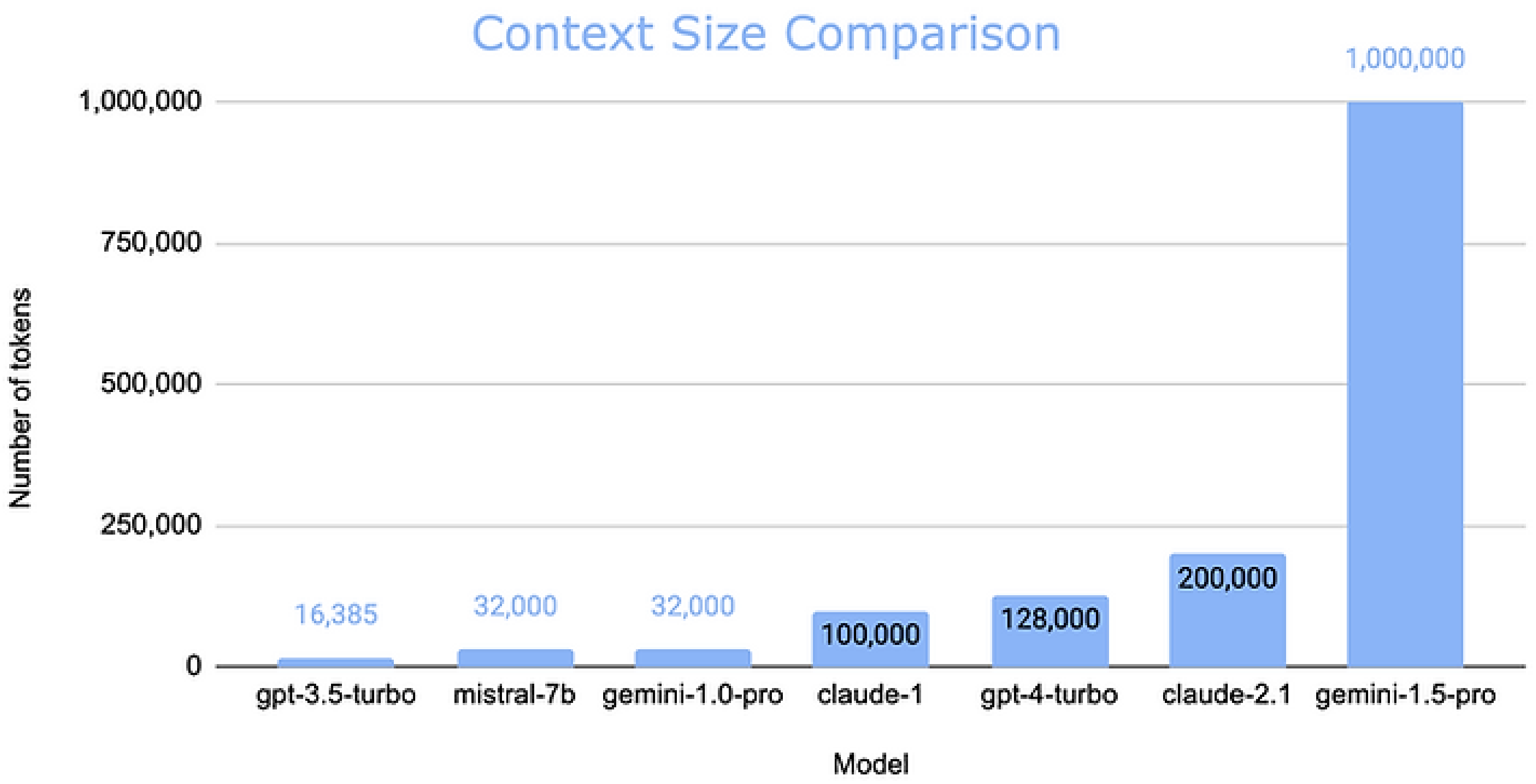
The context window is the amount of text (tokens) an LLM can "see" and process at once to generate meaningful responses.

**Key Components**

- Previous conversation history
- User-provided documents
- System instructions
- Current user query

**Importance**

- Enables coherent conversations
- Helps maintain topic relevance
- Allows document-based responses
- Critical for accuracy




From TRECCANI.it

**Linguistic Context**

- textual elements surrounding a word or phrase within discourse.
- The meaning of a word can change based on the sentences that precede or follow it

**Extralinguistic Context**

- This includes the physical, temporal, social, and cultural circumstances in which communication occurs
- Factors such as location, time, relationships between participants, and cultural norms play a crucial role in interpreting messages

## What about LLMs?

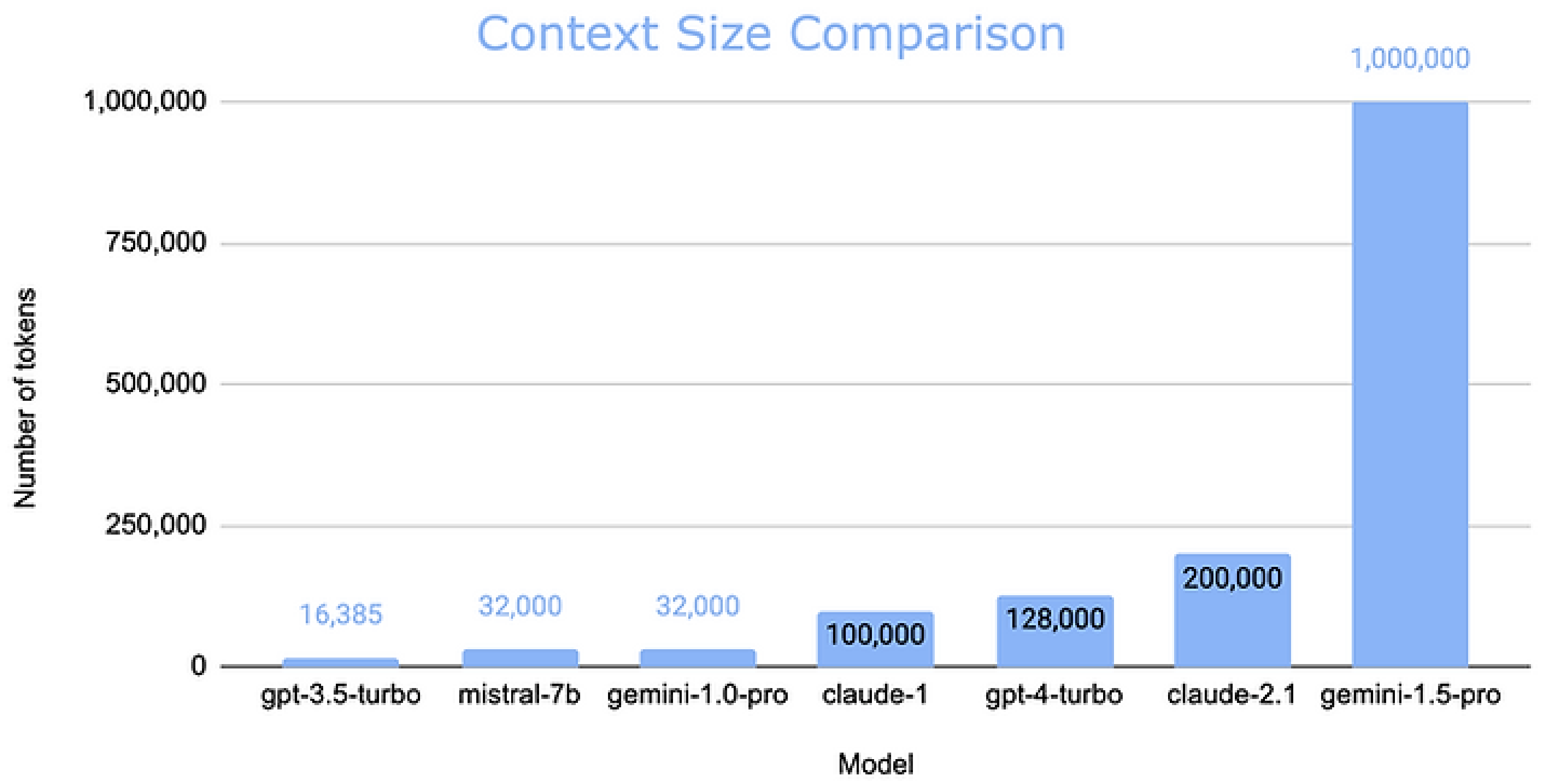
The context window is the amount of text (tokens) an LLM can "see" and process at once to generate meaningful responses.

**Key Components**

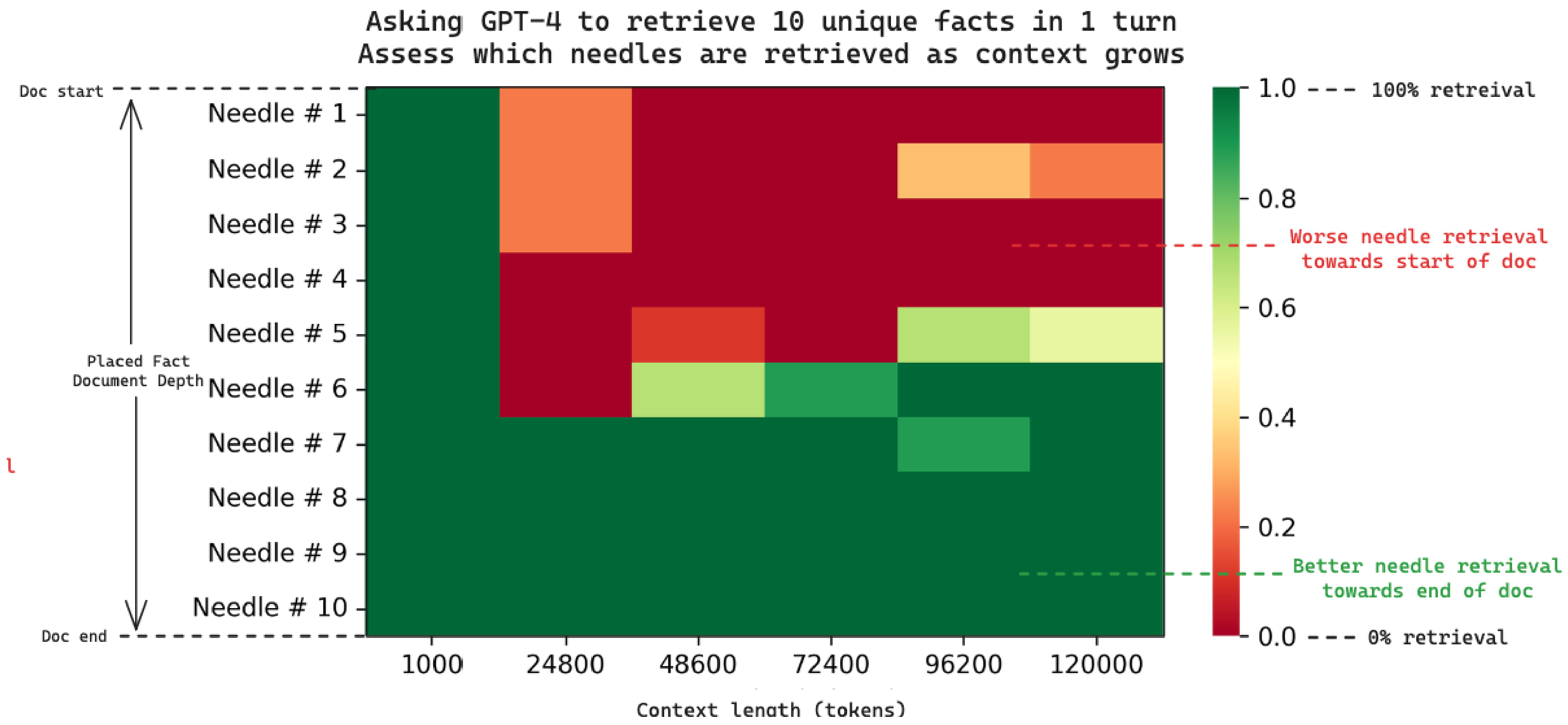
- Previous conversation history
- User-provided documents
- System instructions
- Current user query

**Importance**

- Enables coherent conversations
- Helps maintain topic relevance
- Allows document-based responses
- Critical for accuracy



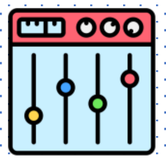
# 3.2 The Needle in a Haystack


**In-Context Learning (ICL)**
- It consists on feeding context with direct prompts
  - Few shot Learning
  - One shot Learning
  - Zero shot Learning


**Retrieval Augemented Generation (RAG)**
- The LLM generates responses sing prompts and vectorized documents
- It allows updated knowledge from external sources


**Fine-Tuning**
- Adapting the model to a specific domain
- Improving it for specific performance
- Incorporates knowledge internally



**Tool Augmentation**
- Access data in real time
- Integration of APIs and Services
- External Search

## 3.4 In-Context Learning

**Zero-shot Learning**

- ❑ The model performs a task without any prior examples.
- ❑ Relies solely on its pre-trained knowledge.

**One-shot Learning**

- ❑ The model is given **a single example** to understand the task.
- ❑ Helps the model generalize from minimal context.

**Few-shot Learning**

- ❑ The model is given **a small number of examples** (typically 2-5) to understand the task.
- ❑ Improves performance by providing more context than one-shot.

**Chain of Thought (CoT)**

- ❑ The model breaks down complex problems into intermediate reasoning steps.
- ❑ Mimics human-like problem-solving by "thinking aloud.“

**Key Differences**

- **Zero-shot**: No examples, relies on generalization.
- **One-shot**: One example, minimal context.
- **Few-shot**: Few examples, improves with more context.
- **CoT**: Focuses on reasoning steps, ideal for complex tasks.

# 3.5 Prompting Techniques – In-Context Learning

### Zero-shot prompt

Prompting without examples

Classify the review into neutral, negative, or positive.

**Text:** I think the movie was decent.
**Sentiment:** ...

### One-shot prompt

Prompting with a single example

Classify the review into neutral, negative, or positive.

**Text:** I think the movie was okay.
**Sentiment:** Neutral

**Text:** I think the movie was decent.
**Sentiment:**

### Few-shot prompt

Prompting with more than one example

Classify the review into neutral, negative, or positive.

**Text:** I think the movie was okay.
**Sentiment:** Neutral

**Text:** I think the movie was amazing!
**Sentiment:** Positive

**Text:** I think the movie was terrible...
**Sentiment:** Negative

**Text:** I think the movie was decent.
**Sentiment:**

### One-shot prompt

Prompting with a single example

**Q:**
Marco has 8 colored pencils. His friend gives him 3 more boxes with 2 pencils each. How many pencils does he have now?

**A:**
The answer is 14.

**Q:**
The library has 45 books. They lend 12 books and receive 5 new donations. How many books do they have now?

↓

**A:**
The answer is 42. ✗

### Chain-of-thought prompt

Prompting with a reasoning example

**Q:**
Marco has 8 colored pencils. His friend gives him 3 more boxes with 2 pencils each. How many pencils does he have now?

**A:**
Marco starts with 8 pencils.
He gets 3 boxes of 2 pencils = 6 more pencils.
8 + 6 = 14 pencils total.

The answer is 14.

**Q:**
The library has 45 books. They lend 12 books and receive 5 new donations. How many books do they have now?

↓

**A:** ✓
The library starts with 45 books.
They lend out 12 books: 45 - 12 = 33 books.
They receive 5 new books: 33 + 5 = 38 books.

The answer is 38.

## Objective

- **LLM Prompting**
  - Working with different LLM providers (Google, MistralAI, Groq)
  - Zero Shot Learning
  - Few Shot Learning
  - CoT
- **Adding Memory**
- **Sentiment Analysis**
  - Comparison with RoBERTA
  - Dataset:
    - Rotten Tomatoes movie reviews
    - Binary classification: positive (1) vs negative (0) reviews
    - Training set + Test set for evaluation
- Using Tools
  - Tavily AI
- Agents