Retrieval-Augmented Generation with Multilingual Documents

Kevin Duh

Johns Hopkins University

Human Language Technology
Center of Excellence

Motivation: Writing Reports
(deep research)

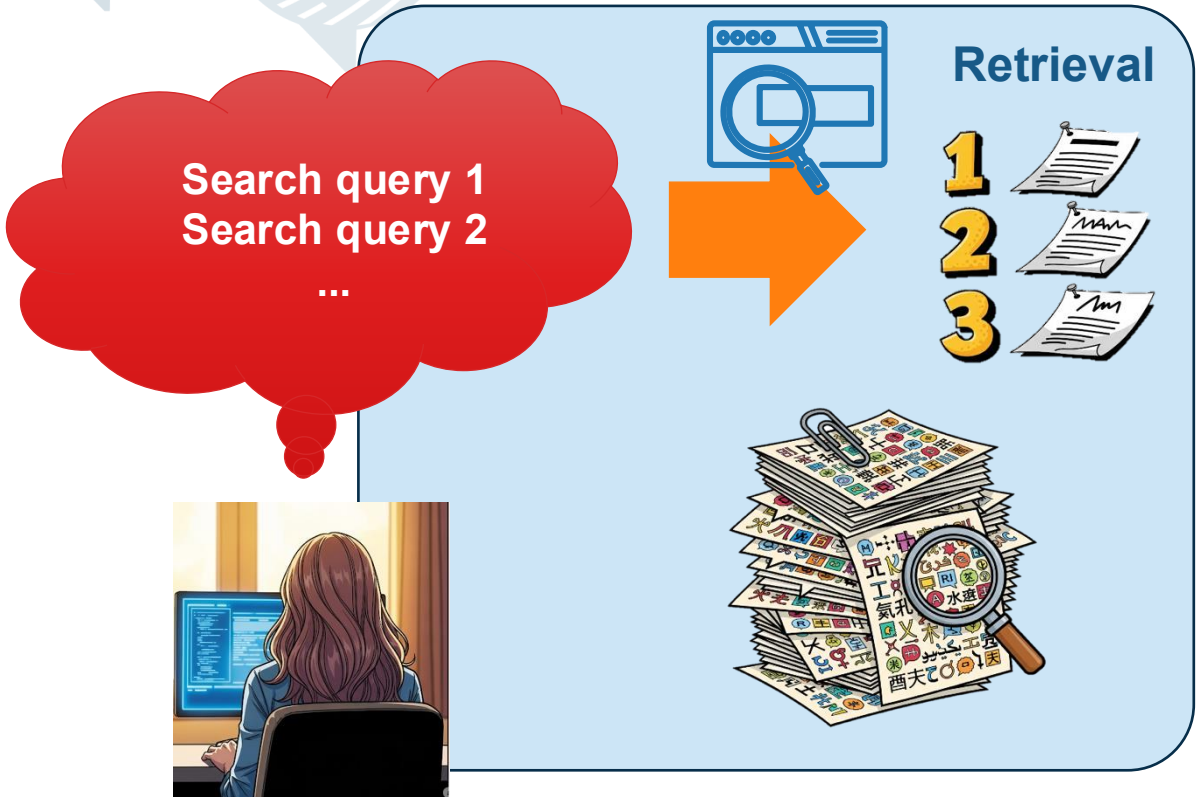
Multilingual documents

Report Request

Give me a report on tsunami preparedness efforts for countries across the Pacific in past decade.

Report Generation System

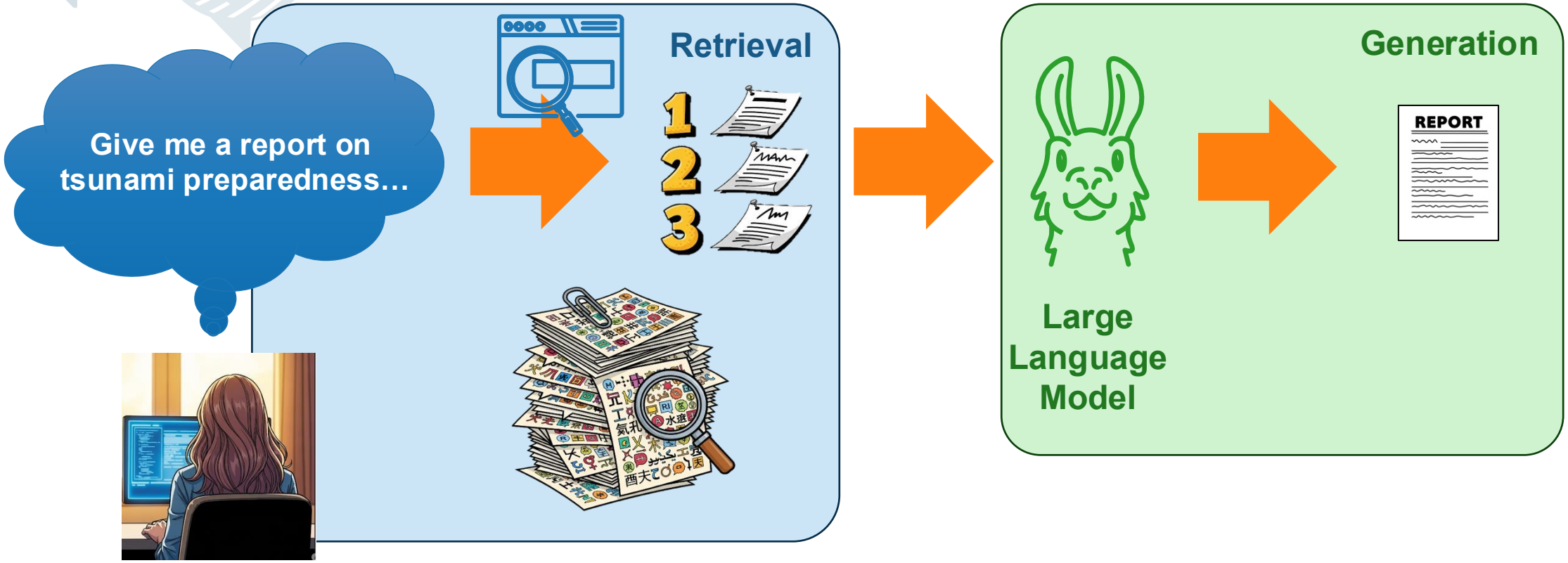
Manual process



Effort needed in:

- Formulating queries
- Reading documents
- Synthesizing info
- Writing

Retrieval-Augmented Generation (RAG)



Talk Outline

- Problem Formulation
  - Novel aspects
  - Evaluation metrics
- System Design
- Research Questions

Problem Formulation

Report Request

Background: As an archaeologist leading an expedition in South America, I require insights into the "Mysteries of Machu Picchu's Architecture" ...

Problem Statement: Produce a report on the mysteries of Machu Picchu's architecture. The focus of the report is on speculations and theories regarding the construction methods and architectural marvels of Machu Picchu….

Report Generation System

Generated Report

Machu Picchu is a city with polished stone walls, built in accordance with the classic architecture of the Incas, with the first building of this complex being "Intihuatana", the Sun Temple, a room with three windows and an area of Machu Picchu, which archaeologists call the Holy Zone [1]. The city was built on a ridge 2,430 meters above sea level, and its architecture has been puzzling researchers for a long time, with the buildings in the city being built of stones, but without concrete or any other fastening solutions [2]…

Complex information need

Report Request

Background: As an archaeologist leading an expedition in South America, I require insights into the "Mysteries of Machu Picchu's Architecture" ...

Problem Statement: Produce a report on the mysteries of Machu Picchu's architecture. The focus of the report is on speculations and theories regarding the construction methods and architectural marvels of Machu Picchu….

Multilingual collection

Report Generation System

Long-form output, with citations

Generated Report

Machu Picchu is a city with polished stone walls, built in accordance with the classic architecture of the Incas, with the first building of this complex being "Intihuatana", the Sun Temple, a room with three windows and an area of Machu Picchu, which archaeologists call the Holy Zone [1]. The city was built on a ridge 2,430 meters above sea level, and its architecture has been puzzling researchers for a long time, with the buildings in the city being built of stones, but without concrete or any other fastening solutions [2]…

Novel aspects about this problem

- Report Writing using RAG with Multilingual Docs
  - Complex info need
  - Long-form output
  - Multilingual collection

  } Combination leads to unique challenges
- Related problems:
  - Question Answering with RAG
  - RAG in languages other than English
  - Long-form RAG

Data Annotation

Report Request (Problem Statement): Produce a report on the mysteries of Machu Picchu's architecture. The focus of the report is on speculations and theories regarding the construction methods…

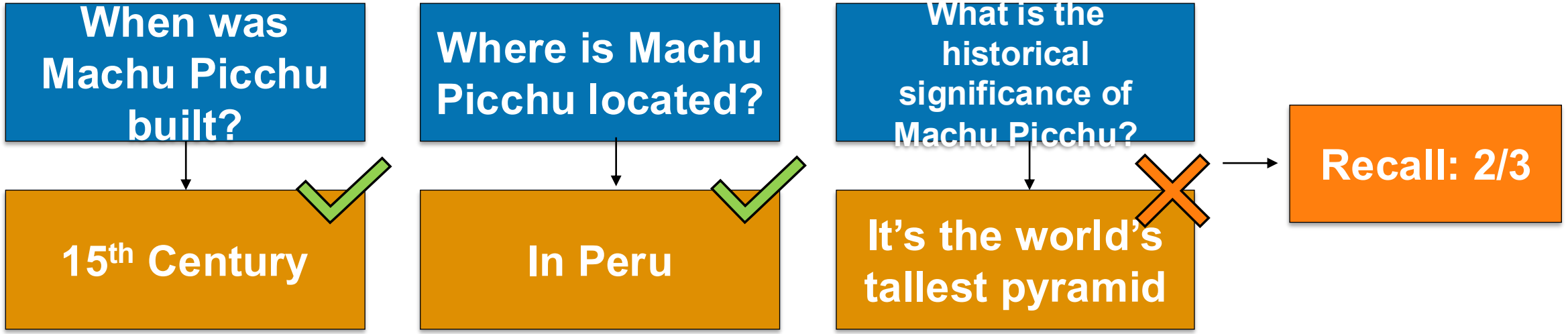
Annotated Nugget Questions/Answers:

Q1	When was Machu Picchu built?	OR	• 15th Century • 1440 • 1400s • in 1450 • around 1440
Q2	How did the fault lines beneath the site of Machu Picchu facilitated stone construction?	AND	• breaking gigantic granite stones into manageable pieces • By providing ample manageable building materials • providing a natural source of water • served as a drainage system

Our evaluation metric: ARGUE

1. How well do reports **recover relevant information**?

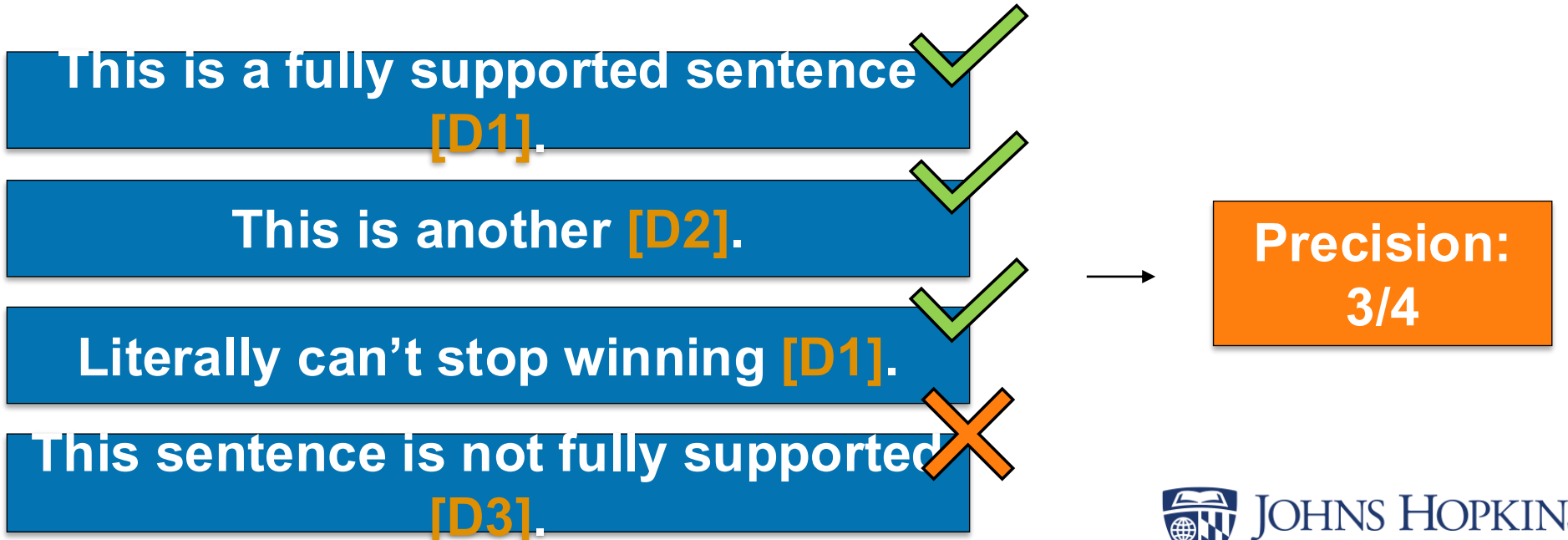
→ **Nugget Recall**: proportion of nugget questions correctly answered by the report



Our evaluation metric: ARGUE

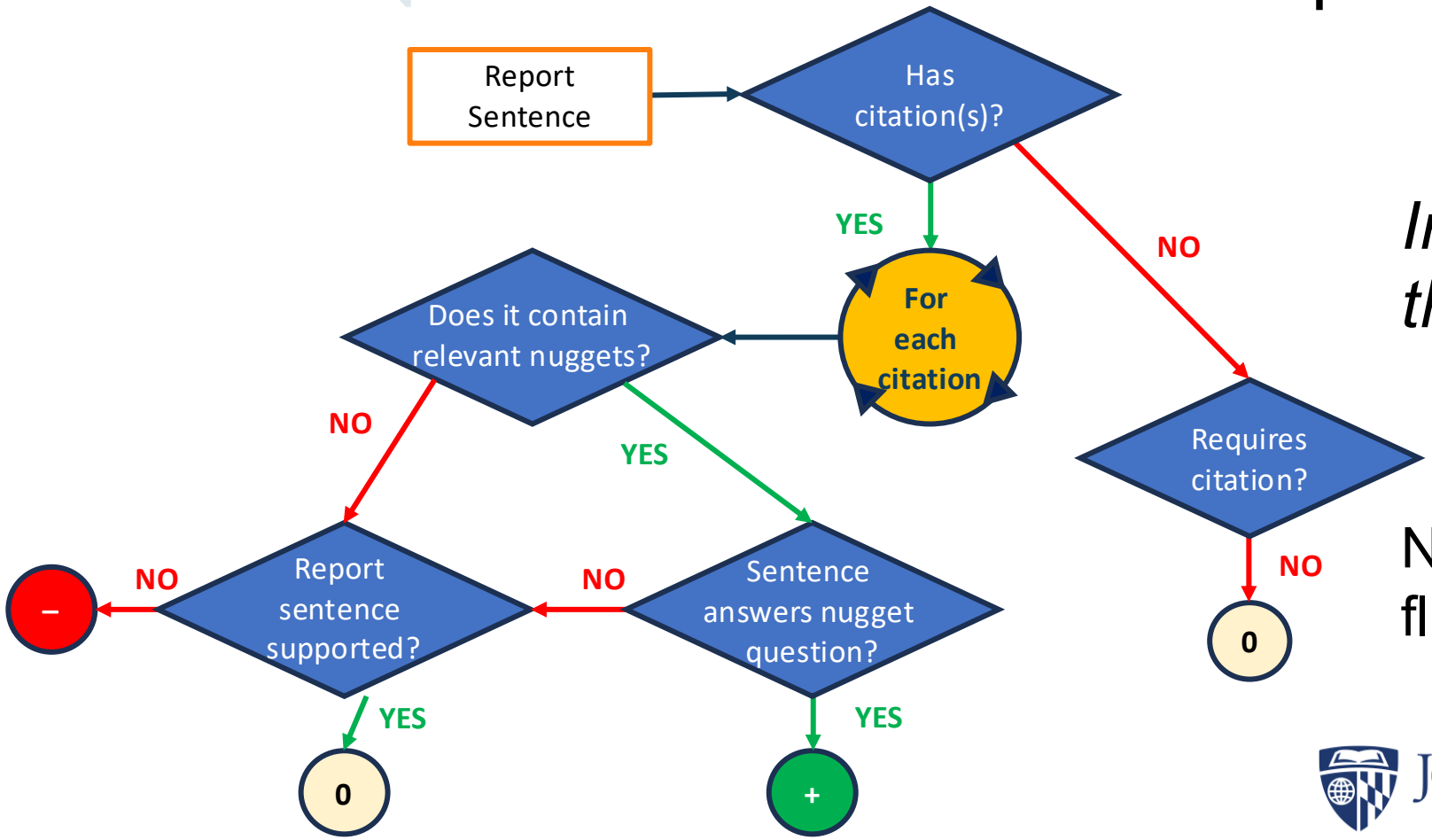
2. How well do reports **provide appropriate citation**?

→ **Citation Precision**: proportion of sentences fully supported by their cited documents



Our evaluation metric: ARGUE

ARGUE F1 score: combines citation precision & nugget recall



In practice, we compute these using LLM-as-judge

Note: ARGUE does not assess fluency, coherence, redundancy

Johns Hopkins University | Human Language Technology Center of Excellence

See: Mayfield et al. (2024) On The Evaluation Of Machine-Generated Reports

Data: TREC NeuCLIR 2024

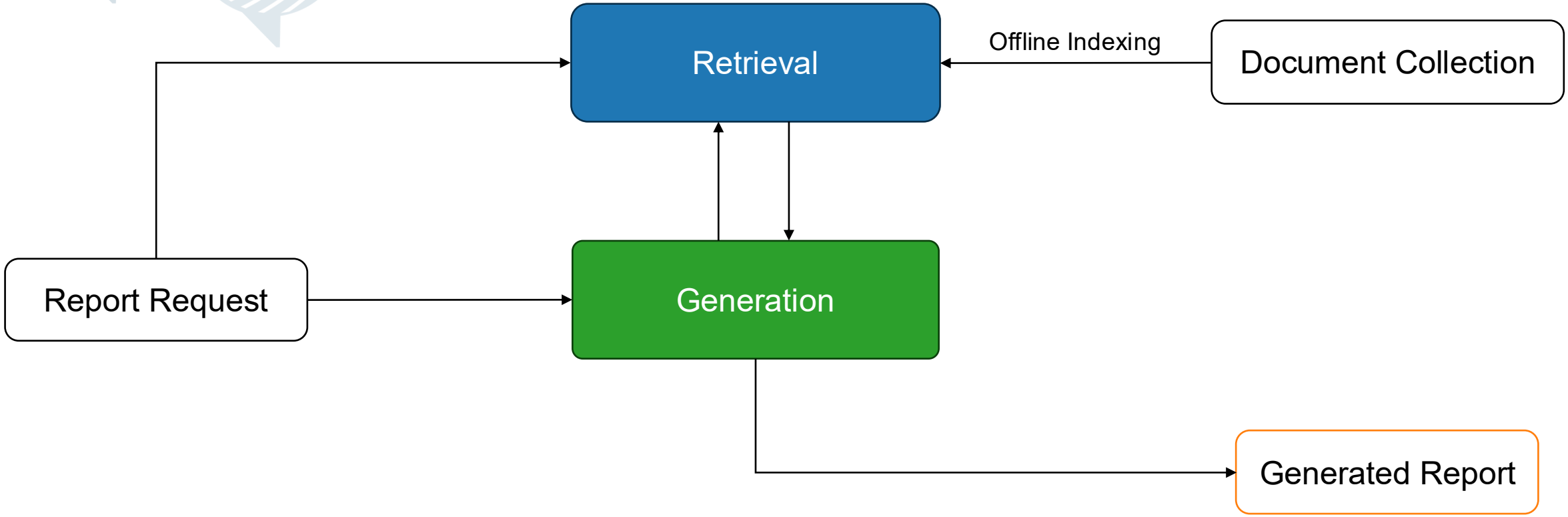
- Report requests in English
  - 3 for dev
  - 19 for test
- Multilingual document collection:
  - Chinese: 3 million docs
  - Farsi: 2 million docs
  - Russian: 5 million docs
  - from Common Crawl News spanning 2016-2021
- Manually created nuggets
- (Other datasets: TREC RAGTIME 2025, etc.)

See: Lawrie et al. (2024) Overview Of The TREC 2024 NeuCLIR Track

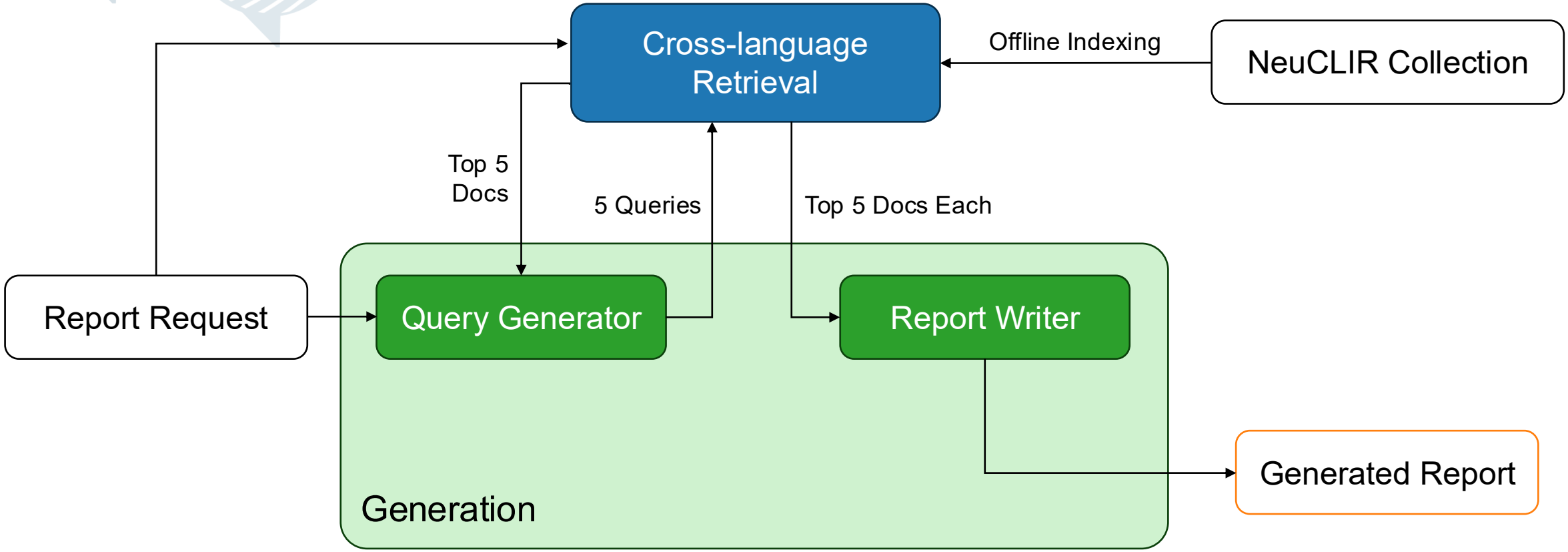
Talk Outline

- Problem Formulation
- System Design
  - Baseline System
  - Details about Retrieval Component
- Research Questions

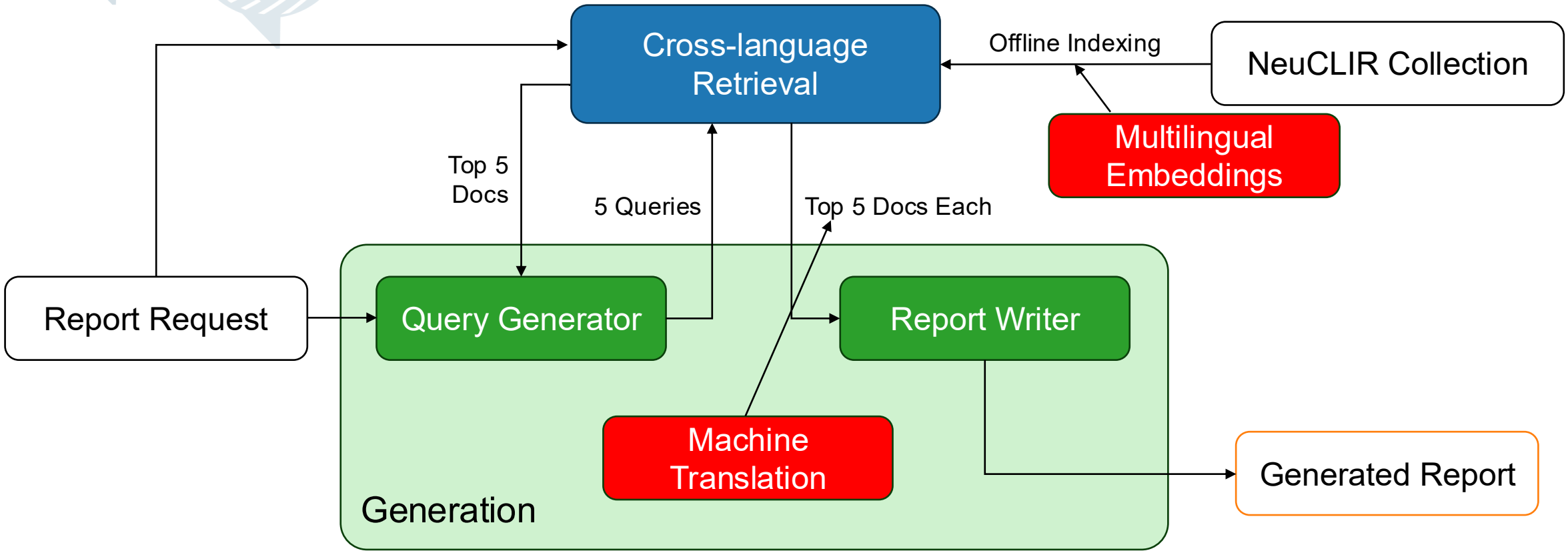
Overall System Design



Baseline System

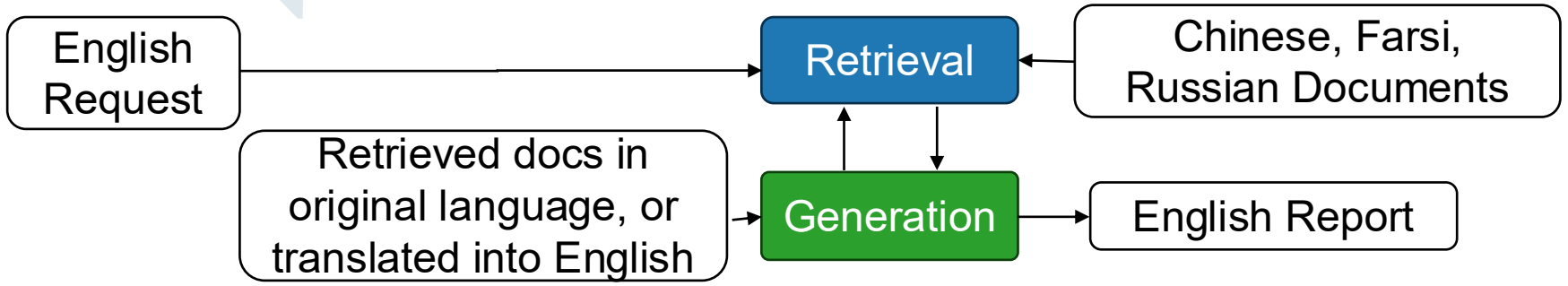


Where do we cross the language barrier?

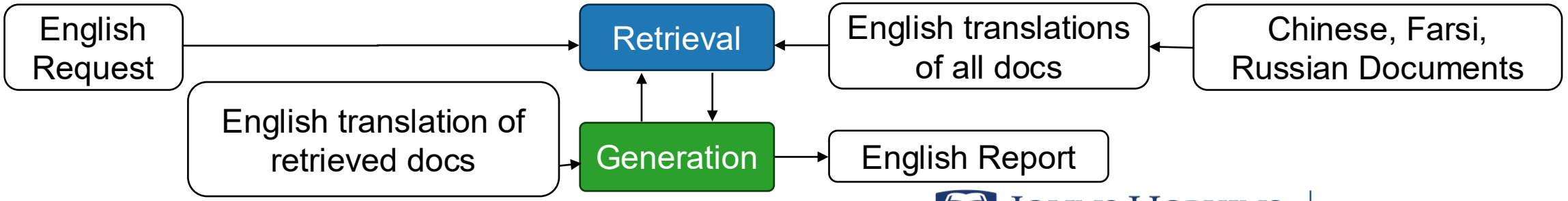


Alternative System Designs (1)

Our baseline system:

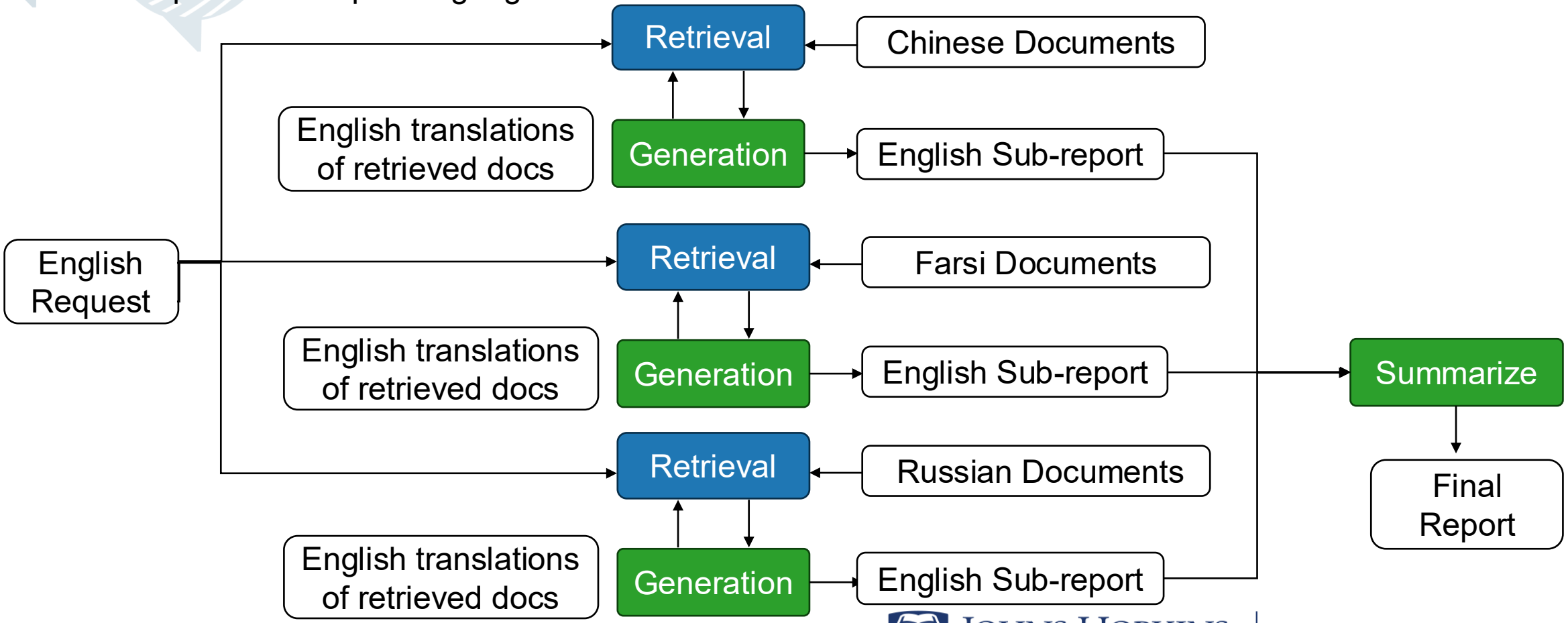


Alternative 1: Machine translation of everything first, map into English monolingual RAG problem



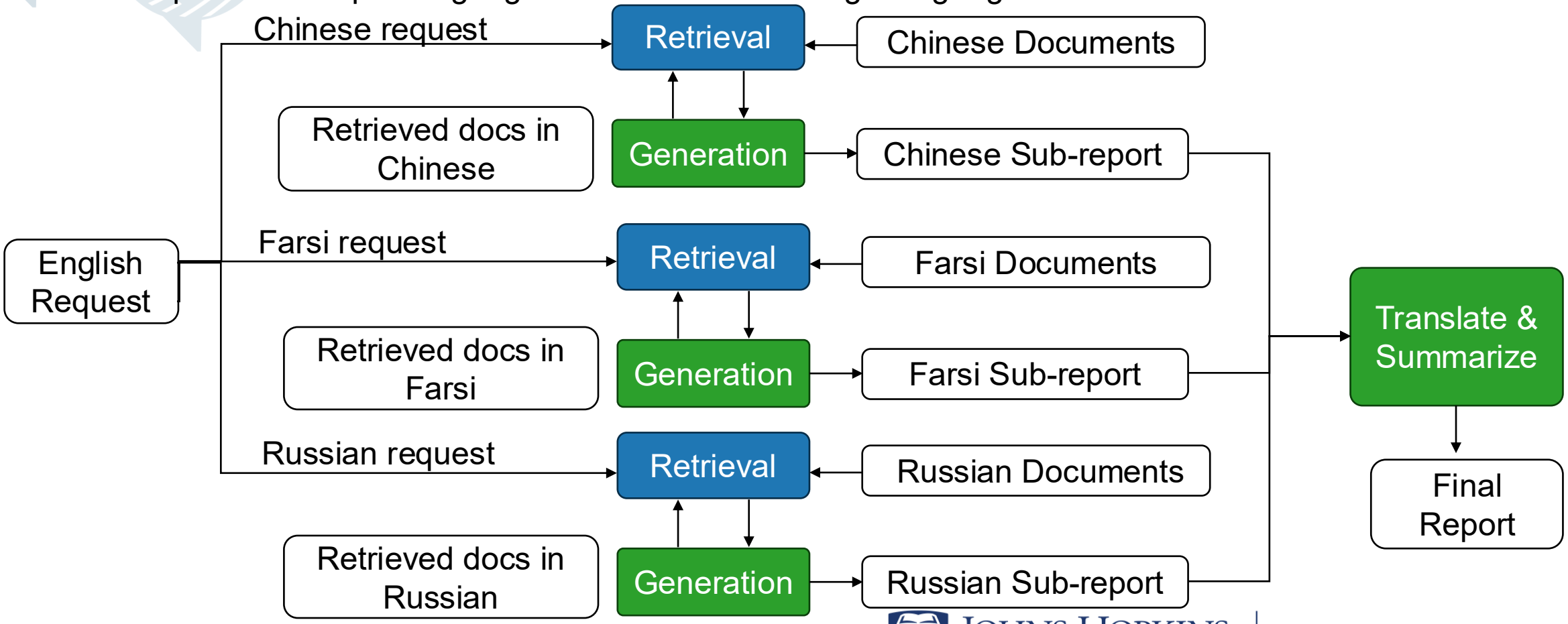
Alternative System Designs (2)

Alternative 2: Separate RAG per language



Alternative System Designs (3)

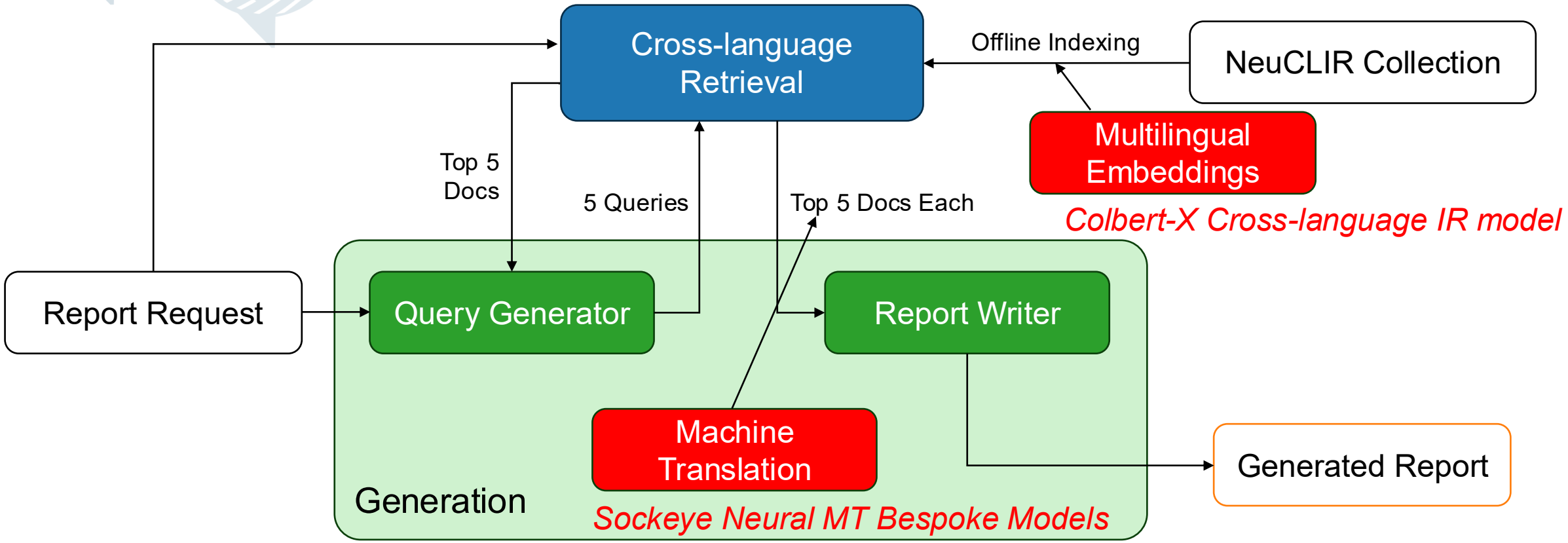
Alternative 3: Separate RAG per language AND do RAG in foreign language



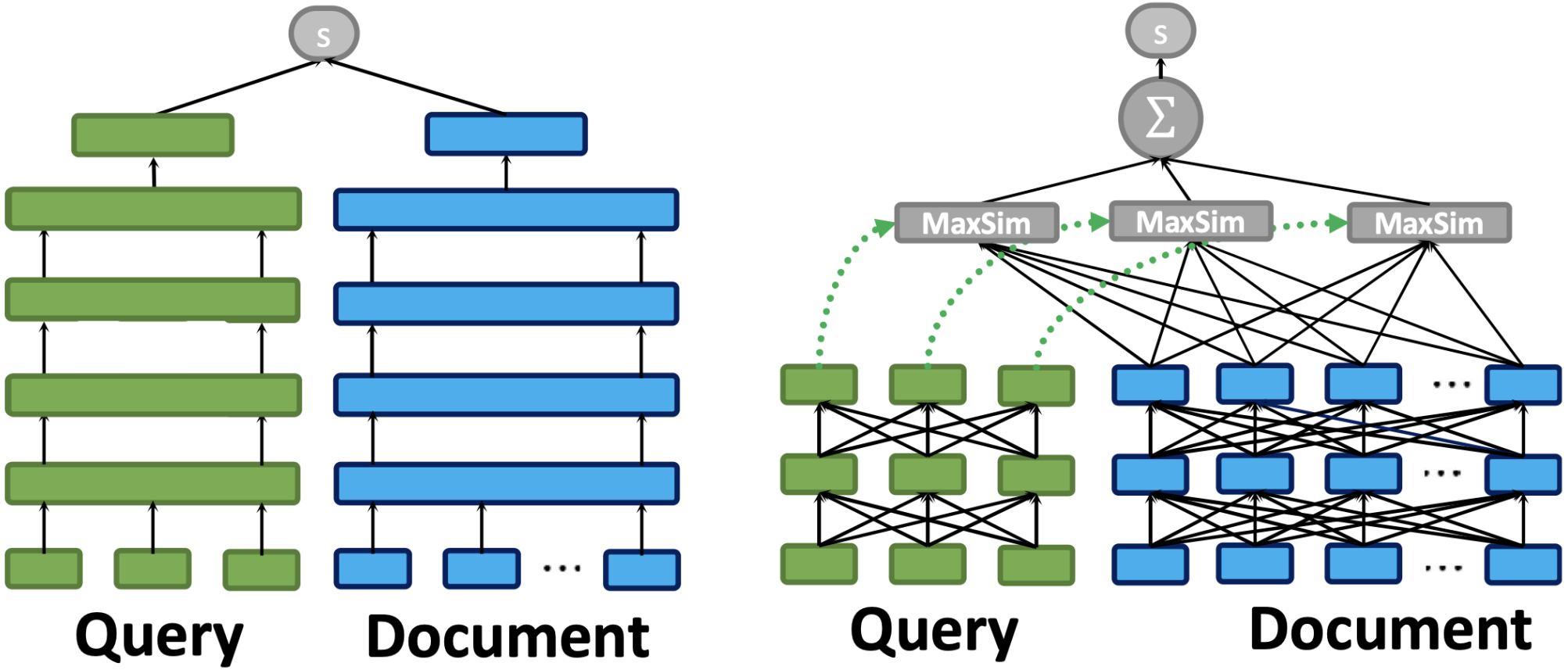
System Design Considerations

- Do we have strong retrieval component that can search across multiple languages at once?
- Do we have deployment constraints (index time, query time)?
- Do we trust LLMs to handle non-English documents?
- Do we have strong machine translation component?
- etc.

Where do we cross the language barrier?

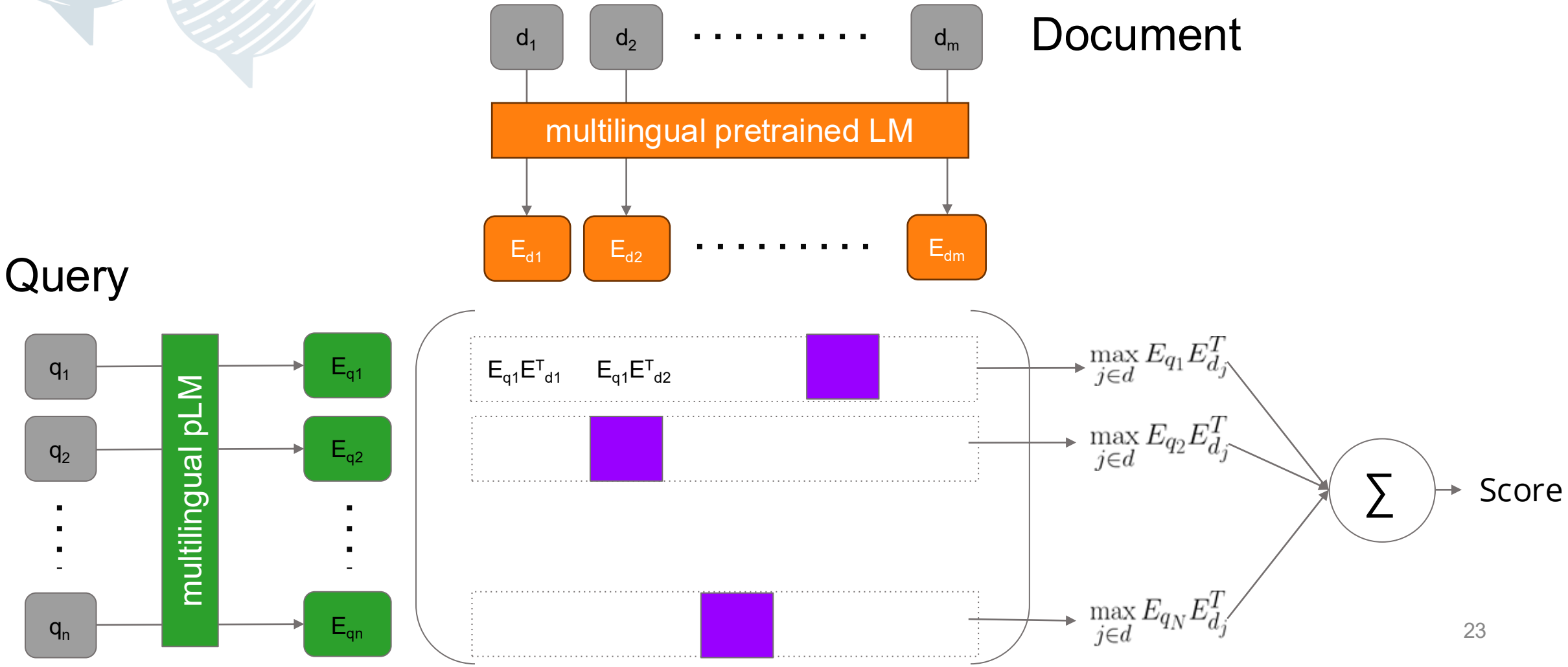


Embeddings & Neural Nets for Retrieval



From: Khattab (SIGIR2020) https://arxiv.org/pdf/2004.12832.pdf

Our cross-language retrieval model: ColBERT-X



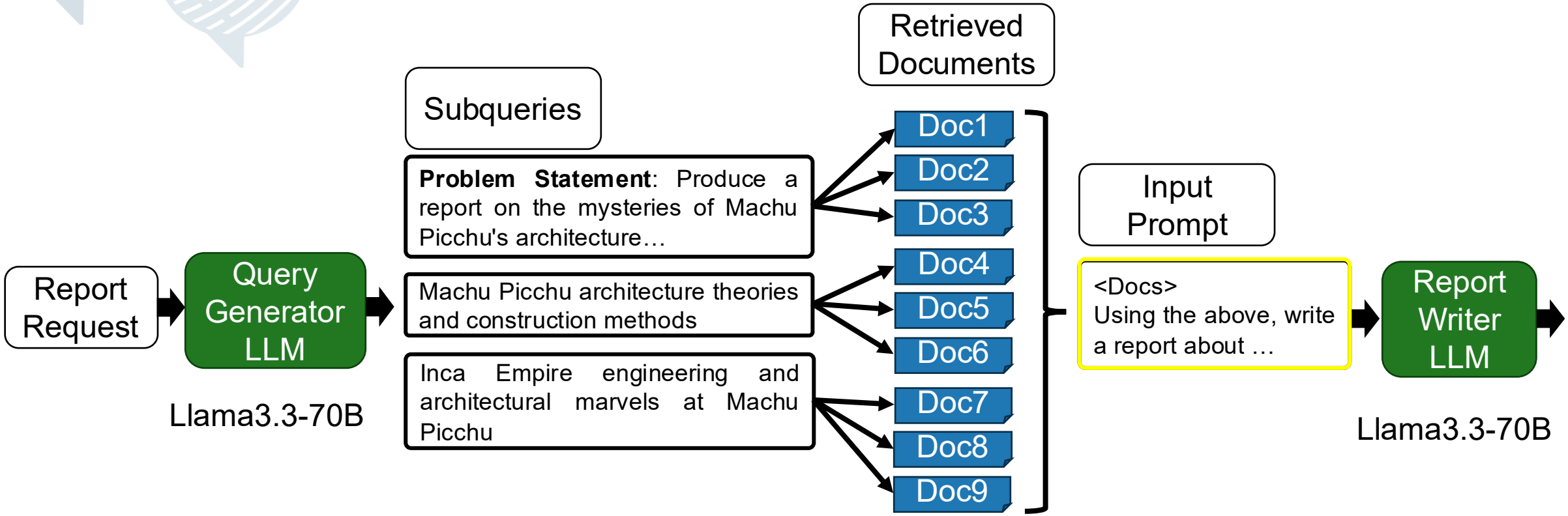
See: Yang et al. (2023) Neural Methods For Cross-Language Information Retrieval

Talk Outline

- Problem Formulation
- System Design
- Research questions
  - Document representation
  - Citation accuracy in reports
  - Agentic workflows

Research Question 1:
What is a good document representation in this multilingual RAG setting?

Baseline system (again): GPT-Researcher



See: Duh et al. (2025) HLTCOE At LiveRAG: GPT-Researcher Using ColBERT Retrieval

Prompt Template for Report Writer LLM

Information:
Source 1:
{document}

Source 2:
{document}

How should we represent the documents in the prompt?

Using the above information, provide supporting facts for the query: "**{problem statement}**" in a detailed report -- The report should focus on the answer to the query, should be well structured, informative,
and contain ALL facts and numbers in the supporting information in at most **{N}** words.

Please follow all of the following guidelines in your report:
- You MUST determine your own concrete and valid opinion based on the given information. Do NOT defer to general and meaningless conclusions.
- You MUST write the report in the following language: English with each sentence on its own line.
- You MUST cite your sources, especially for relevant sentences that answer the question.
- When using information that comes from the documents, use inline notation which refer to the Document ID (e.g. [1] for Source: 1).
- It is important to ensure that the Document ID is a valid string from the information above and that the information in the sentence is present in the document.
- Do not include a list of citations at the end of the document.
- Do not put citations in the middle of the sentence. Always cite at the end of the sentence.
- Write the report in a Objective (impartial and unbiased presentation of facts and findings) tone.
Please do your best, this is very important to my career.

How to represent documents in the prompt: which language?

1. Machine translation (MT) into English

Machu Picchu, a famous site in Peru, South America, has recently been claimed by geologists to solve the mystery of its construction, due to the large amount of stone produced by its fault layer…

2. Source language

南美洲祕魯的著名遺址馬丘比丘，近日有地質學家宣稱解開其建造之謎，由於其斷層產生了大量的石材，讓古印加帝國的人能夠取得建築原料…

How to represent documents in the prompt: how much information?

1. Keep all documents (as is)

Machu Picchu, a famous site in Peru, South America, has recently been claimed by geologists to solve the mystery of its construction, due to the large amount of stone produced by its fault layer…

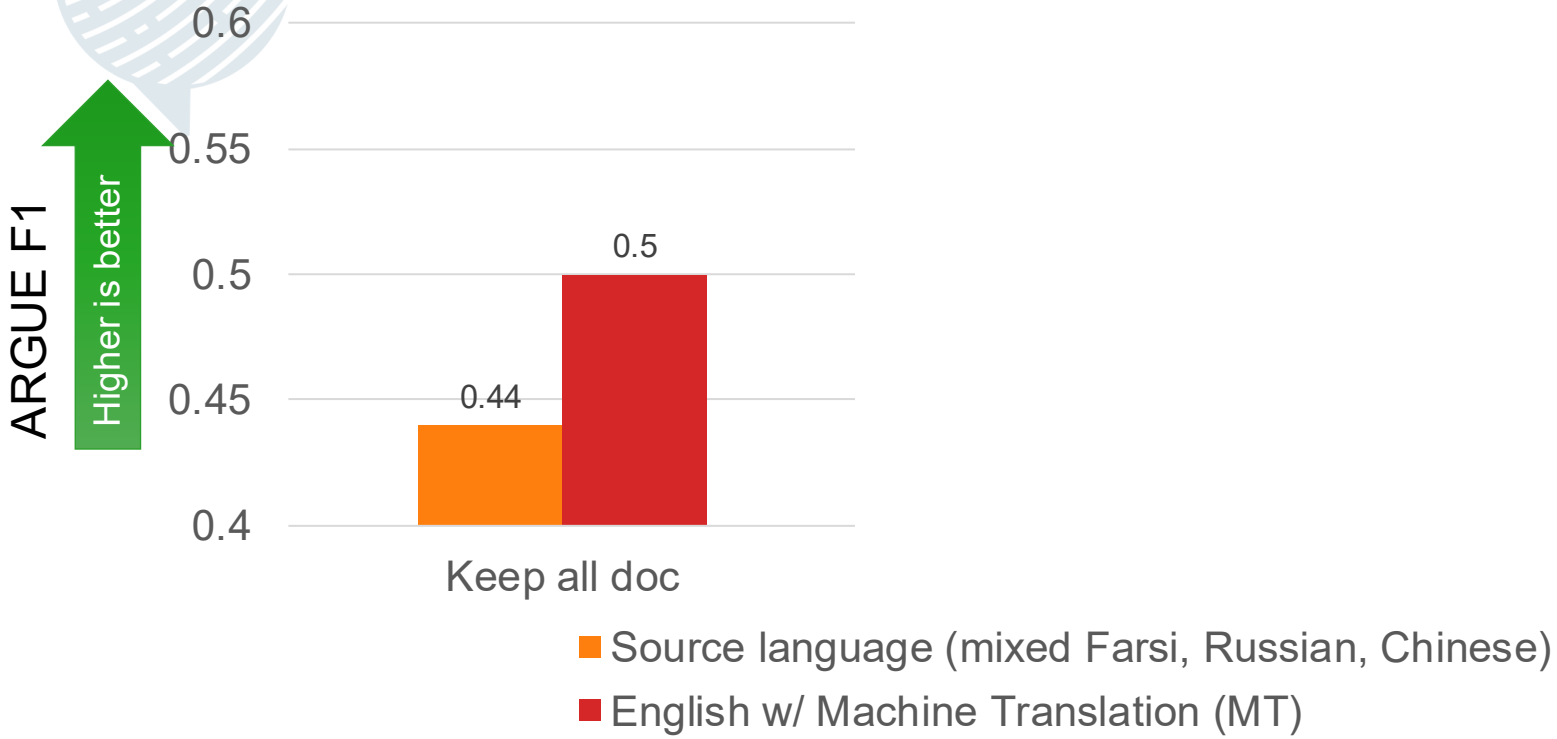
2. Snippets with Embedding filter

[of its construction, due to the large amount of stone]

3. Notetaking: ask another LLM to extract facts

1. Machu Picchu constructed using local stone from fault layer.
2. Fault layer made it easy to break stone into same plane.
3. Geologist Rualdo Menegat mapped structure using satellite photos.

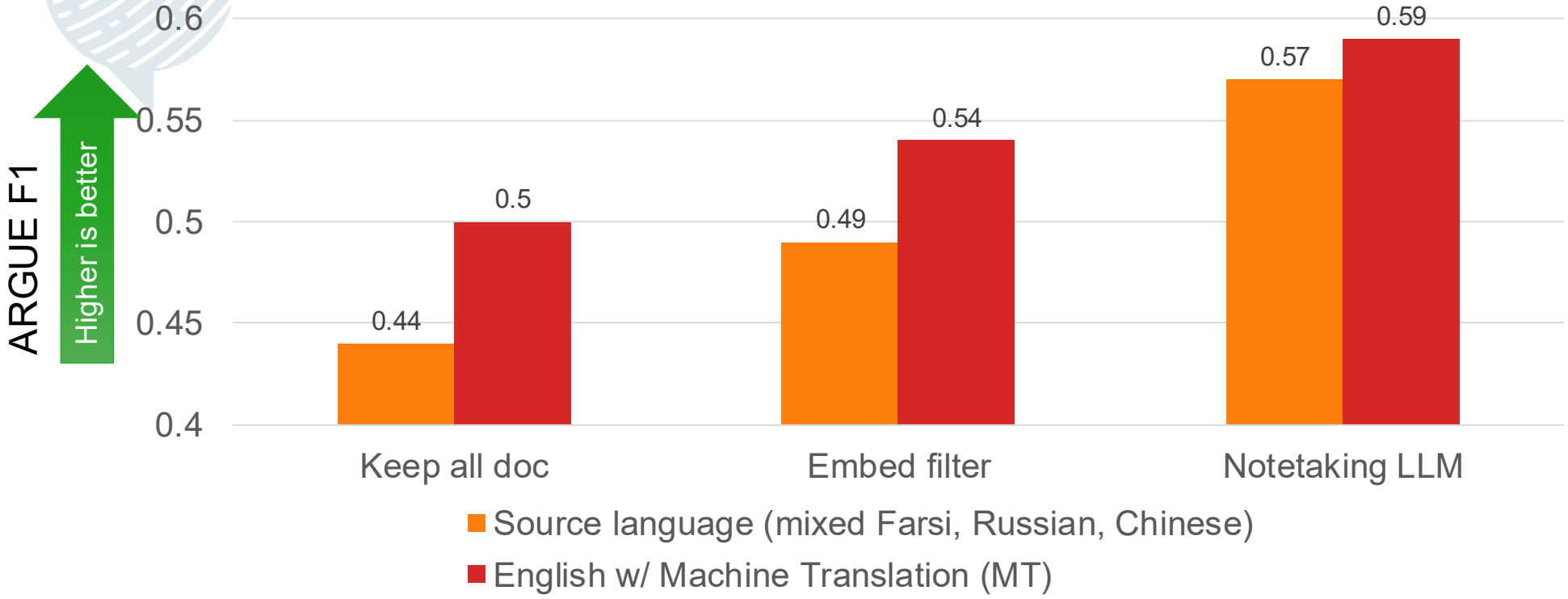
Report quality: ARGUE F1



Findings:

- Documents in English MT better than mixed source language: strong MT is useful

Report quality: ARGUE F1



Findings:

- Documents in English MT better than mixed source language: strong MT is useful
- Notetaking > Embedding > Keep all: more focused information is useful

Research Question 2:
Does multilinguality impact citation accuracy in reports?

Example report

Machu Picchu's architecture is a subject of fascination, with many theories regarding its construction methods [1]. The site was built around 1140 by Pachacutec [2]. The Incas chose a location along a 175km long fault line, which provided abundant stone [3].

Example report*

Machu Picchu's architecture is a subject of fascination, with many theories regarding its construction methods [1]. The site was built around 1140* by Pachacutec [2*]. The Incas chose a location along a 175km long fault line, which provided abundant stone [3].

FAKE! The presence of citation hallucination is dangerous.
It may give the user a false sense of trust.
→ system should flag citations that do not attest the sentence.

Does multilinguality impact citations?

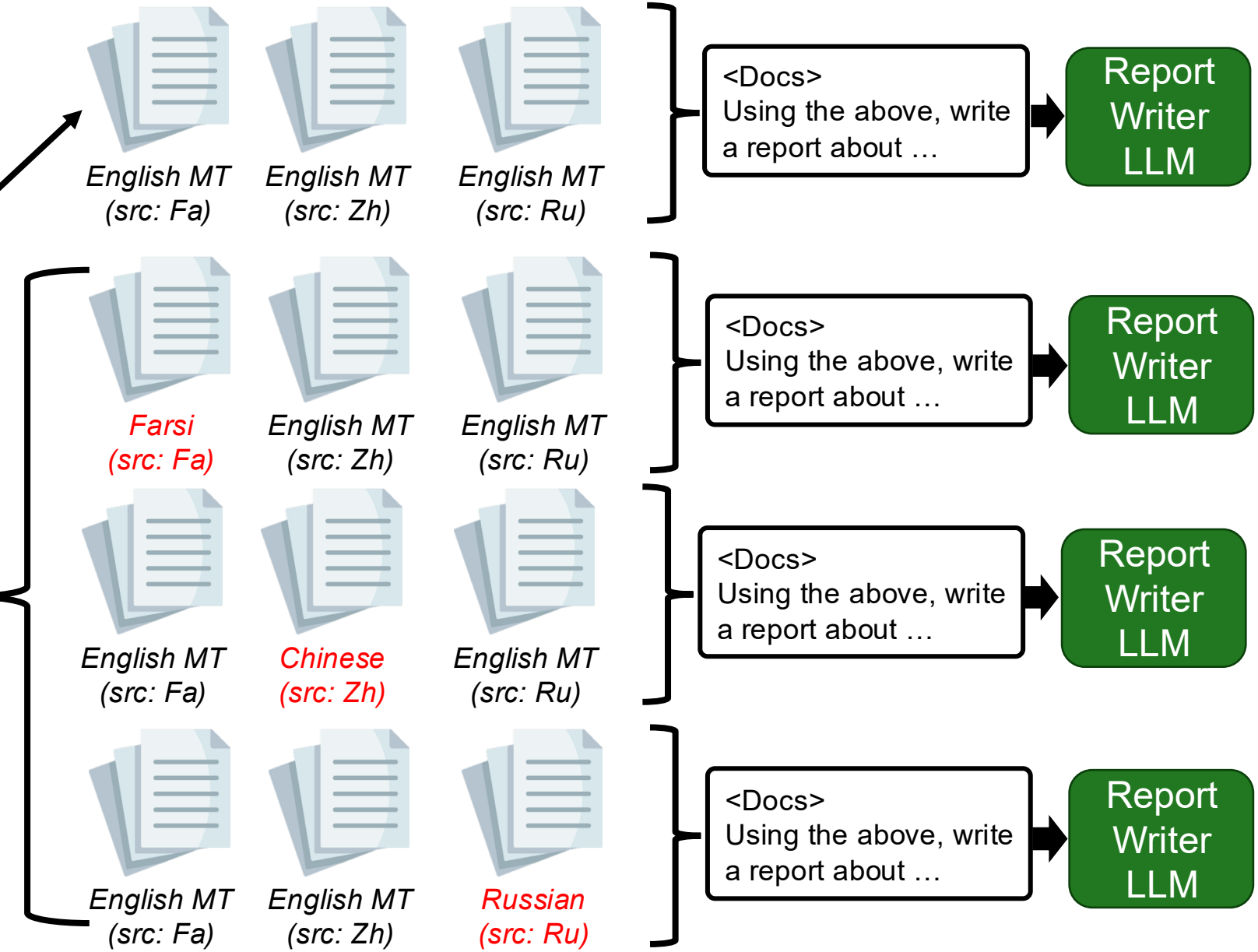
- Citation hallucination is a general problem
- But are there additional complications when using a multilingual collection in RAG?

Experiment Setup: NeuCLIR Report Writing

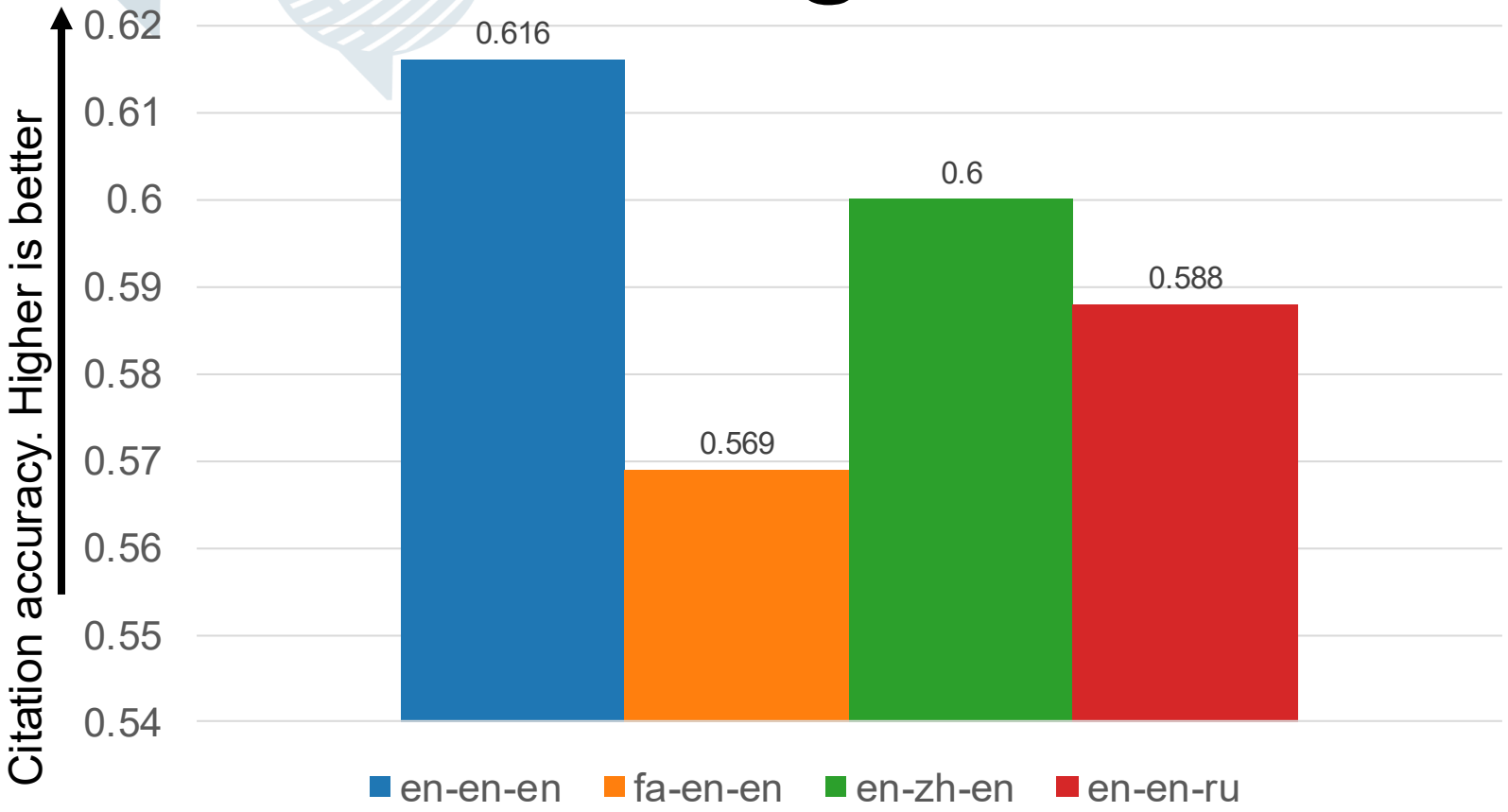
1. Use 3 relevant English MT documents in context

2. Swap out using one document with source

3. Measure citation accuracy: does it drop vs. all-English?



A: Yes. LLM shows preference to cite English documents



Findings:

- When English document is replaced with source, Llama3.3-70B cites it less frequently
- Full story is a bit more complicated…

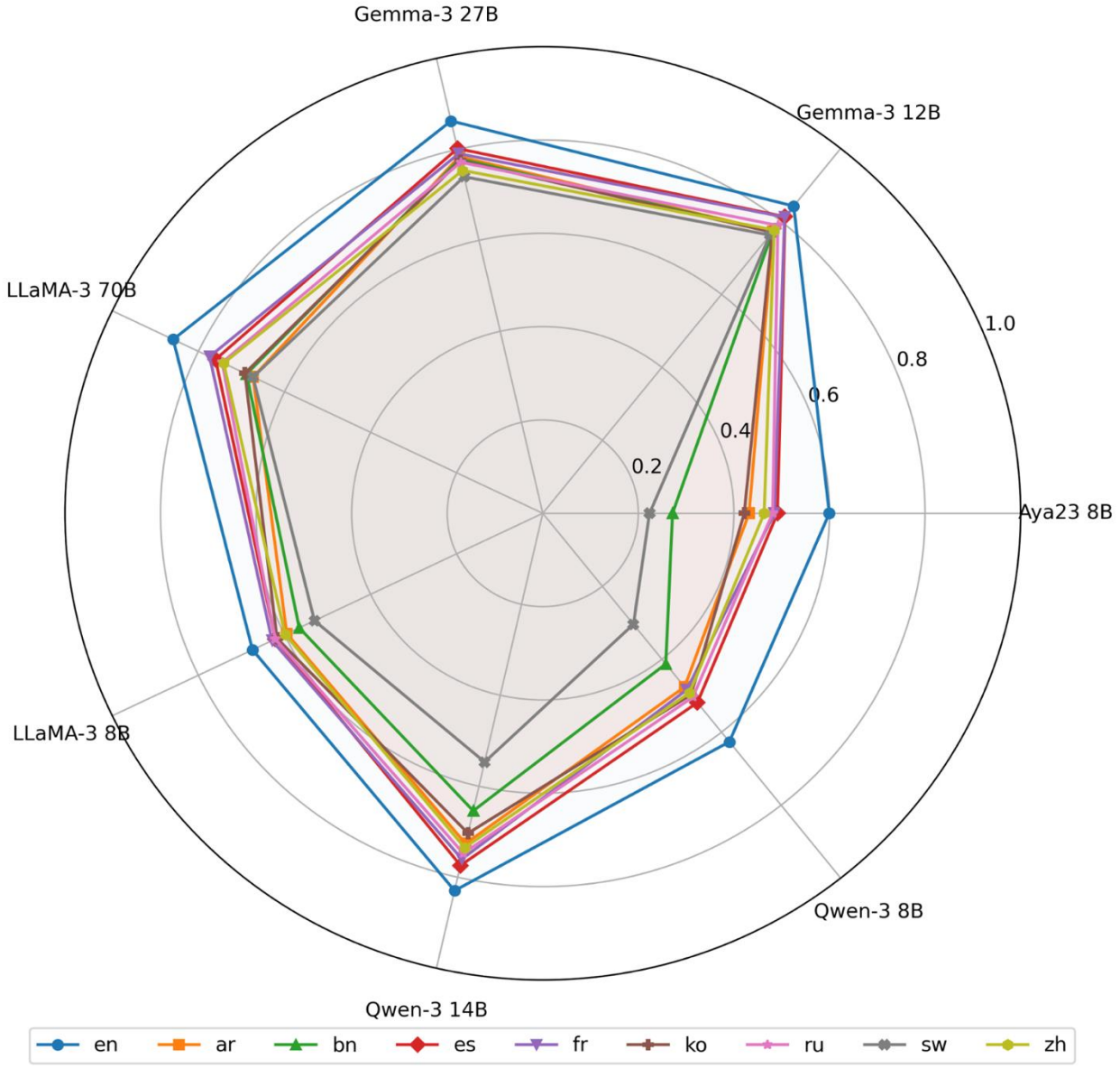
Additional Experiments

- Controlled simulations to further understand the phenomenon
  - Previously, ran LLM on different contexts, generate full report and measured aggregated accuracy
  - Simulation: also run LLM on different contexts, but focus on just the citation token by force decoding and compare accuracy

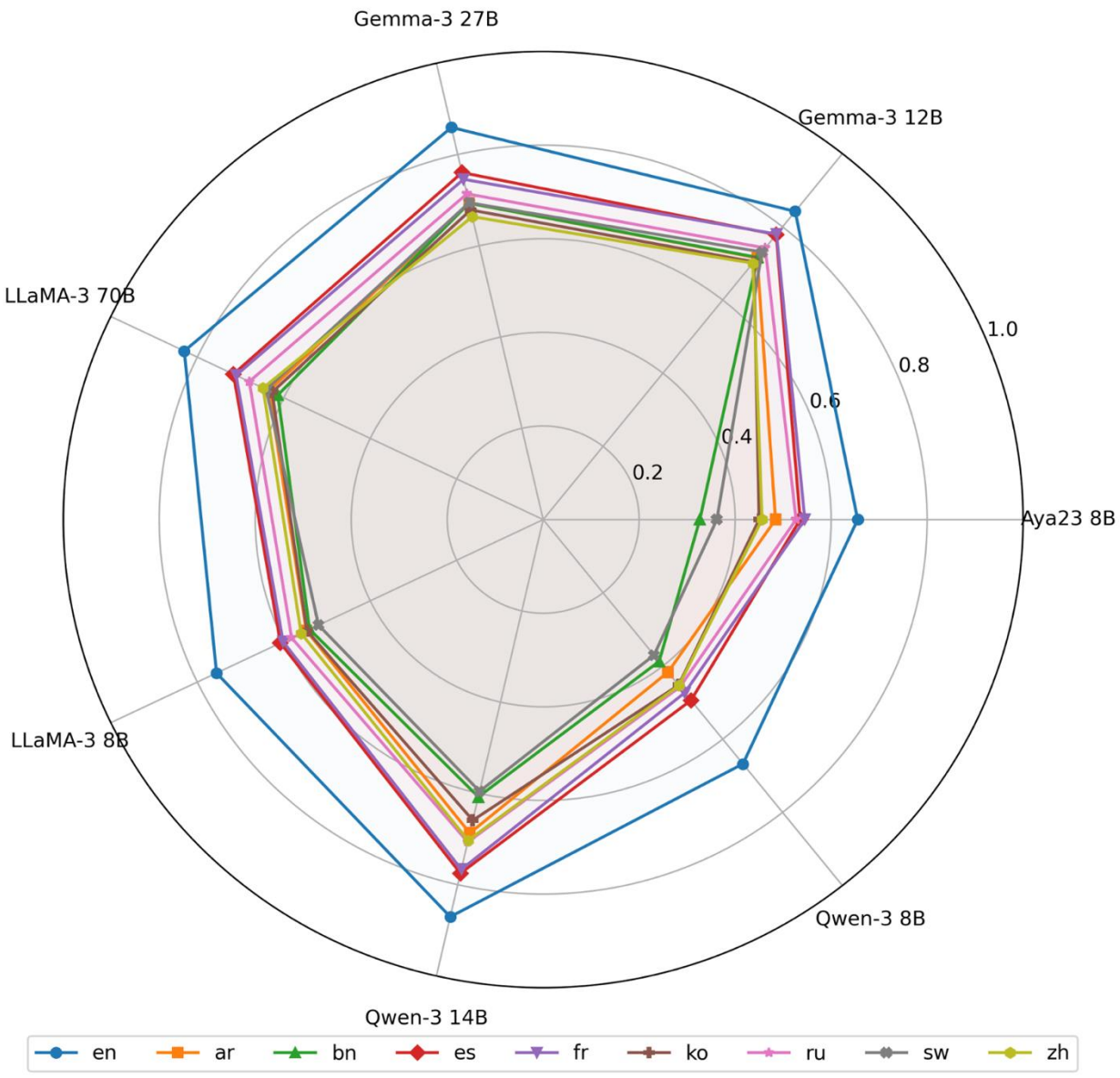
- Different datasets, more LLMs, reverse language directions

See: Ki et al. (2025) Linguistic Nepotism: Trading-Off Quality For Language Preference In Multilingual RAG

Citation Accuracy



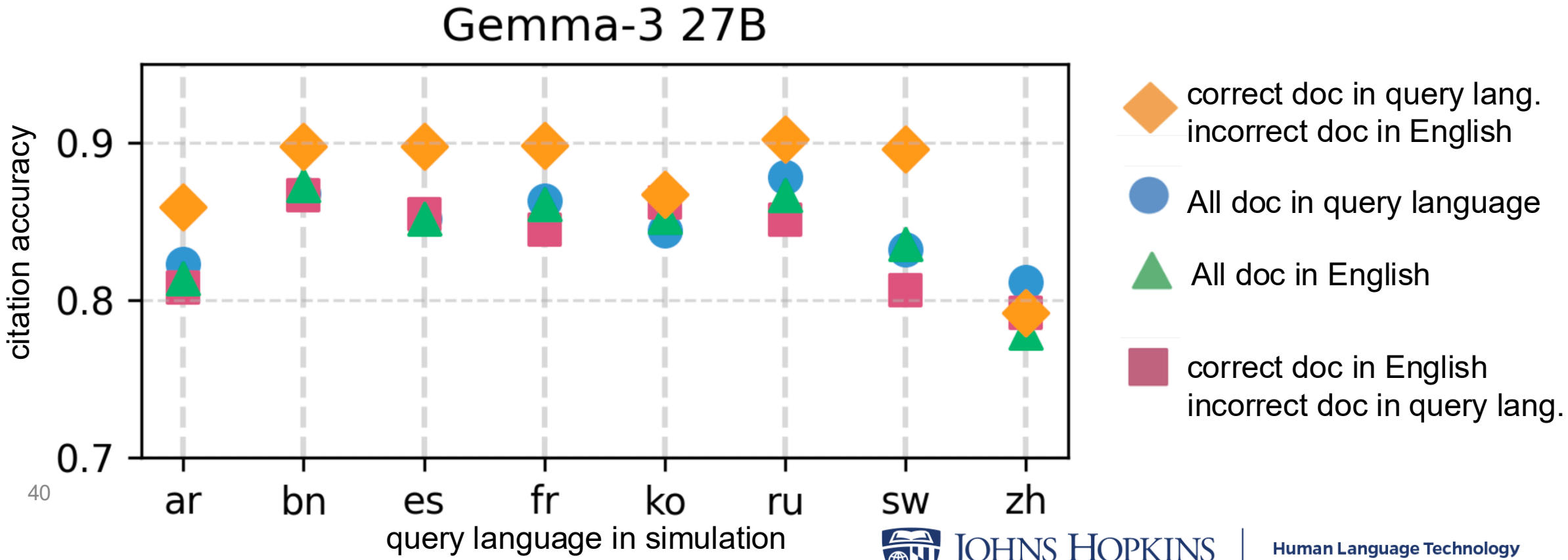
Dataset: ELI5 *(Explain Like I'm Five)*



Dataset: MIRACL

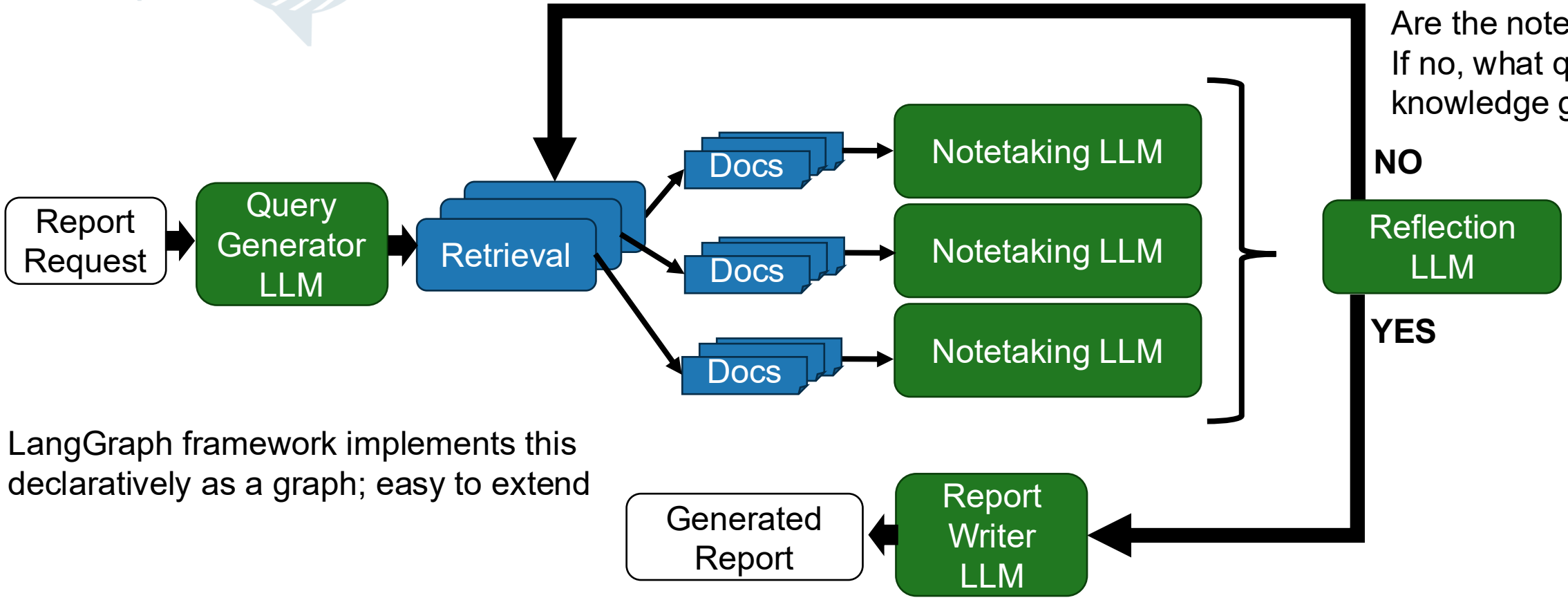
What if: reverse language direction

Query in non-English → LLMs prefer citing docs in query lang.



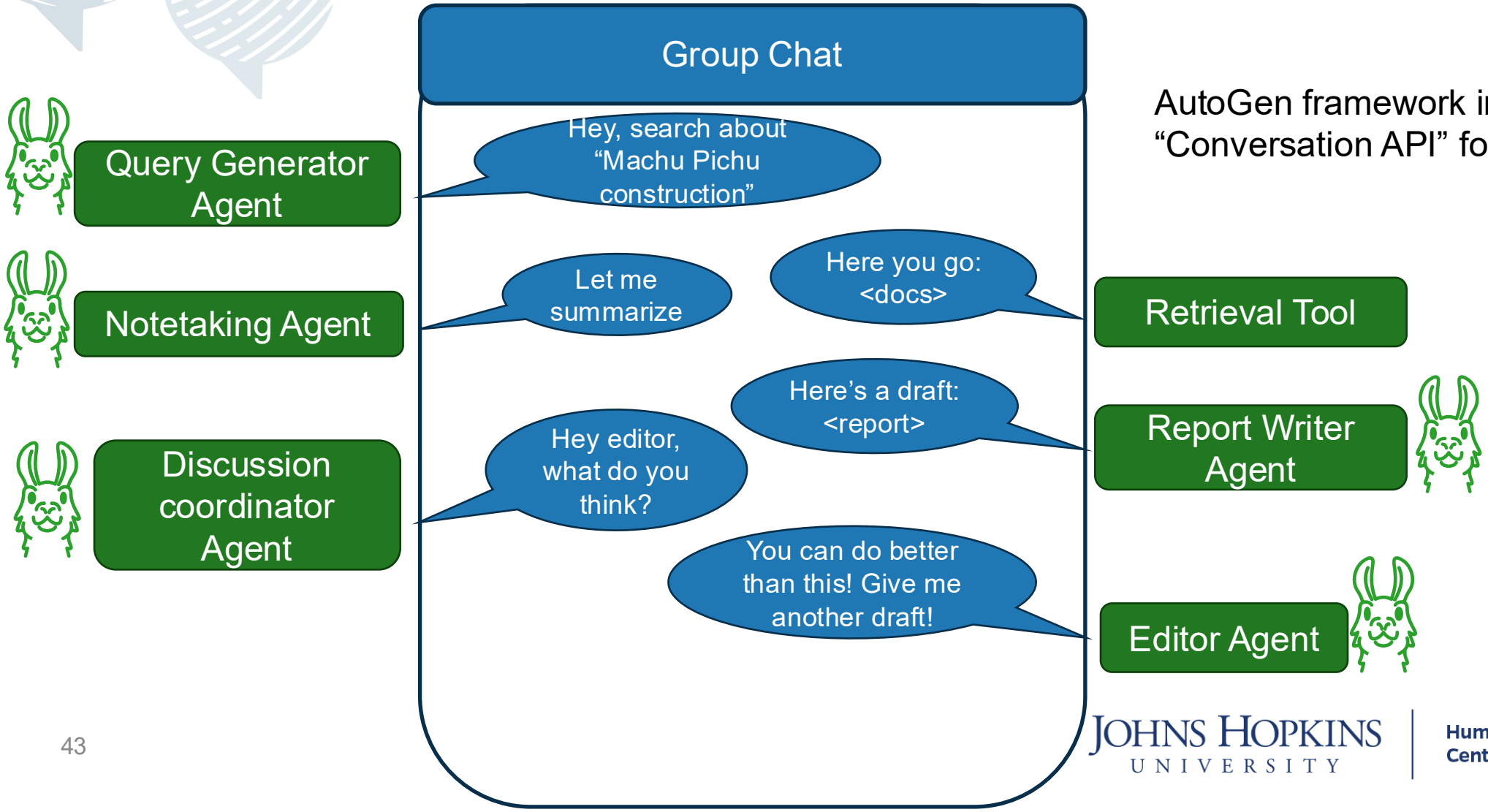
Research Question 3:
Can agentic workflows improve report writing?

LangGraph Researcher



LangGraph framework implements this declaratively as a graph; easy to extend

AutoGen Researcher



AutoGen framework implements a "Conversation API" for agents

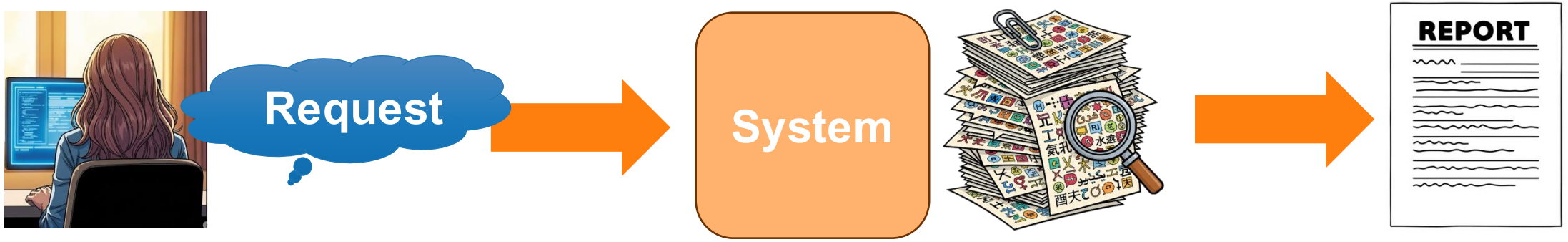
Comparison of frameworks

	ARGUE F1	# LLM calls	Estimated cost per report
GPT-Researcher (with notetaking)	0.59	2 + #Query x #Doc/Query	5 cents
LangGraph Researcher	0.62	#QueryOnAvg x Doc/Query x MaxReflectionLoop	7 - 120 cents
AutoGen Researcher	0.69	Potentially unbounded, in experiments ~24 calls with long prompts	6 - 28 cents

Note: Not a fair comparison due to different hyperparameters & prompts
Agentic frameworks are flexible and can improve results, but need to wary of cost

Conclusion

- RAG with multilingual docs: a novel and challenging problem
  - System design: How to combine IR, MT, LLM
  - How document representation impacts RAG
  - How multilinguality impacts citation accuracy
  - How agentic workflows might be helpful
- Generally, report writing (deep research) is an exciting application
  - Potential for more human-AI collaboration



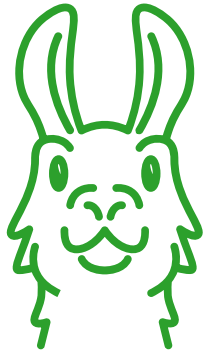
Acknowledgment: SCALE 2025 Summer Workshop Teams

This talk is a summary of a large effort involving three teams of researchers.
At Johns Hopkins University, we have a tradition of summer workshops (hackathons) that bring together researchers to work on a problem together. It's intense and rewarding!

Retrieval

Effective Retrieval for Report Generation

Generation



Faithful and Effective Generation

Evaluation

Reliable and Informative Evaluation

Query How does fingerprint unlock work on phones?

❶ How Touch ID works: To make sense of Apple's fingerprint identity sensor [...]

❷ Fingerprint scanners use arrays of tiny capacitor circuits to collect data [...]

❸ Digital data is analyzed to look for distinctive fingerprint attributes [...]

Relevant K documents

Step 1 : Translate each relevant document

❶ How Touch ID works: To make sense of Apple's fingerprint identity sensor [...] en

⋮

❶ Comment fonctionne Touch ID· Pour fr

❶ كيف يعمل Touch ID· لفهم مستشعر هوية بصمة ar

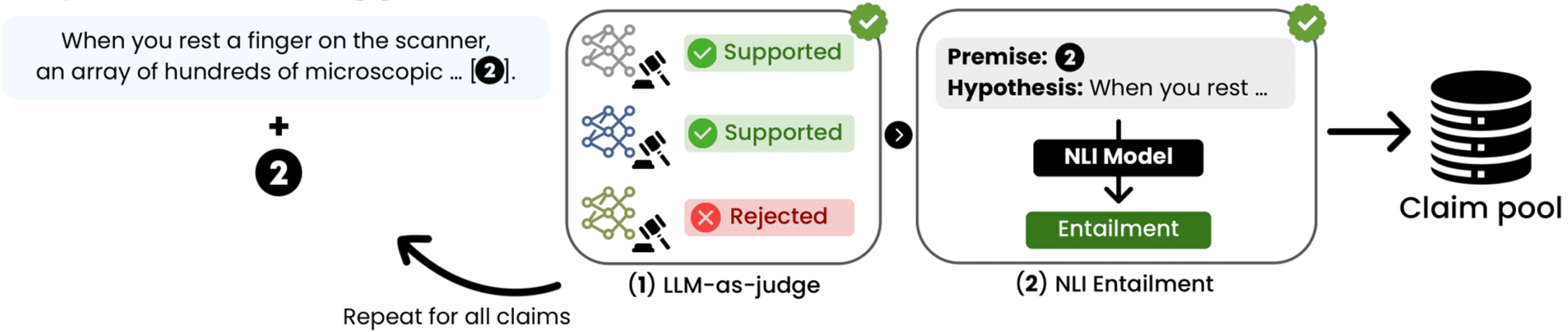
❶ Touch ID 작동 원리 : Apple의 지문 인식 센서를 이해하기 위해서는 [...] ko

Step 2 : Generate *gold* evidence-supported response

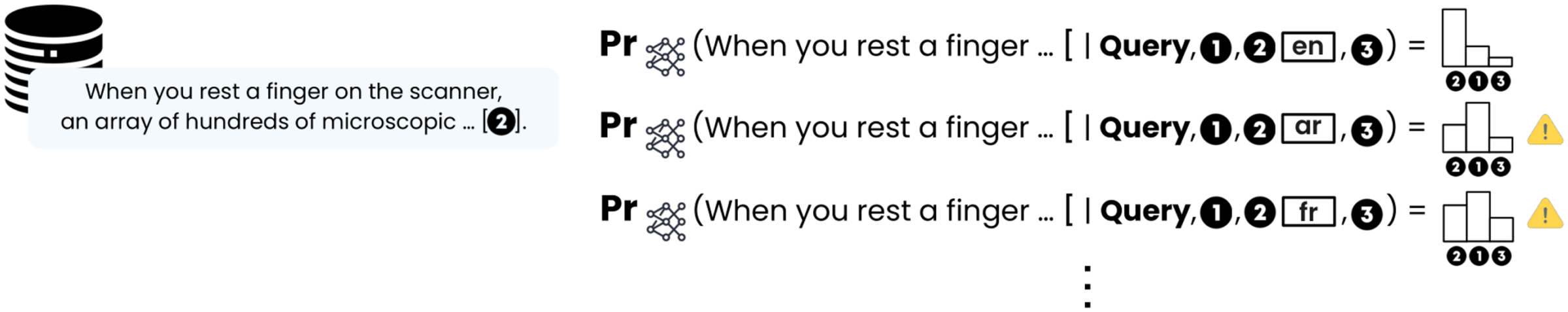
Query + ❶, ❷, ❸

When you rest a finger on the scanner, an array of hundreds of microscopic ... [❷]. An analog-to-digital converter then ... [❶]. ... captures 550 ppi scans ... [❸].

Step 3 : Filter for supported claims



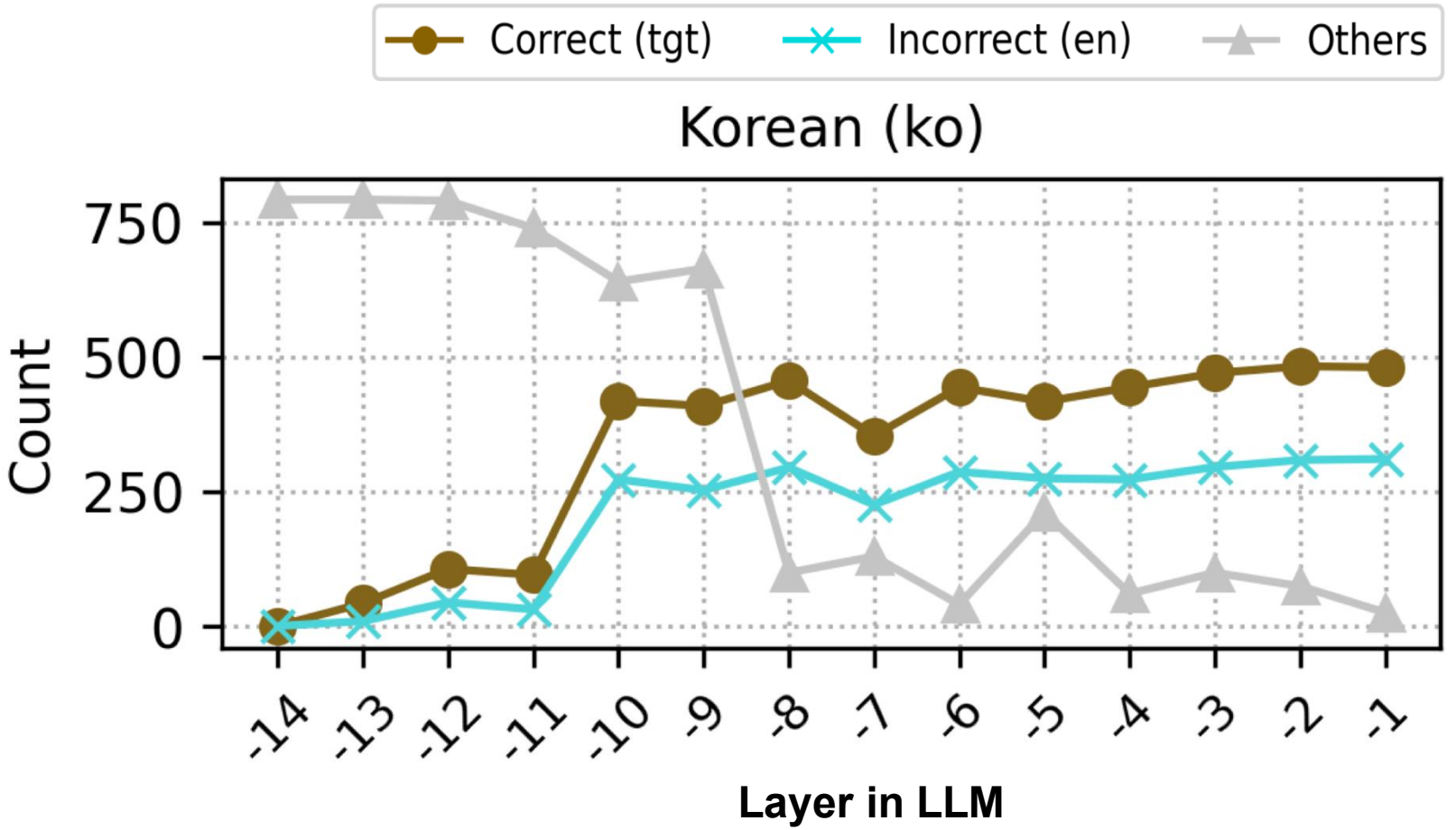
Step 4 : Get next token probability for document ID



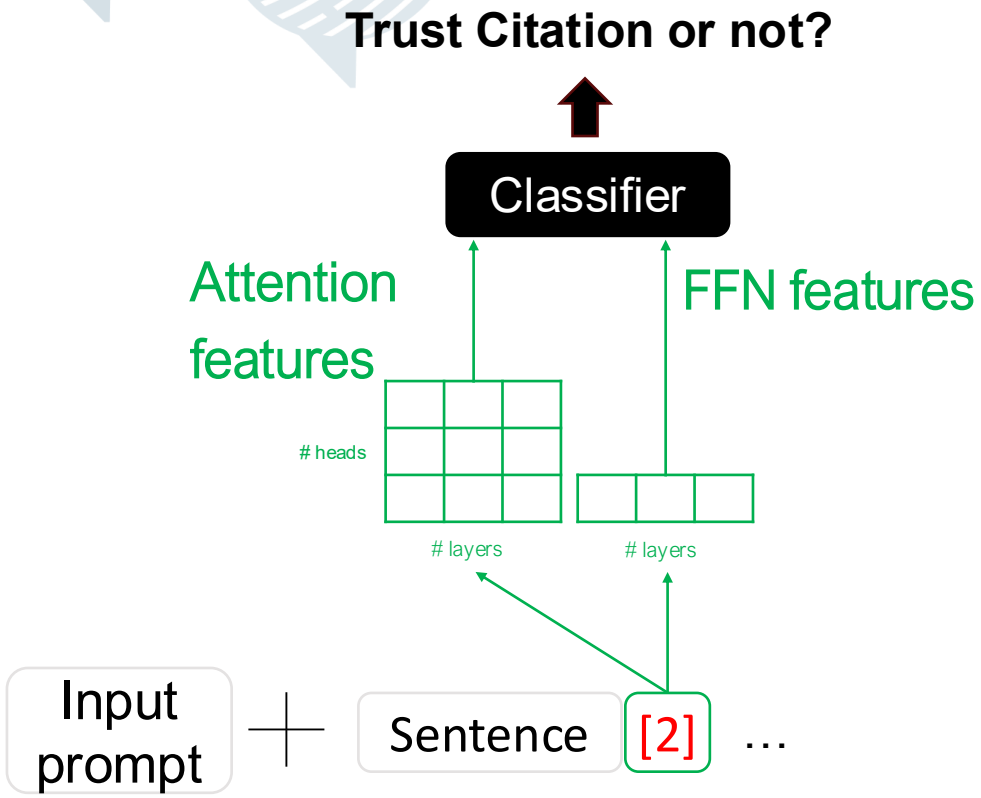
Logit Lens analysis of model internals

How does citation preference unfold during generation?

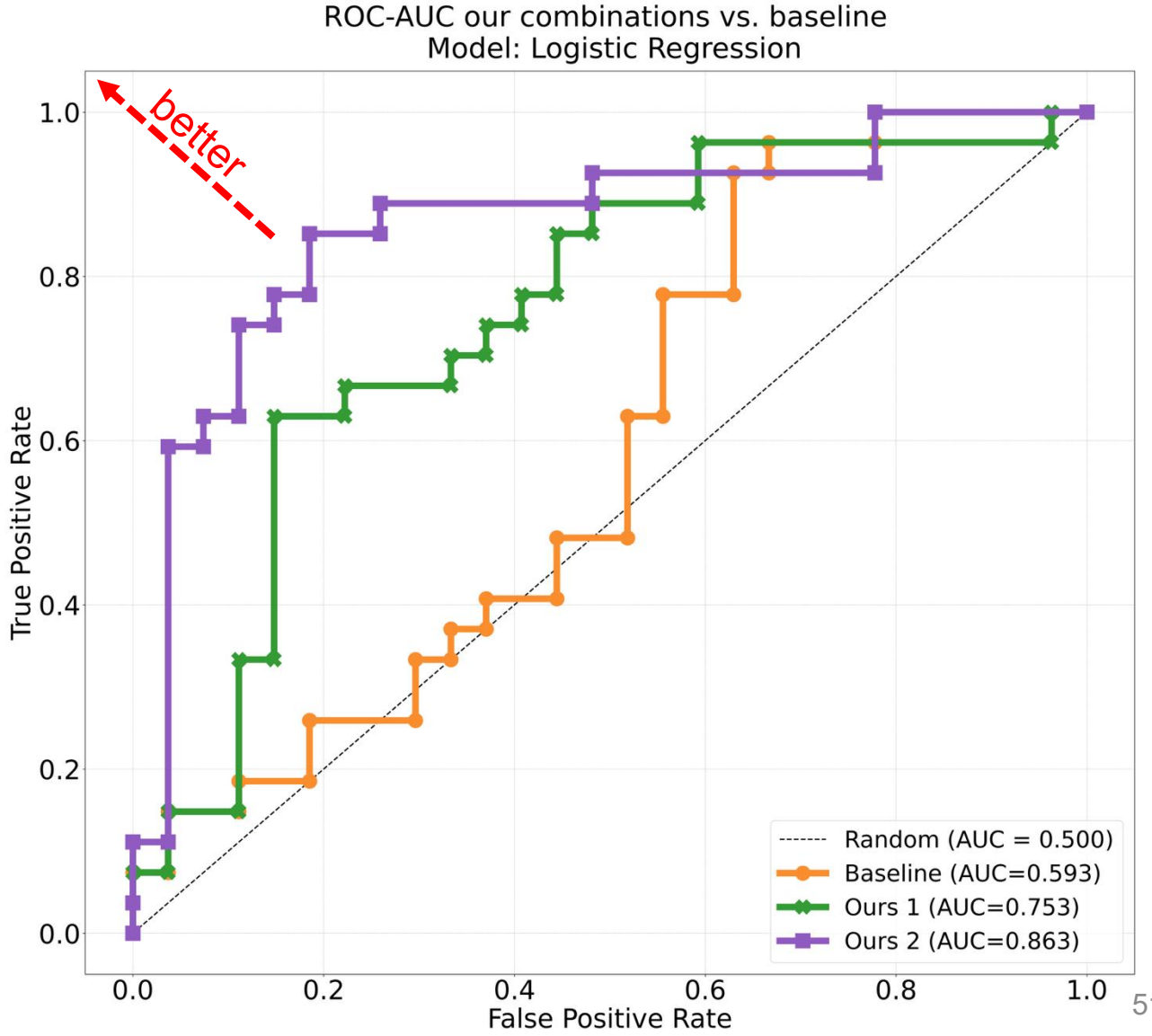
Is there switching between correct target and English as layers progress?



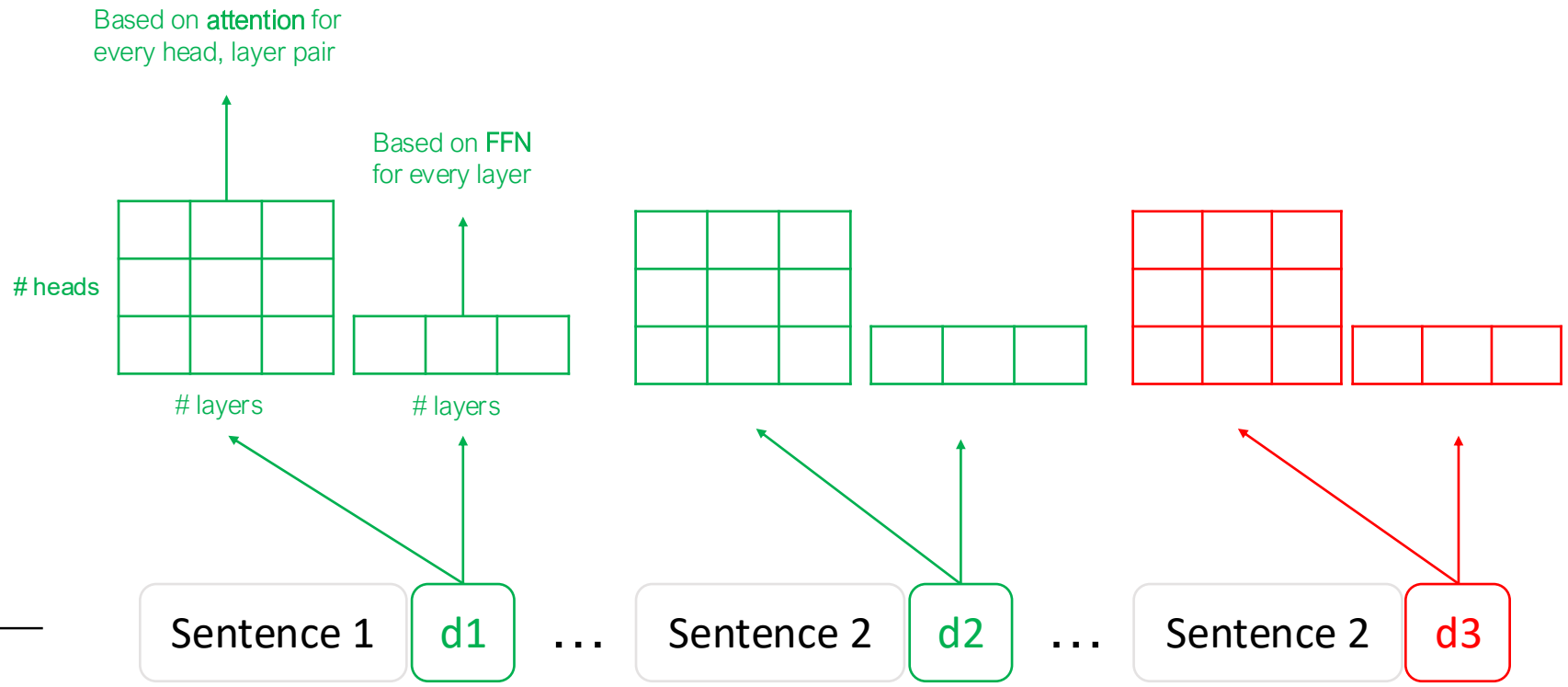
Detection of hallucinated citations



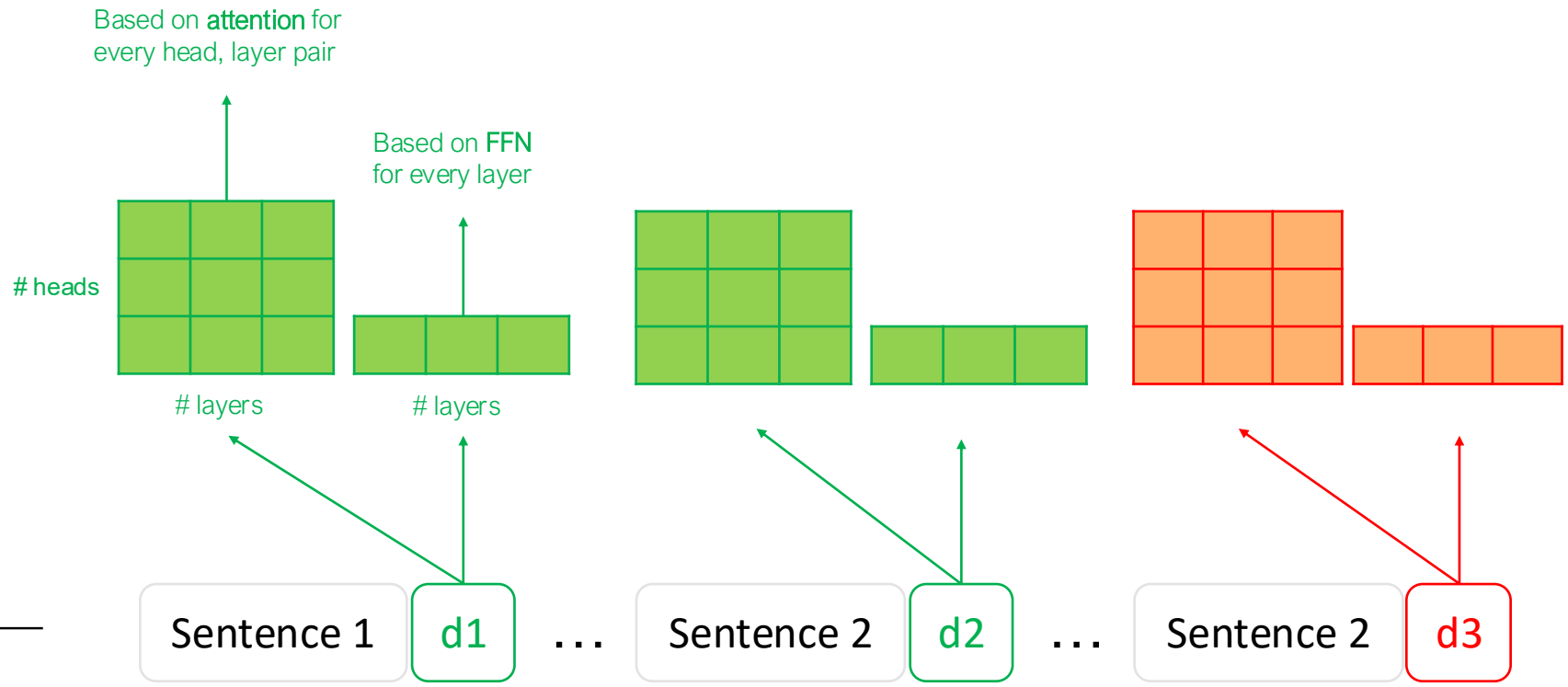
Approach: extract model-internal features, for use in a binary classifier



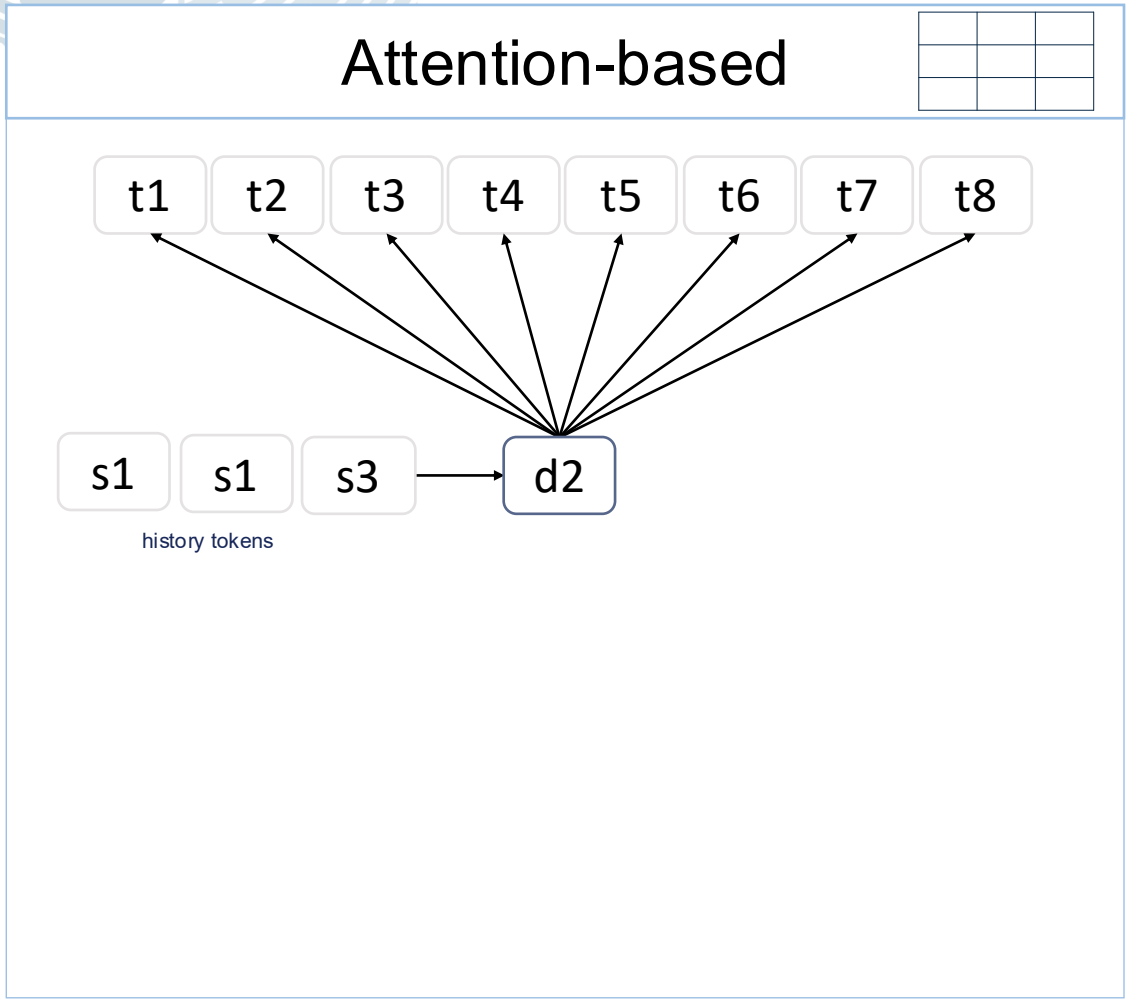
Training Data



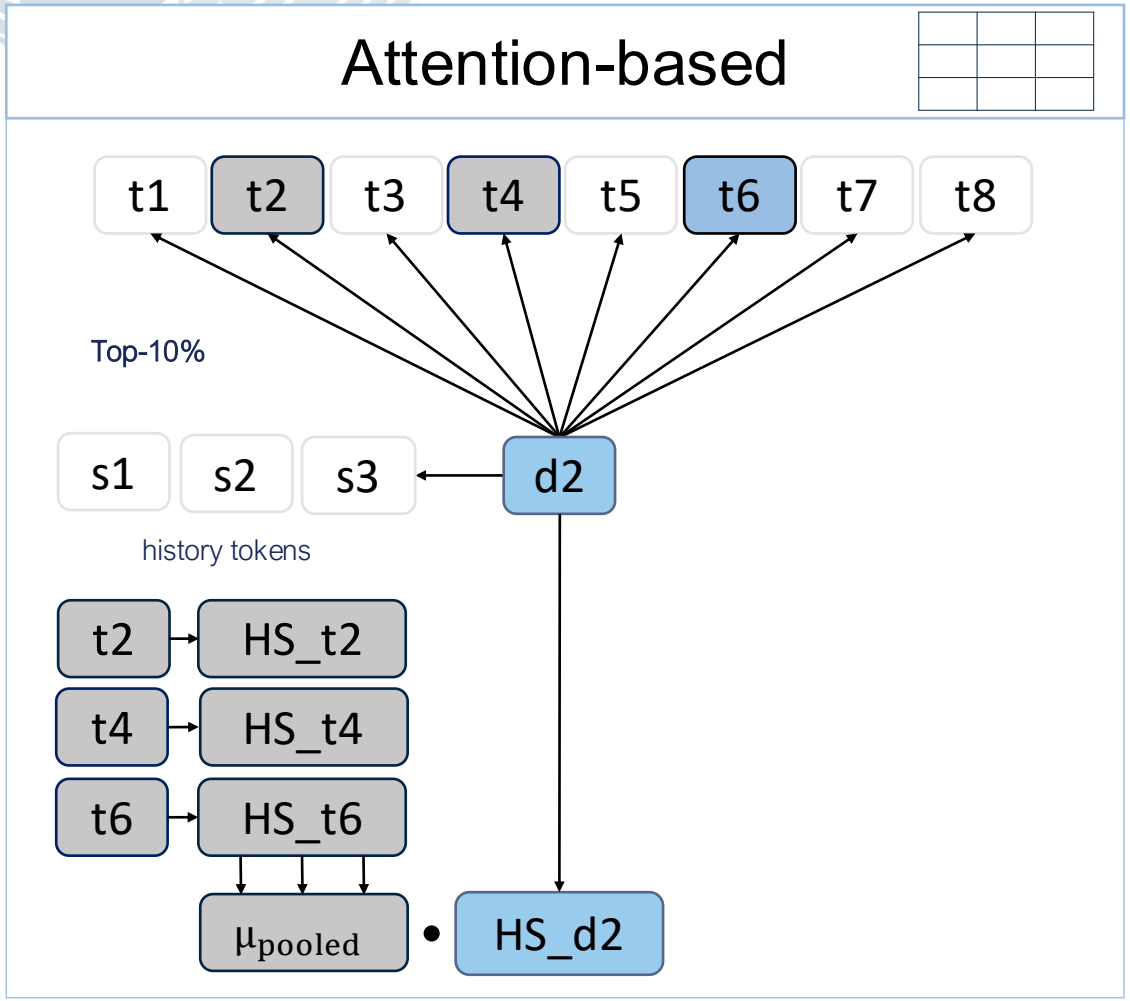
Training Data



Features for Hallucination Detection



Features for Hallucination Detection

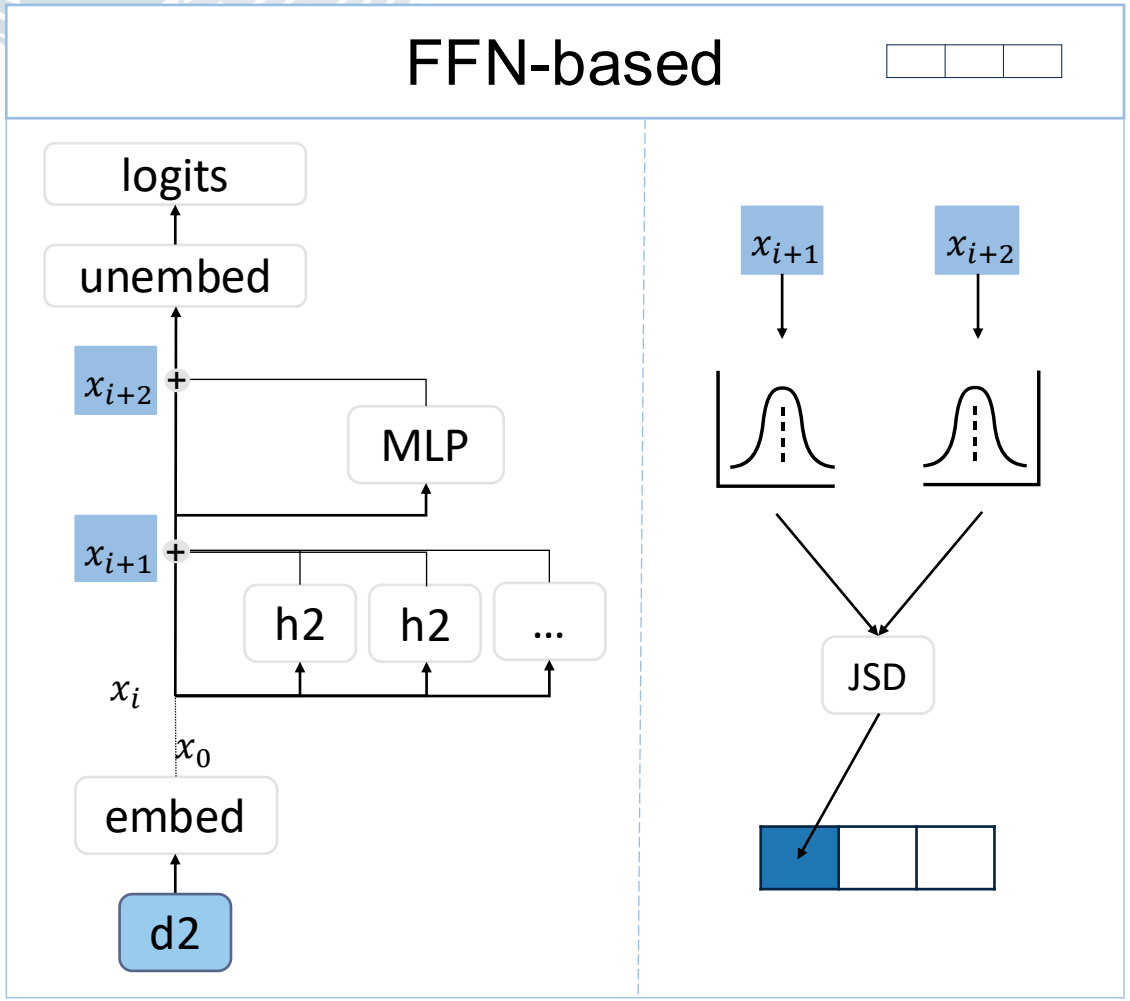


External Context Score (ECS)

→ How grounded is the generated token in the context it attended to the most?

$$ECS_{factual} > ECS_{hallucinated}$$

Features for Hallucination Detection



Parametric Knowledge Score (PKS)

→ How much parametric knowledge is put into the residual stream?

$$PKS_{factual} < PKS_{hallucinated}$$

See: Dassen et al. (2026) FACTUM: Mechanistic Detection of Citation Hallucination in Long-Form RAG
Sun et al. (2025) Re-deep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability