# Attributed Generation: What works and what fails

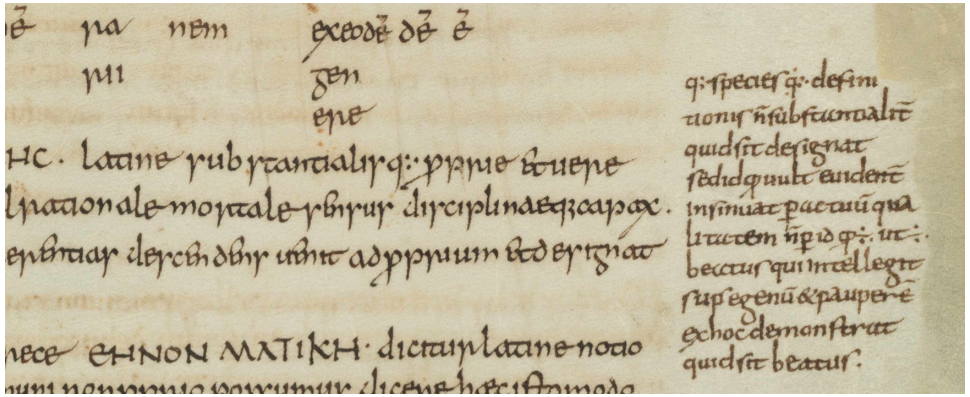
Daniel Khashabi

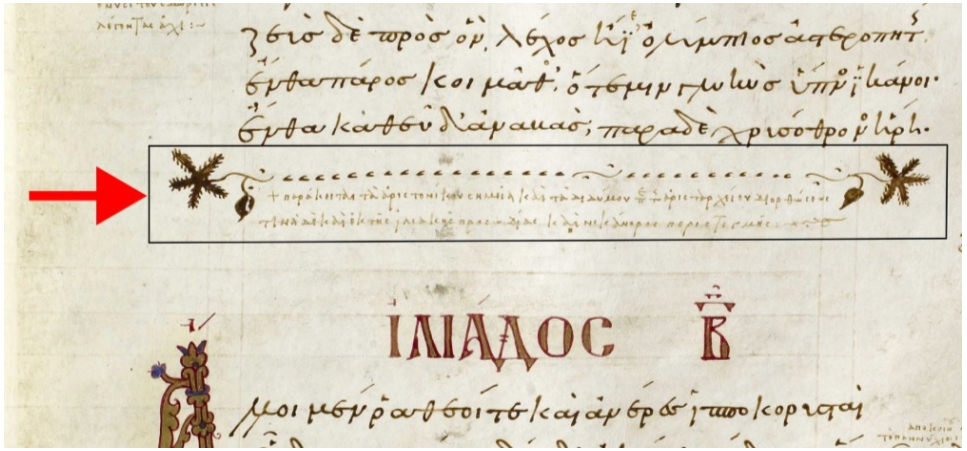
# Attribution: Why?

- A claim is *only* as trustworthy as its source.

# Attribution: Why?

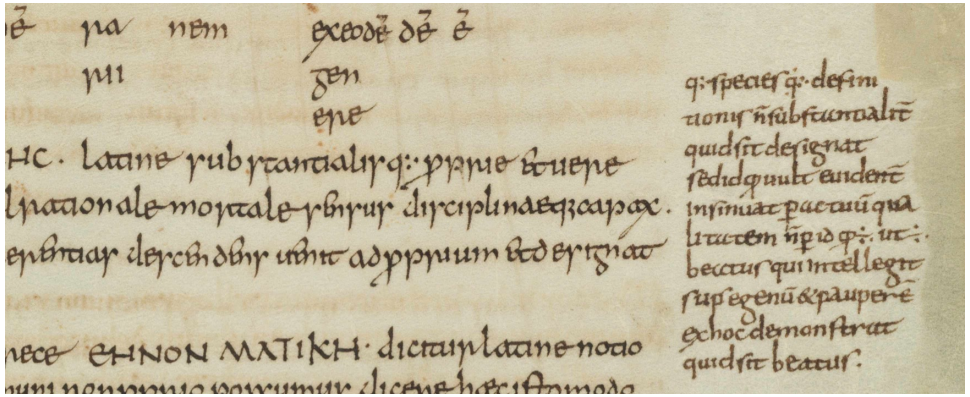
- A claim is *only* as trustworthy as its source.

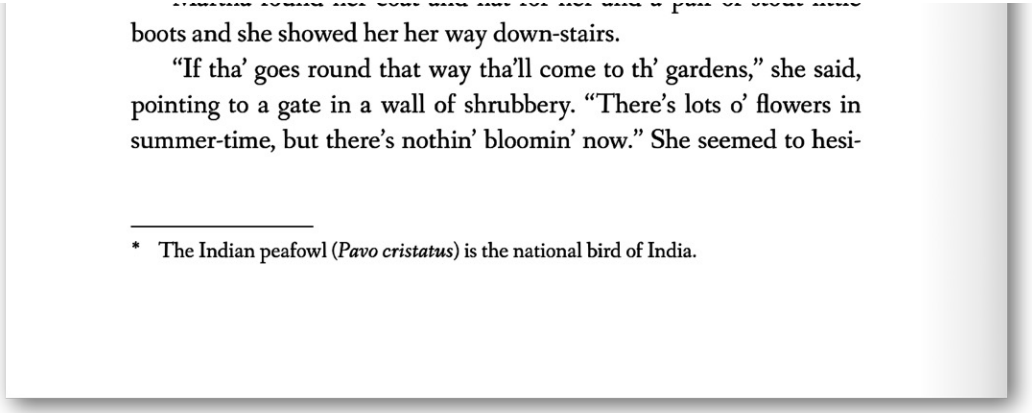




# Attribution: Why?

- A claim is *only* as trustworthy as its source.

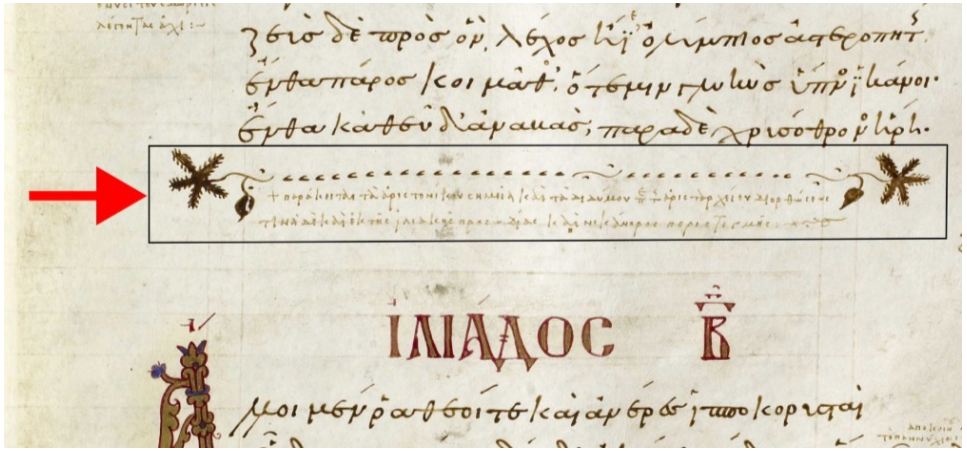


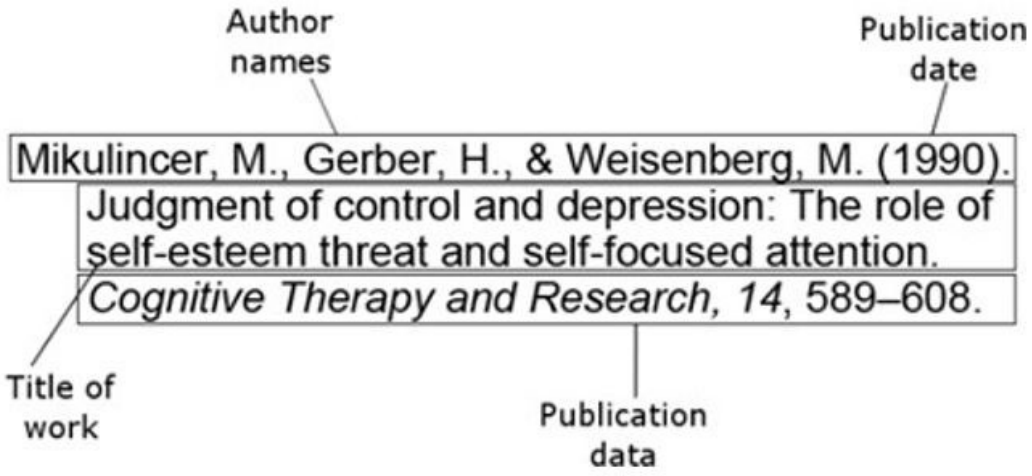


boots and she showed her her way down-stairs.

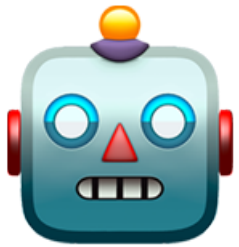
"If tha' goes round that way tha'll come to th' gardens," she said, pointing to a gate in a wall of shrubbery. "There's lots o' flowers in summer-time, but there's nothin' bloomin' now." She seemed to hesi-

* The Indian peafowl (*Pavo cristatus*) is the national bird of India.



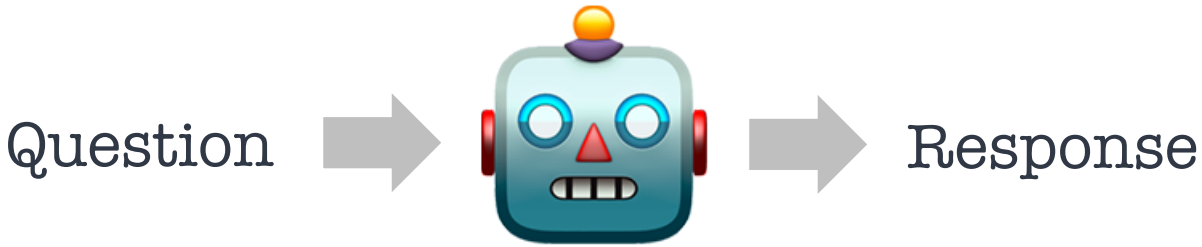


# Vanilla LLMs: Fluent, but Unverifiable

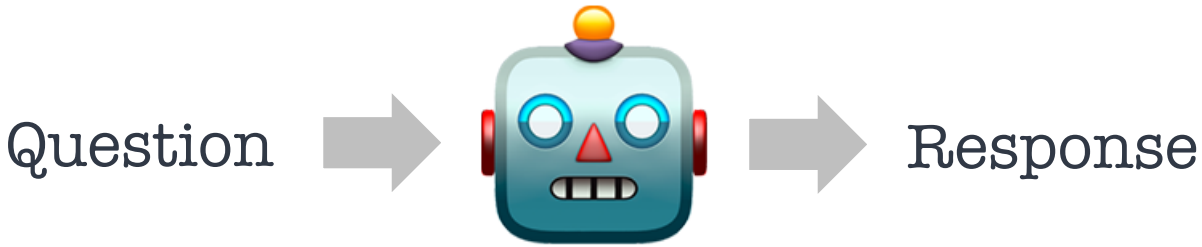
Question →  → Response

# Vanilla LLMs: Fluent, but Unverifiable

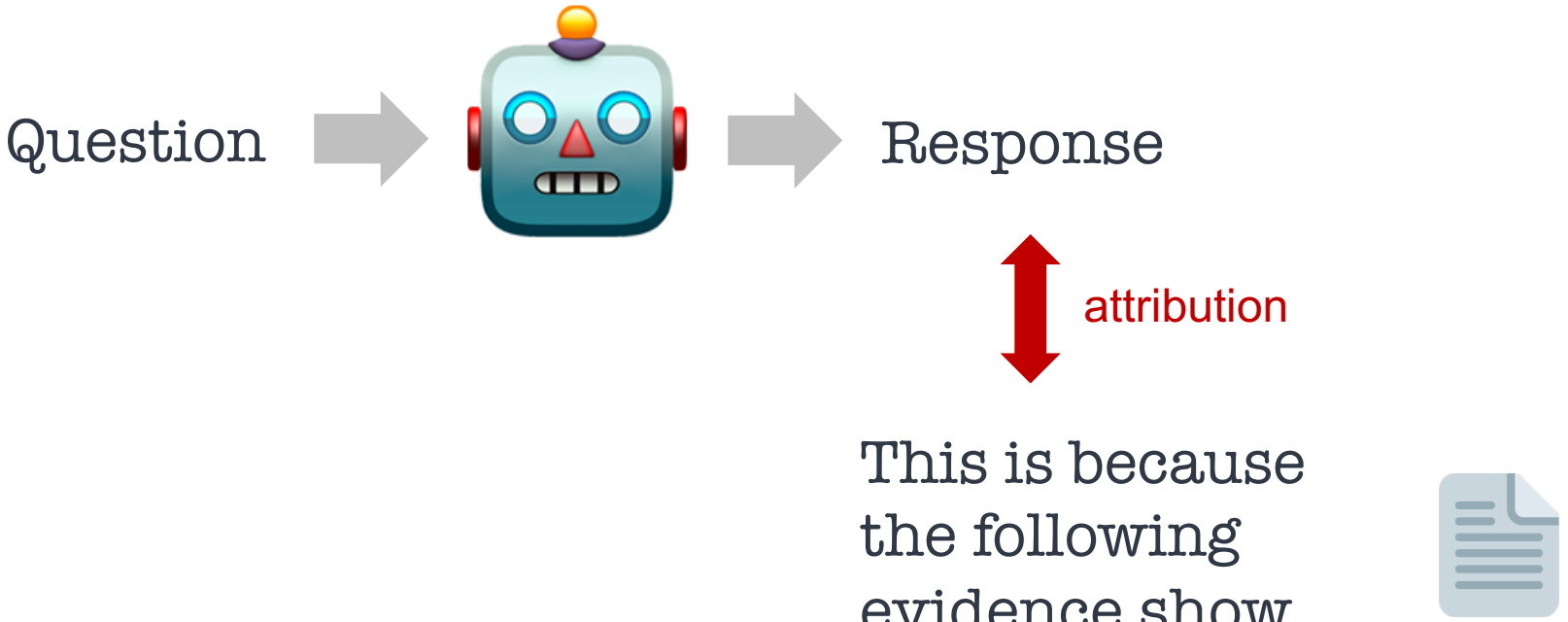
- A vanilla LLM takes a prompt and produces a response
- Becomes a bottleneck:
  - Correct facts cannot be distinguished from hallucinated ones without external verification → There is no audit trail



# Attributed LLM Generation

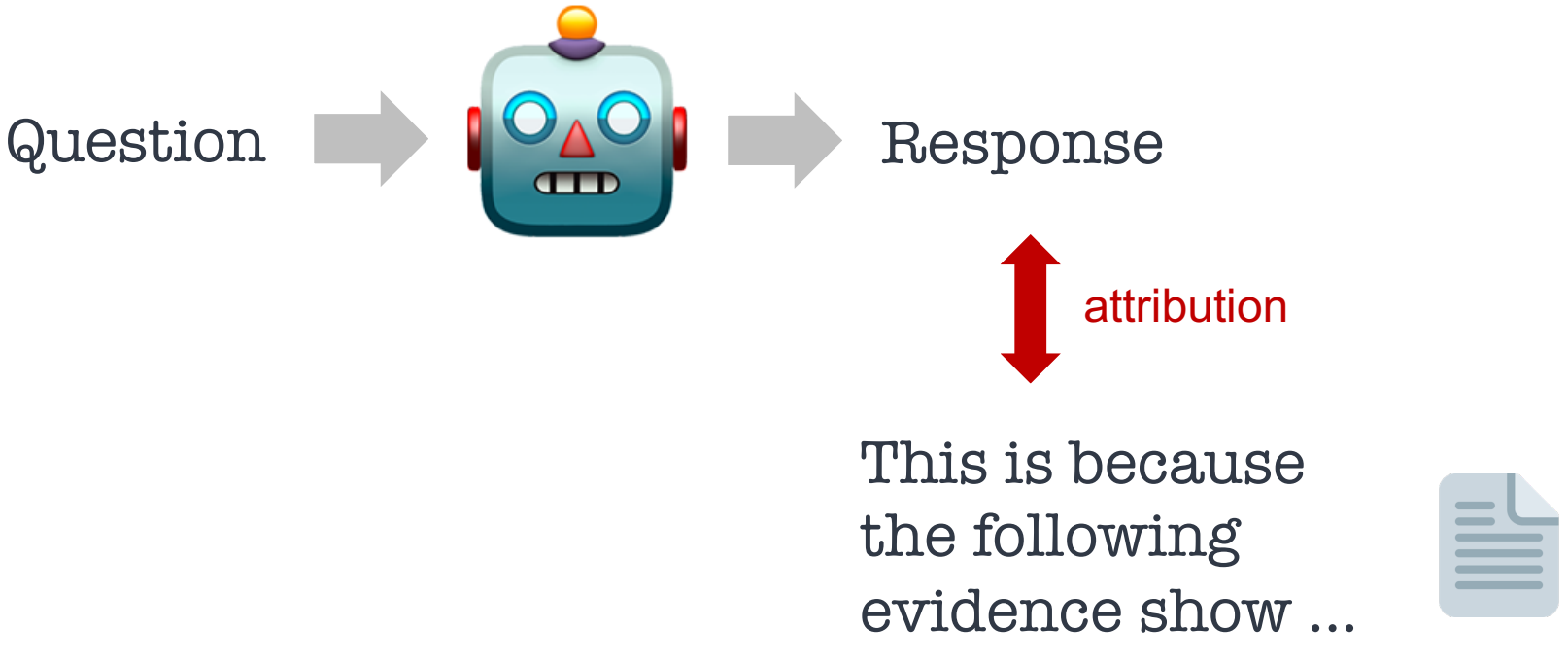


# Attributed LLM Generation



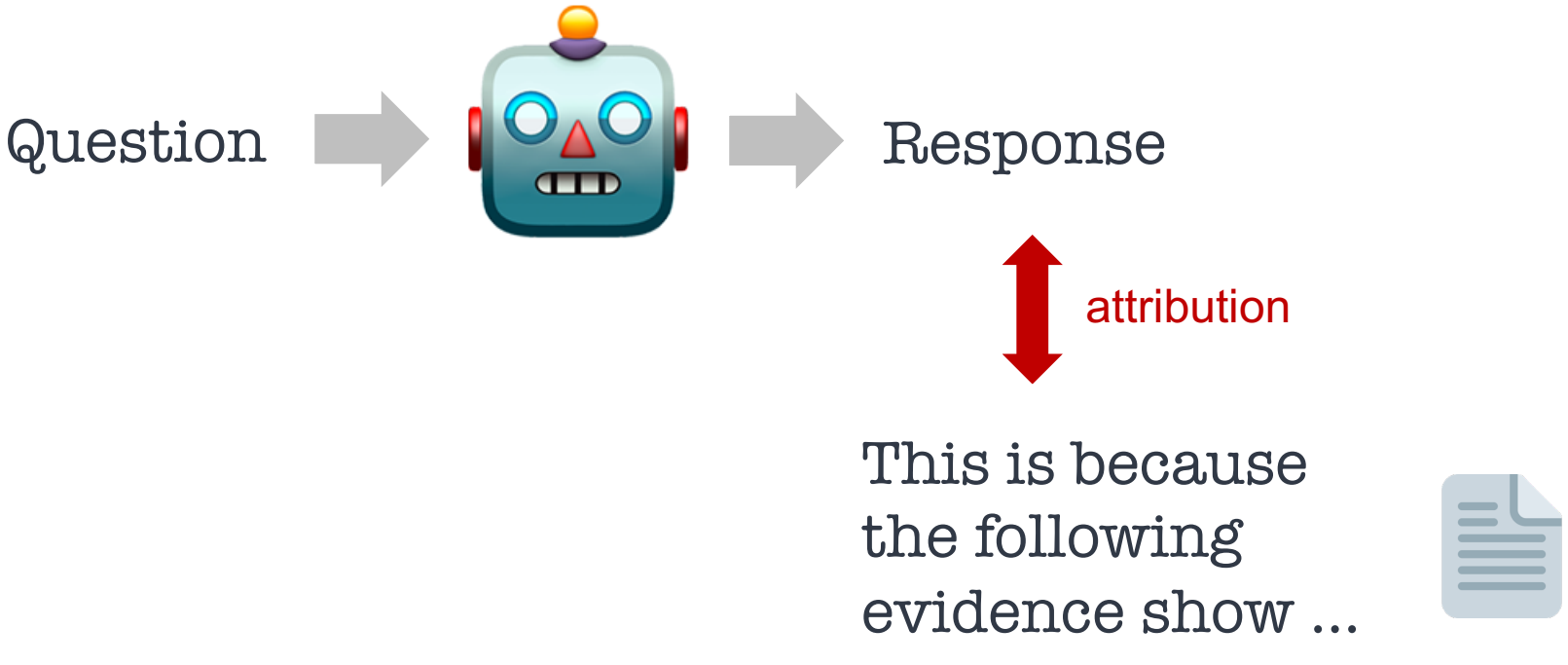
# Attributed LLM Generation

- In attributed generation, each claim is linked to verifiable sources.
- Allows a human inspector to trace a claim to its evidence.
- The ideal is faithful attribution, i.e., the source supports what is claimed.



# Attributed LLM Generation

- In attributed generation, each claim is linked to verifiable sources.
- Allows a human inspector to trace a claim to its evidence.
- The ideal is faithful attribution, i.e., the source supports what is claimed.



# The (Dominant) Paradigms for Attribution

1. **Retrieval-Augmented Generation (RAG)**
2. **Sketching**
3. **Knowledge Bases**

# The (Dominant) Paradigms for Attribution

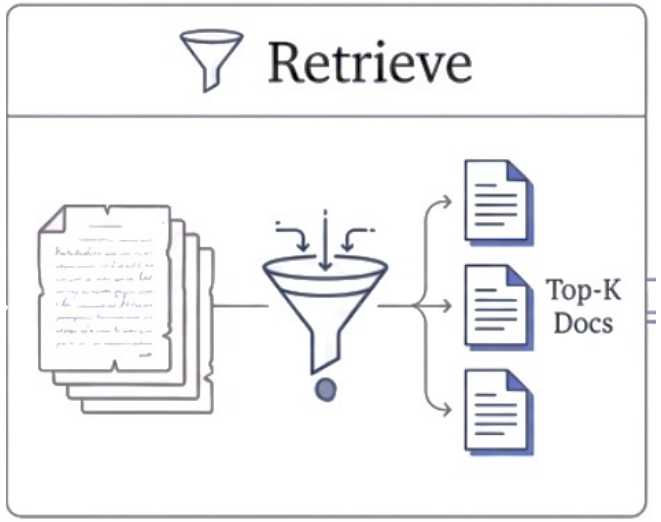
1. **Retrieval-Augmented Generation (RAG)**
2. **Sketching**
3. **Knowledge Bases**

There is no single silver bullet.
Effective attribution depends on properties of your problem/data.

# ① Retrieval-Augmented Generation (RAG)

# ① Retrieval-Augmented Generation (RAG)

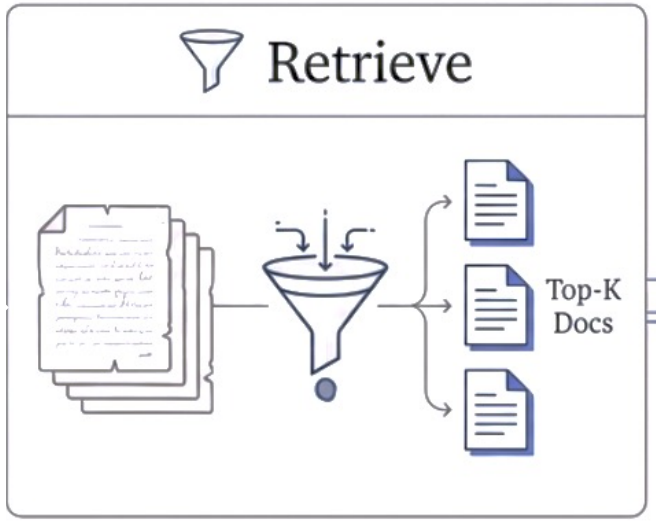
Question




Retrieve most
“relevant” documents
to the given query

# ① Retrieval-Augmented Generation (RAG)

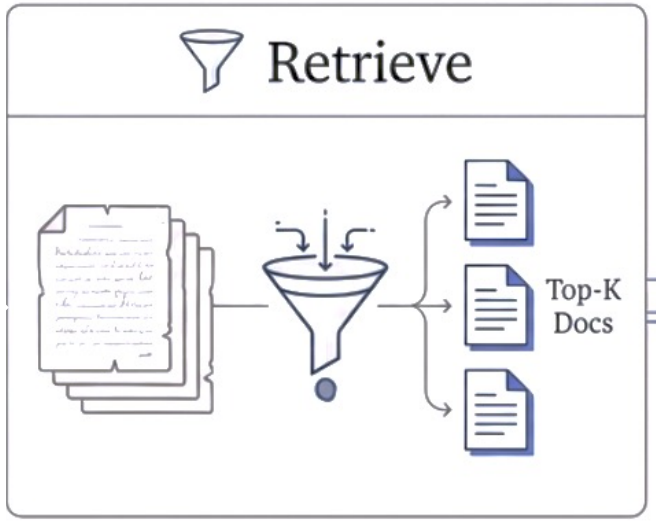
Question




Retrieve most “relevant” documents to the given query

# ① Retrieval-Augmented Generation (RAG)

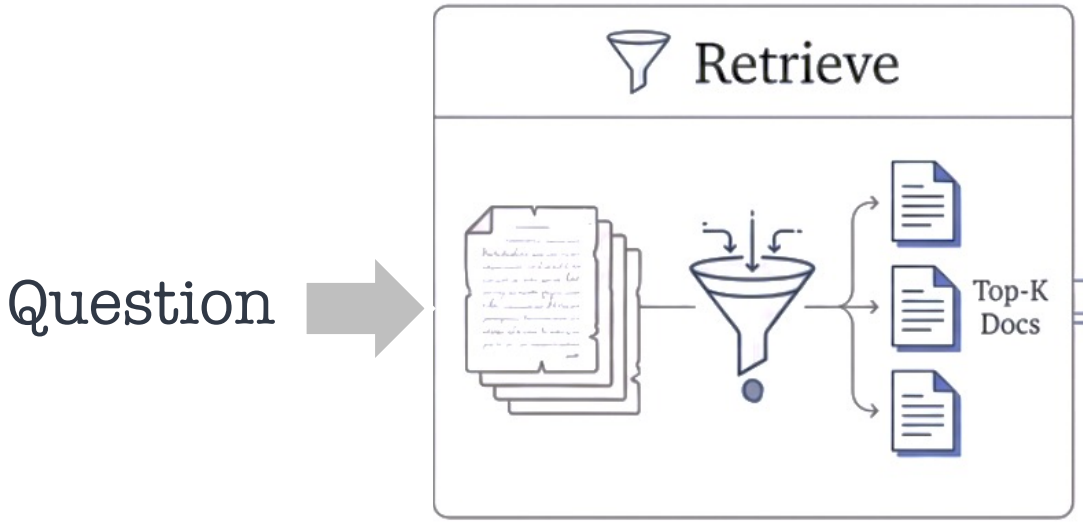
Question



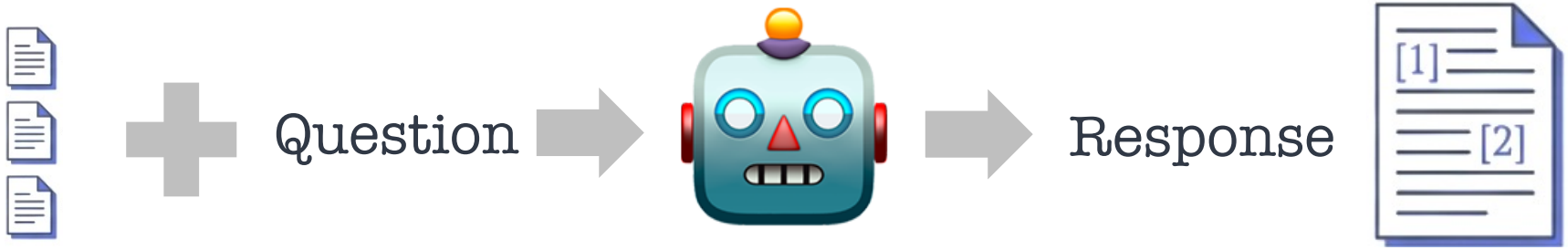

Retrieve most
“relevant” documents
to the given query

Question

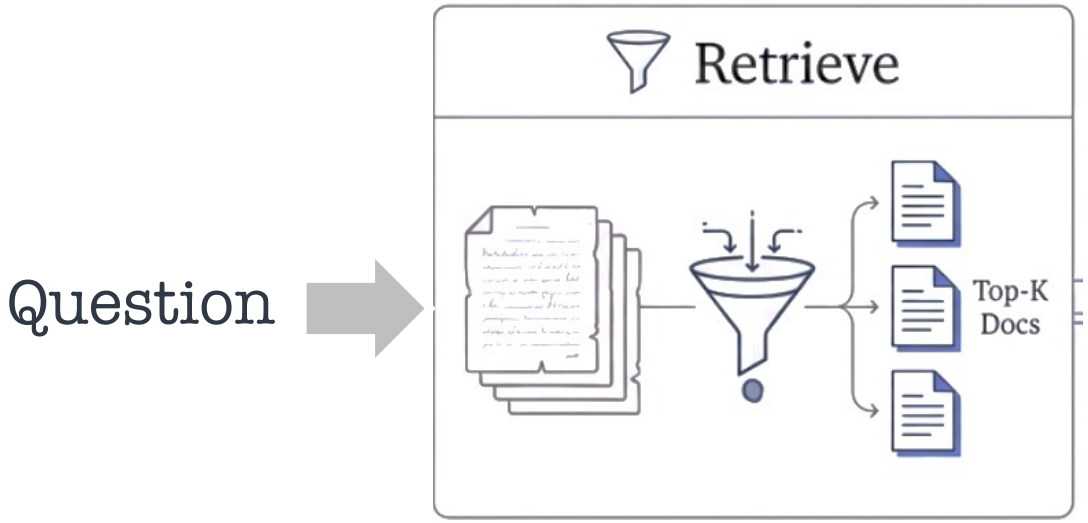
# ① Retrieval-Augmented Generation (RAG)



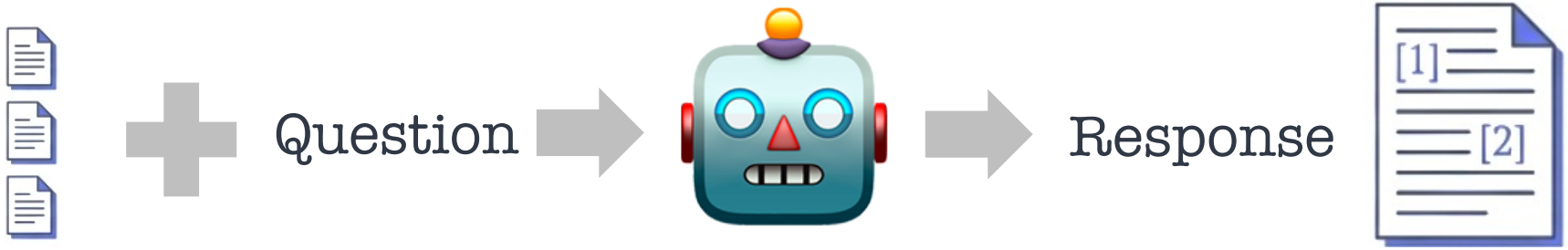

Retrieve most
“relevant” documents
to the given query



# ① Retrieval-Augmented Generation (RAG)

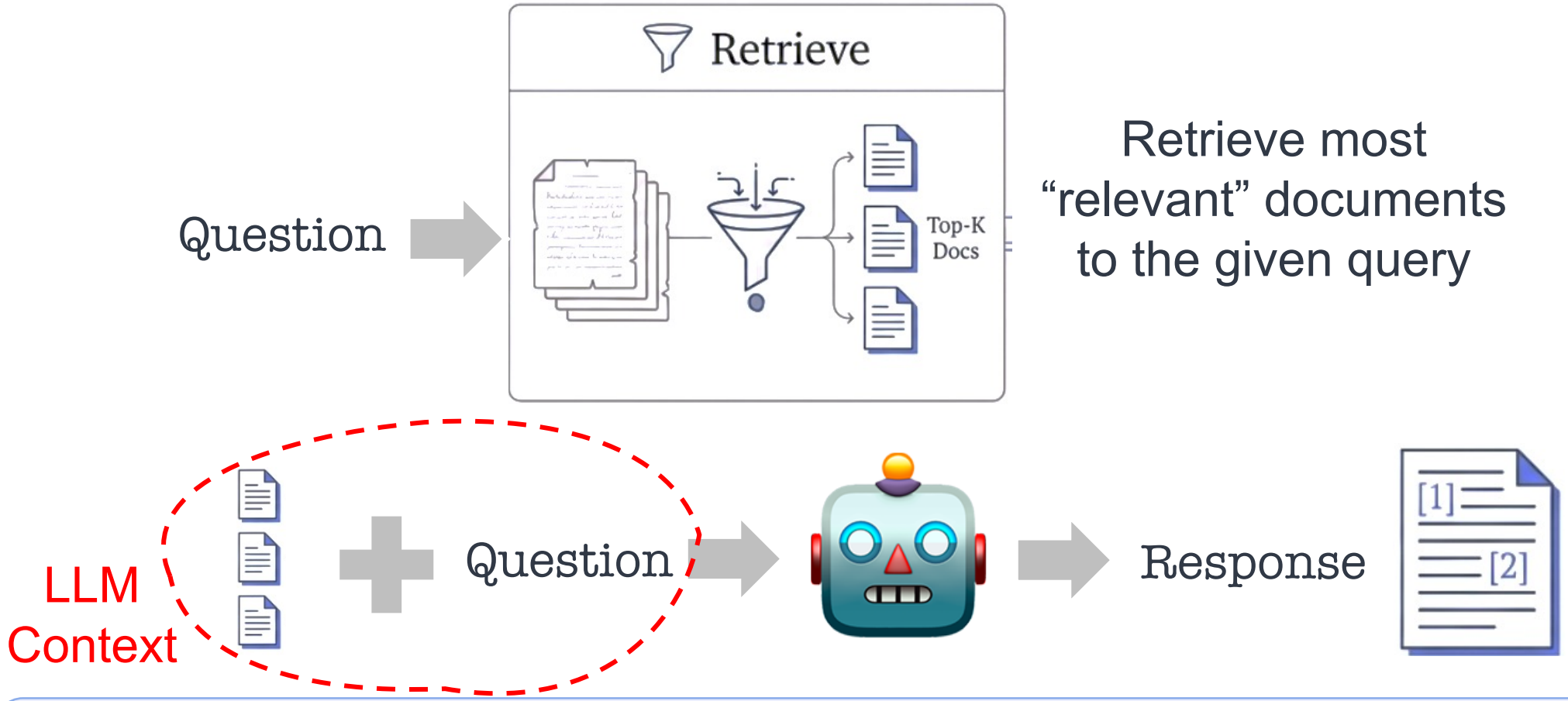



Retrieve most
“relevant” documents
to the given query




RAG is a flexible and powerful mechanism for attribution.

# ① Retrieval-Augmented Generation (RAG)




RAG is a flexible and powerful mechanism for attribution.

# ① RAG Challenge: Long-Context Biases

# ① RAG Challenge: Long-Context Biases

| Models | Claimed Length | Effective Length |
|---|---|---|
| Llama2 (7B) | 4K | - |
| Gemini-1.5-Pro | 1M | >128K |
| GPT-4 | 128K | 64K |
| Llama3.1 (70B) | 128K | 64K |
| Qwen2 (72B) | 128K | 32K |
| Command-R-plus (104B) | 128K | 32K |
| GLM4 (9B) | 1M | 64K |
| Llama3.1 (8B) | 128K | 32K |
| GradientAI/Llama3 (70B) | 1M | 16K |
| Mixtral-8x22B (39B/141B) | 64K | 32K |
| Yi (34B) | 200K | 32K |
| Phi3-medium (14B) | 128K | 32K |
| Mistral-v0.2 (7B) | 32K | 16K |
| LWM (7B) | 1M | <4K |
| DBRX (36B/132B) | 32K | 8K |
| Together (7B) | 32K | 4K |
| LongChat (7B) | 32K | <4K |
| LongAlpaca (13B) | 32K | <4K |

Hsieh et al. RULER: What's the Real Context Size of Your Long-Context Language Models? COLM.

# ① RAG Challenge: Long-Context Biases

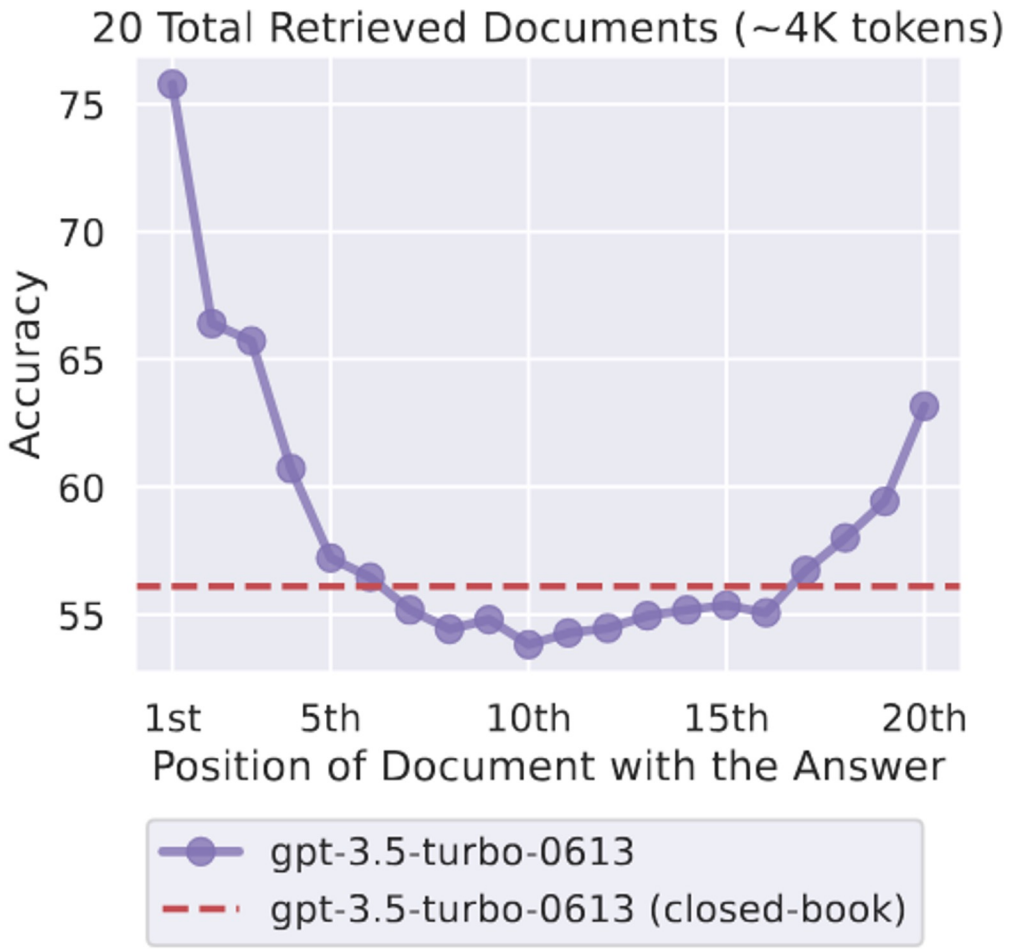
The effective context window of models is much smaller than what they're claimed to be.

| Models | Claimed Length | Effective Length |
|---|---|---|
| Llama2 (7B) | 4K | - |
| Gemini-1.5-Pro | 1M | >128K |
| GPT-4 | 128K | 64K |
| Llama3.1 (70B) | 128K | 64K |
| Qwen2 (72B) | 128K | 32K |
| Command-R-plus (104B) | 128K | 32K |
| GLM4 (9B) | 1M | 64K |
| Llama3.1 (8B) | 128K | 32K |
| GradientAI/Llama3 (70B) | 1M | 16K |
| Mixtral-8x22B (39B/141B) | 64K | 32K |
| Yi (34B) | 200K | 32K |
| Phi3-medium (14B) | 128K | 32K |
| Mistral-v0.2 (7B) | 32K | 16K |
| LWM (7B) | 1M | <4K |
| DBRX (36B/132B) | 32K | 8K |
| Together (7B) | 32K | 4K |
| LongChat (7B) | 32K | <4K |
| LongAlpaca (13B) | 32K | <4K |

Hsieh et al. RULER: What's the Real Context Size of Your Long-Context Language Models? COLM.

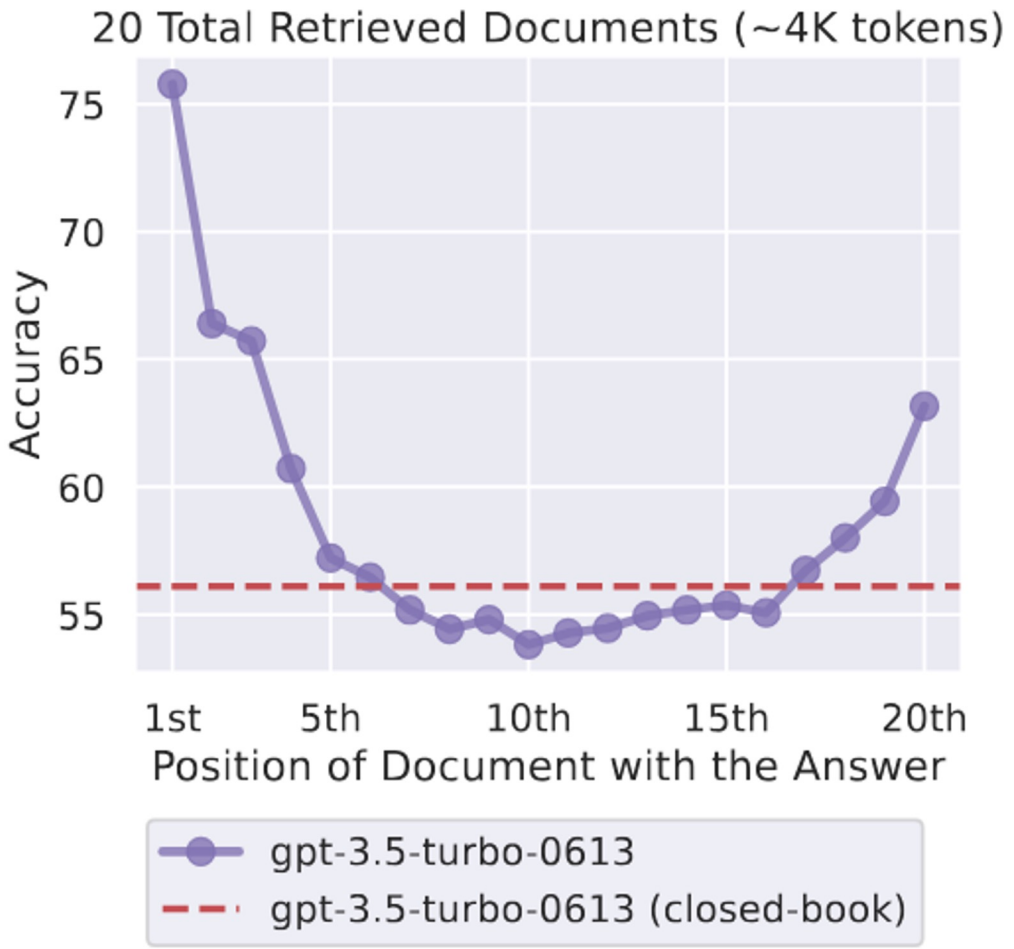
# ① RAG Challenge: Long-Context Biases

# ① RAG Challenge: Long-Context Biases




Liu et al. (2024). Lost in the middle: How language models use long contexts, TACL.

# ① RAG Challenge: Long-Context Biases




Liu et al. (2024). Lost in the middle: How language models use long contexts, TACL.

Models attend more to documents at the beginning and end of the context

# ① RAG Challenge: Linguistic Nepotism

Ki et al. (2026) Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG, ICML.

# ① RAG Challenge: Linguistic Nepotism

- LLMs have preferences for languages (e.g., English > Greek).
  - Question (input) in English
  - **Two** evidence docs; **only one** relevant

Ki et al. (2026) Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG, ICML.

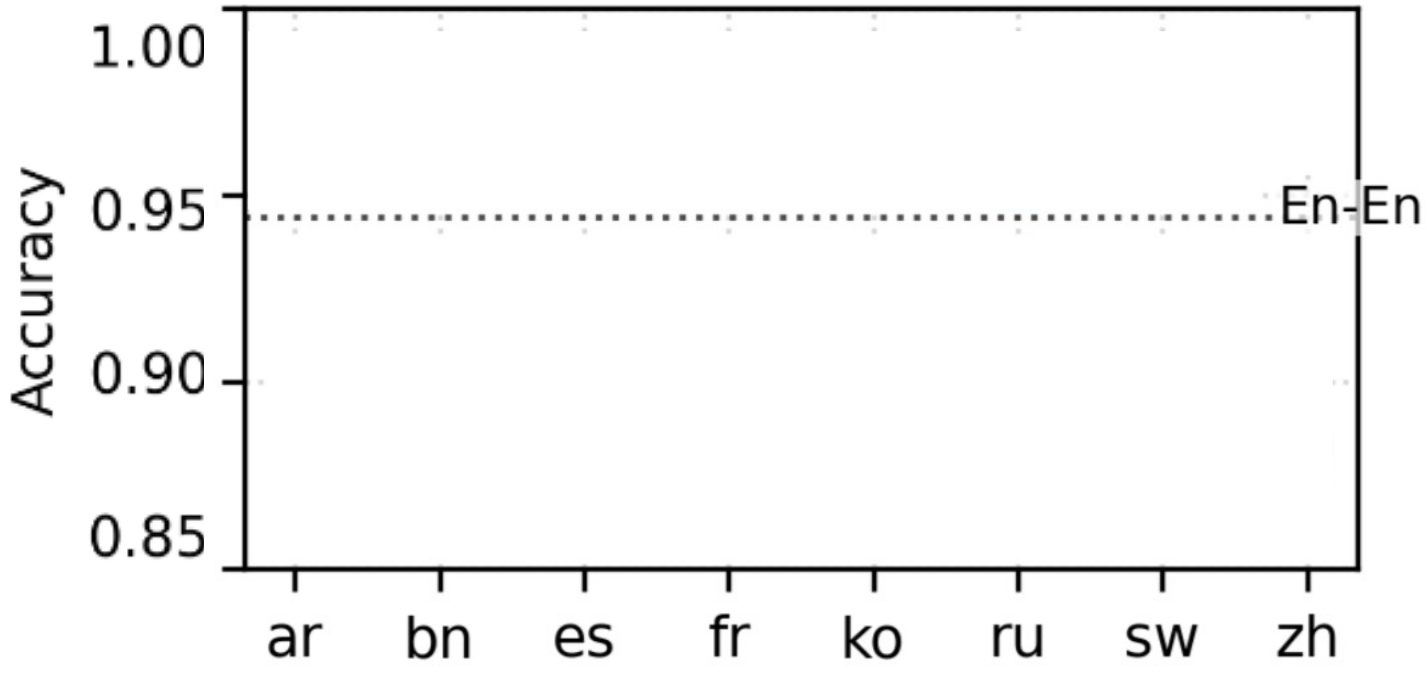
# ① RAG Challenge: Linguistic Nepotism

- LLMs have preferences for languages (e.g., English > Greek).
  - Question (input) in English
  - **Two** evidence docs; **only one** relevant

Ki et al. (2026) Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG, ICML.

# ① RAG Challenge: Linguistic Nepotism

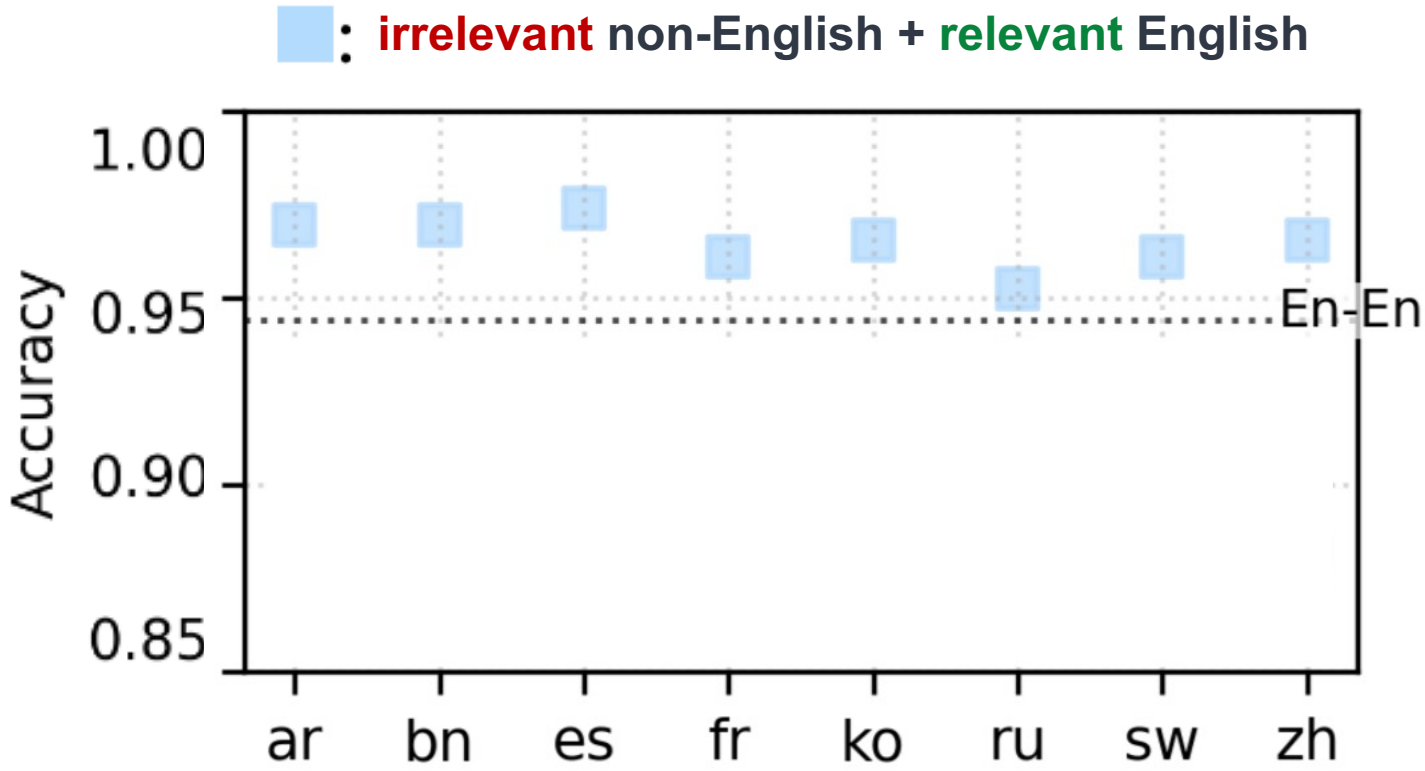
- LLMs have preferences for languages (e.g., English > Greek).
  - Question (input) in English
  - **Two** evidence docs; **only one** relevant




Ki et al. (2026) Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG, ICML.

# ① RAG Challenge: Linguistic Nepotism

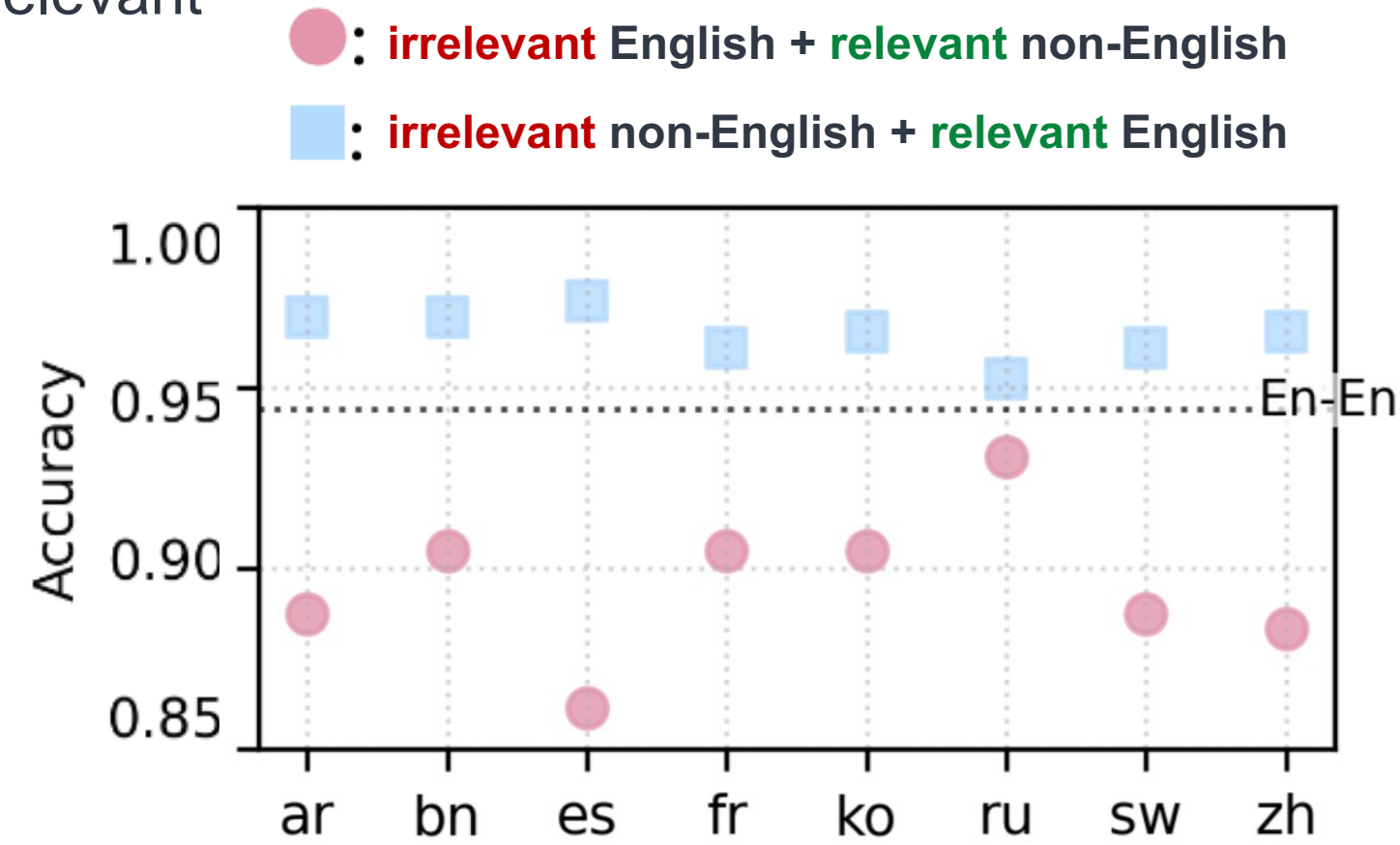
- LLMs have preferences for languages (e.g., English > Greek).
  - Question (input) in English
  - **Two** evidence docs; **only one** relevant




Ki et al. (2026) Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG, ICML.

# ① RAG Challenge: Linguistic Nepotism

- LLMs have preferences for languages (e.g., English > Greek).
  - Question (input) in English
  - **Two** evidence docs; **only one** relevant

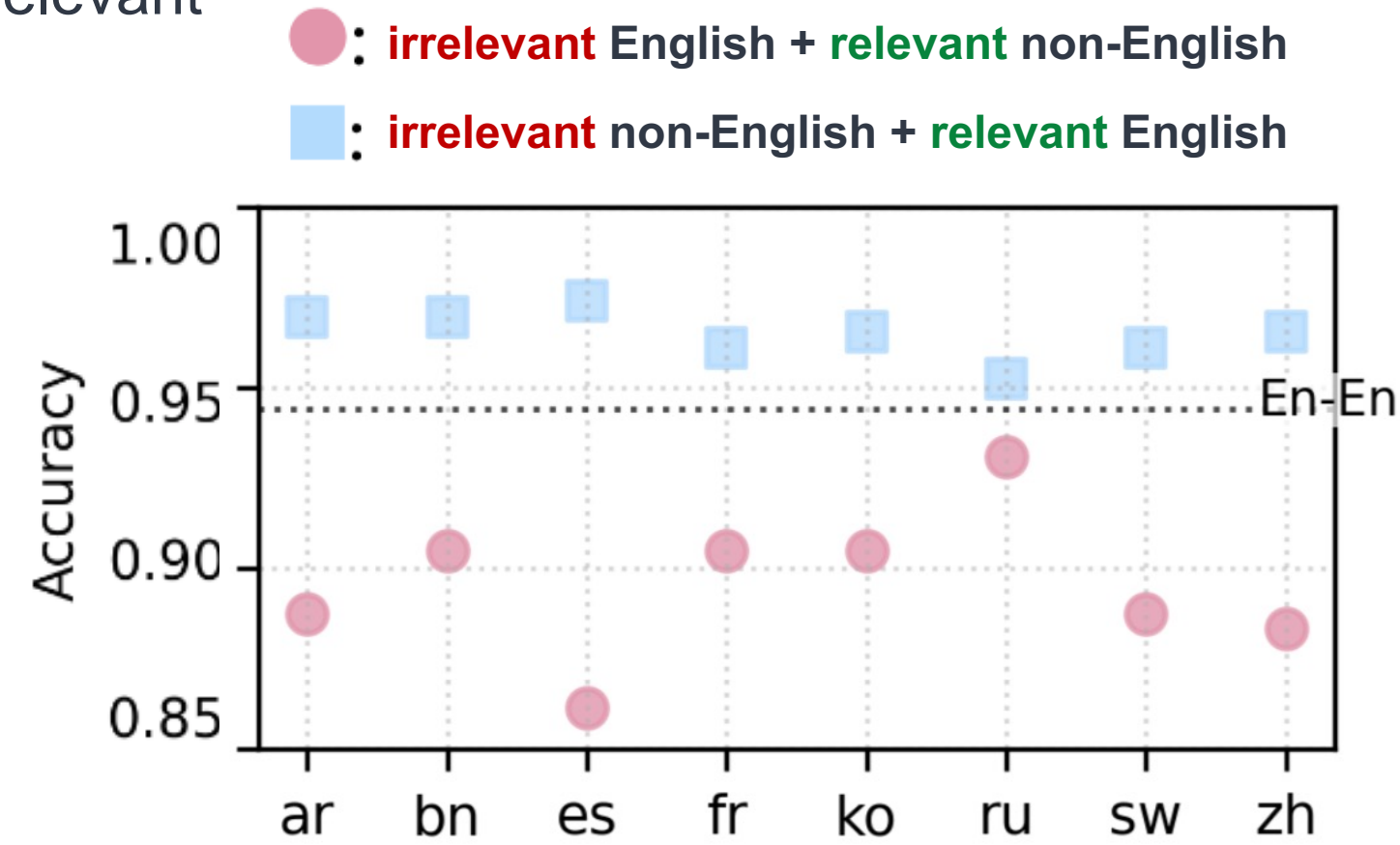



Ki et al. (2026) Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG, ICML.

# ① RAG Challenge: Linguistic Nepotism

- LLMs have preferences for languages (e.g., English > Greek).
  - Question (input) in English
  - **Two** evidence docs; **only one** relevant

When forced to choose, models will actively prefer **English** docs over **non-English docs**. (at the cost of relevance)




Ki et al. (2026) Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG, ICML.

# ② Attribution via Sketching

Marone and Van Durme (2023). Data Portraits: Recording Foundation Model Training Data, NeurIPS.
Liu et al. (2024). Infinigram: Scaling Unbounded N-gram Language Models to a Trillion Tokens, COLM.

2

# Attribution via Sketching

Trusted scholarly data

Marone and Van Durme (2023). Data Portraits: Recording Foundation Model Training Data, NeurIPS.
Liu et al. (2024). Infinigram: Scaling Unbounded N-gram Language Models to a Trillion Tokens, COLM.

# ② Attribution via Sketching

Trusted scholarly data

Sketch := A compressed representation of the data

Marone and Van Durme (2023). Data Portraits: Recording Foundation Model Training Data, NeurIPS.
Liu et al. (2024). Infinigram: Scaling Unbounded N-gram Language Models to a Trillion Tokens, COLM.

# ② Attribution via Sketching

Trusted scholarly data

Sketch := A compressed representation of the data

Very fast lookup of (near) quoted statements.

Question

Water is an unusual substance in many ways, and one of its peculiarities is that it has its lowest density at 4 °C. As water cools from room temperature, it becomes denser and denser until it reaches 4 °C. After that, as it continues to cool, it becomes less dense again.

Marone and Van Durme (2023). Data Portraits: Recording Foundation Model Training Data, NeurIPS.
Liu et al. (2024). Infinigram: Scaling Unbounded N-gram Language Models to a Trillion Tokens, COLM.

# ② Attribution via Sketching

Trusted scholarly data

Sketch := A compressed representation of the data

Very fast lookup of (near) quoted statements.

Question

Water is an unusual substance in many ways, and one of its peculiarities is that it has its lowest density at 4 °C. As water cools from room temperature, it becomes denser and denser until it reaches 4 °C. After that, as it continues to cool, it becomes less dense again.

- Fast mechanism to identify (near) quotes from the trusted data
- Brittle to how the ideas are phrased

Marone and Van Durme (2023). Data Portraits: Recording Foundation Model Training Data, NeurIPS.
Liu et al. (2024). Infinigram: Scaling Unbounded N-gram Language Models to a Trillion Tokens, COLM.

# ② Attribution via Sketching

Trusted scholarly data

Sketch := A compressed representation of the data

Very fast lookup of (near) quoted statements.

Question

Water is an unusual substance in many ways, and one of its peculiarities is that it has its lowest density at 4 °C. As water cools from room temperature, it becomes denser and denser until it reaches 4 °C. After that, as it continues to cool, it becomes less dense again.

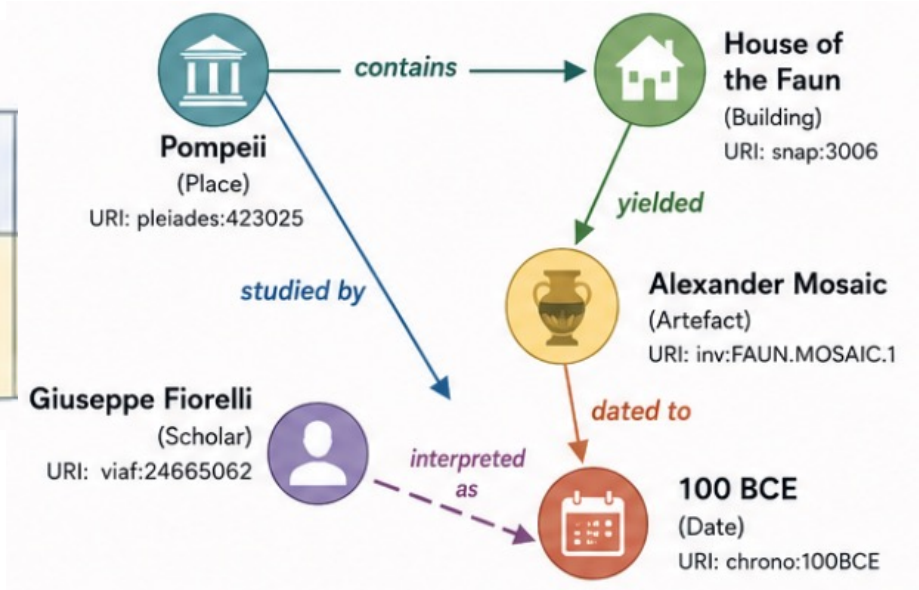
- Fast mechanism to identify (near) quotes from the trusted data
- Brittle to how the ideas are phrased

Marone and Van Durme (2023). Data Portraits: Recording Foundation Model Training Data, NeurIPS.
Liu et al. (2024). Infinigram: Scaling Unbounded N-gram Language Models to a Trillion Tokens, COLM.

# ③ Attribution via Structured Knowledge Bases

**Table: Artefact_Find**

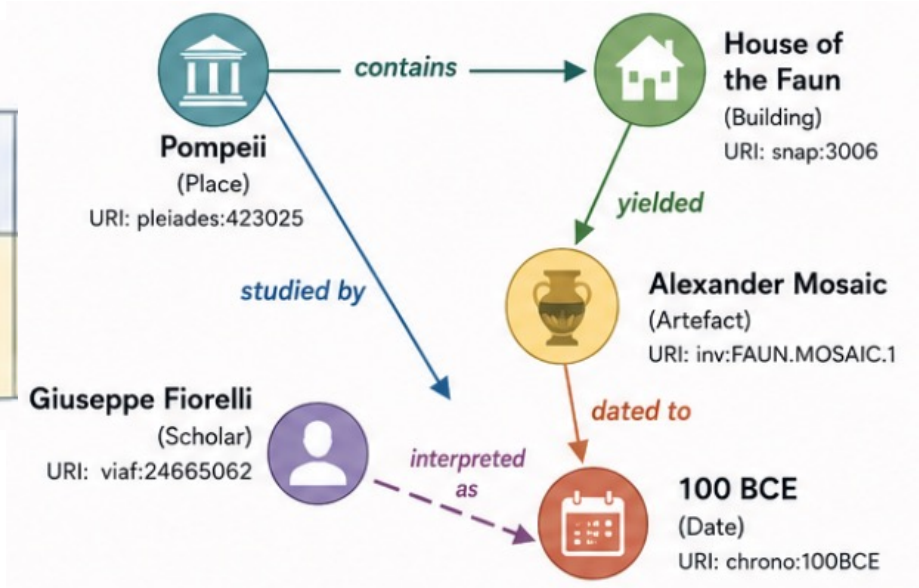
| find_id | artefact_id | artefact_name | find_spot | site_id | site_name | material | date_from | date_to |
|---|---|---|---|---|---|---|---|---|
| 83421 | INV:FAUN. MOSAIC.1 | Alexander Mosaic | House of the Faun | PLEIADES: 423025 | Pompeii | Tesserae (Mosaic) | -150 | -100 |



# ③ Attribution via Structured Knowledge Bases

**Table: Artefact_Find**

| find_id | artefact_id | artefact_name | find_spot | site_id | site_name | material | date_from | date_to |
|---|---|---|---|---|---|---|---|---|
| 83421 | INV:FAUN. MOSAIC.1 | Alexander Mosaic | House of the Faun | PLEIADES: 423025 | Pompeii | Tesserae (Mosaic) | -150 | -100 |

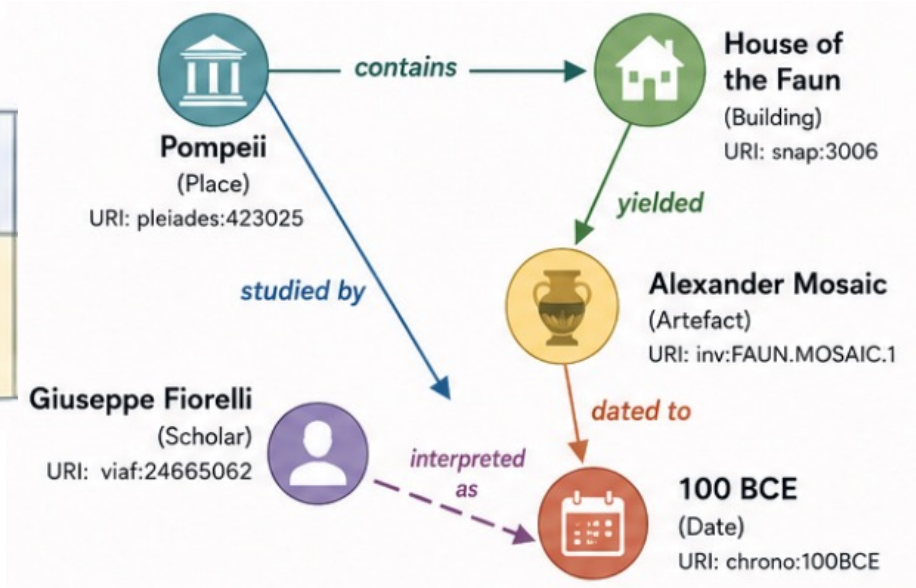



Question → 🤖 → Executable command against the knowledge base → ⌨

# ③ Attribution via Structured Knowledge Bases

Table: Artefact_Find

| find_id | artefact_id | artefact_name | find_spot | site_id | site_name | material | date_from | date_to |
|---|---|---|---|---|---|---|---|---|
| 83421 | INV:FAUN. MOSAIC.1 | Alexander Mosaic | House of the Faun | PLEIADES: 423025 | Pompeii | Tesserae (Mosaic) | -150 | -100 |




Question → 🤖 → Executable command against the knowledge base → 

- Obvious attribution: The results of the execution is the citation.
- Challenge: Historical artifacts resist rigid schemas.

# The (Dominant) Paradigms for Attribution

There is no single silver bullet.
Effective attribution depends on properties of your problem/data.

# The (Dominant) Paradigms for Attribution

**Retrieval-Augmented Generation (RAG)**

Retrieves documents at query time

Flexible and powerful

Suffers from various biases

There is no single silver bullet.
Effective attribution depends on properties of your problem/data.

# The (Dominant) Paradigms for Attribution

**Retrieval-Augmented Generation (RAG)**

Retrieves documents at query time

Flexible and powerful

Suffers from various biases

**Sketching**

Tests whether the output exists verbatim in a known corpus

Precise

Brittle (low recall)

There is no single silver bullet.
Effective attribution depends on properties of your problem/data.

# The (Dominant) Paradigms for Attribution

**Retrieval-Augmented Generation (RAG)**

Retrieves documents at query time

Flexible and powerful

Suffers from various biases

**Sketching**

Tests whether the output exists verbatim in a known corpus

Precise

Brittle (low recall)

**Knowledge Bases**

Grounds answers in queryable databases.

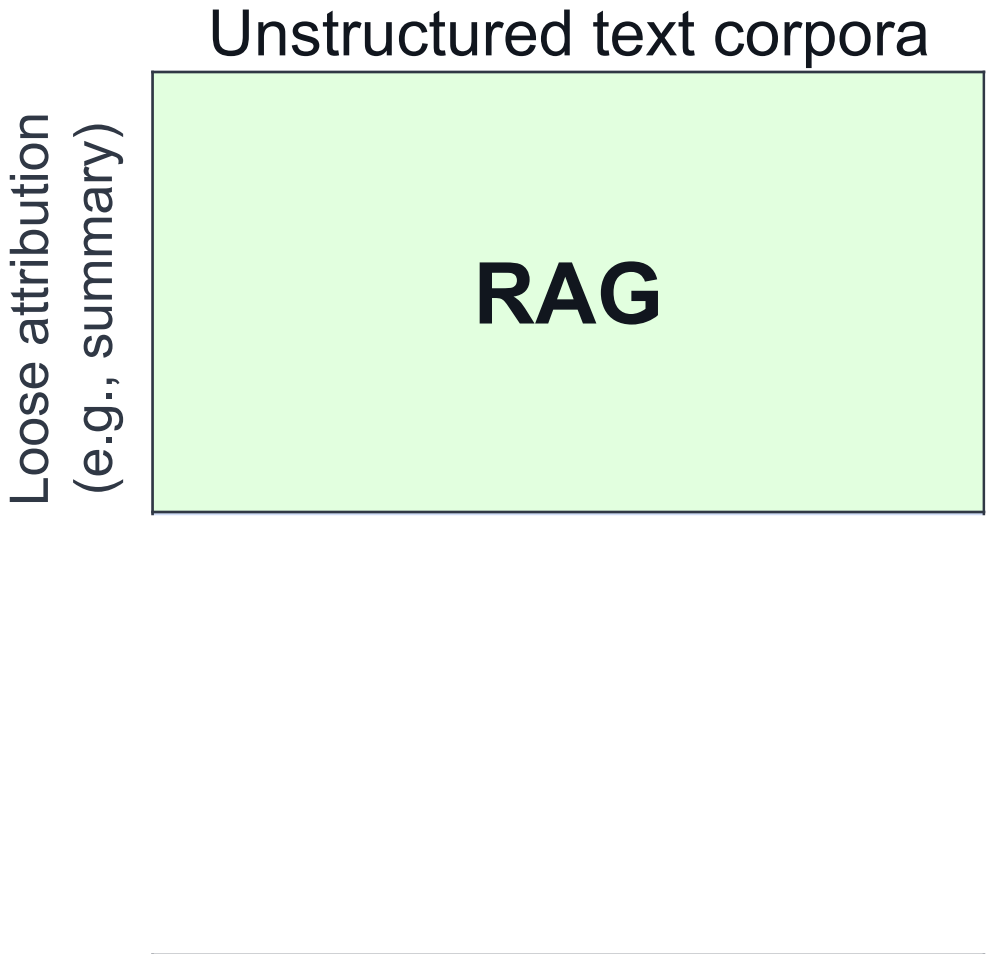
Verifiability by construction

Requires upfront data/schema curation

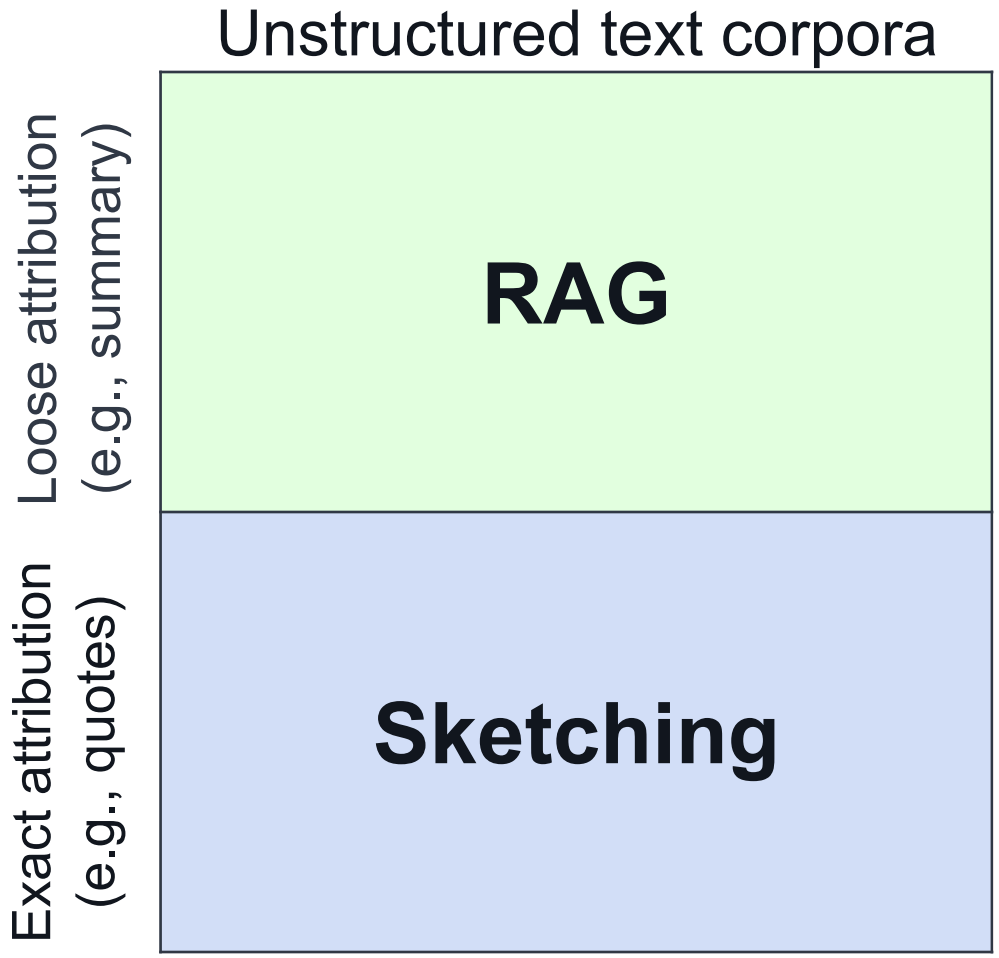
There is no single silver bullet.
Effective attribution depends on properties of your problem/data.

# Summary: Decision Tree of Paradigms

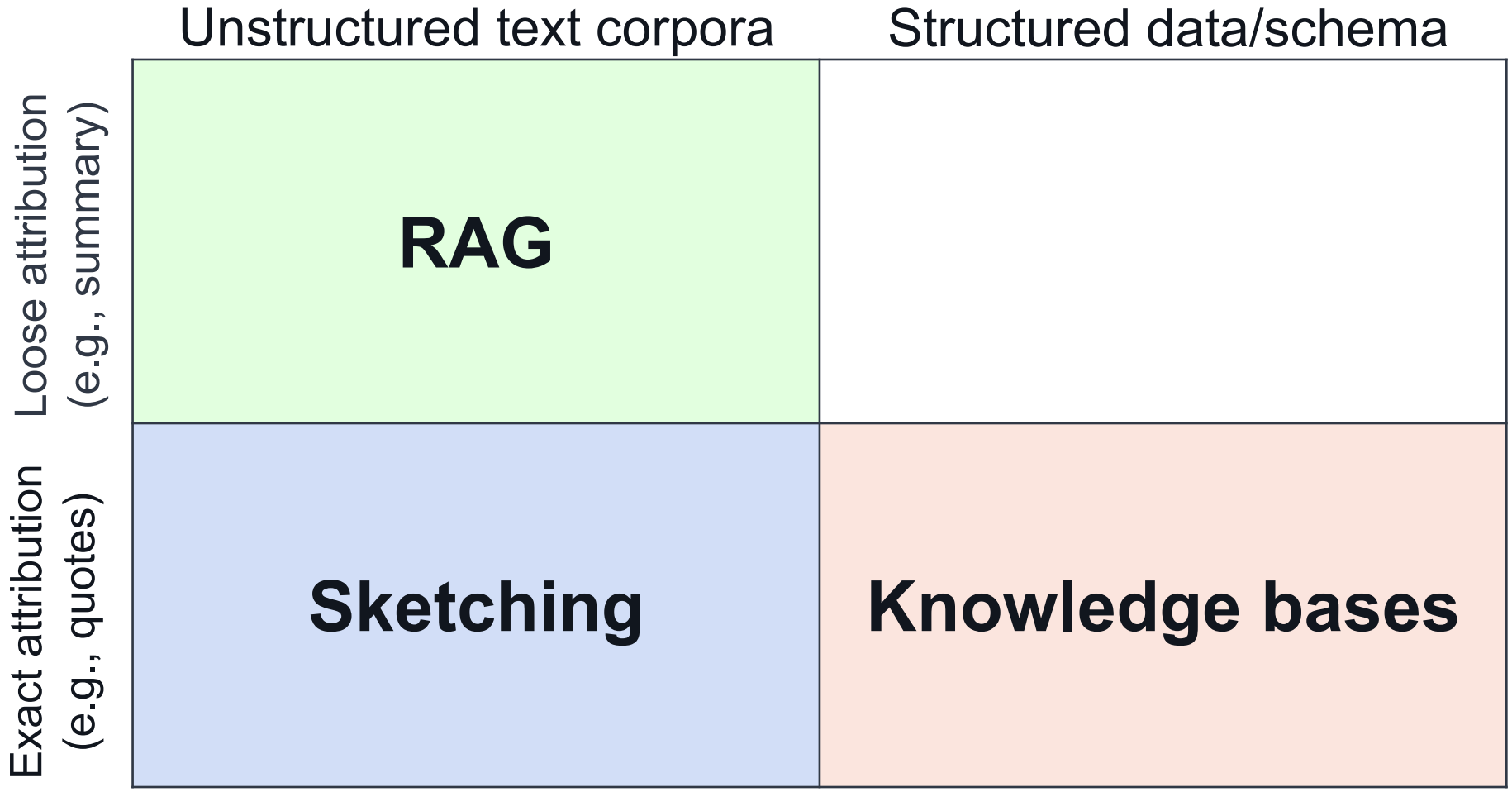
# Summary: Decision Tree of Paradigms



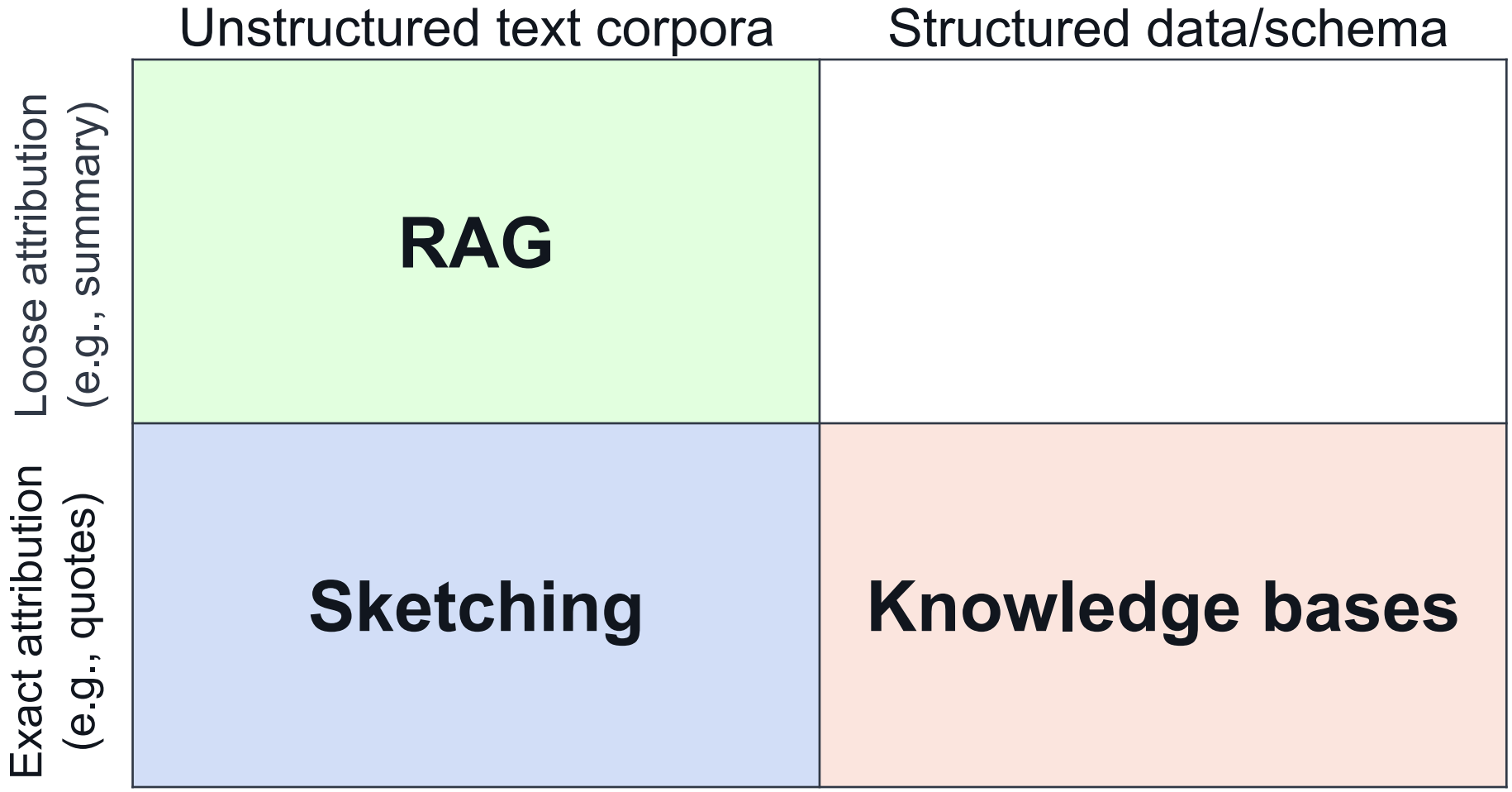
# Summary: Decision Tree of Paradigms



# Summary: Decision Tree of Paradigms



# Summary: Decision Tree of Paradigms




Match the tool to your archive!

# Closing Thoughts

# Closing Thoughts

- Attribution is a solvable problem, under specific conditions.

- Current tools work best for: English, well-digitized, large-scale.
- Humanities often has the opposite conditions.

- Potential source of impact: curating high-quality, digitized, multilingual corpora to to develop attribution tools on top of.