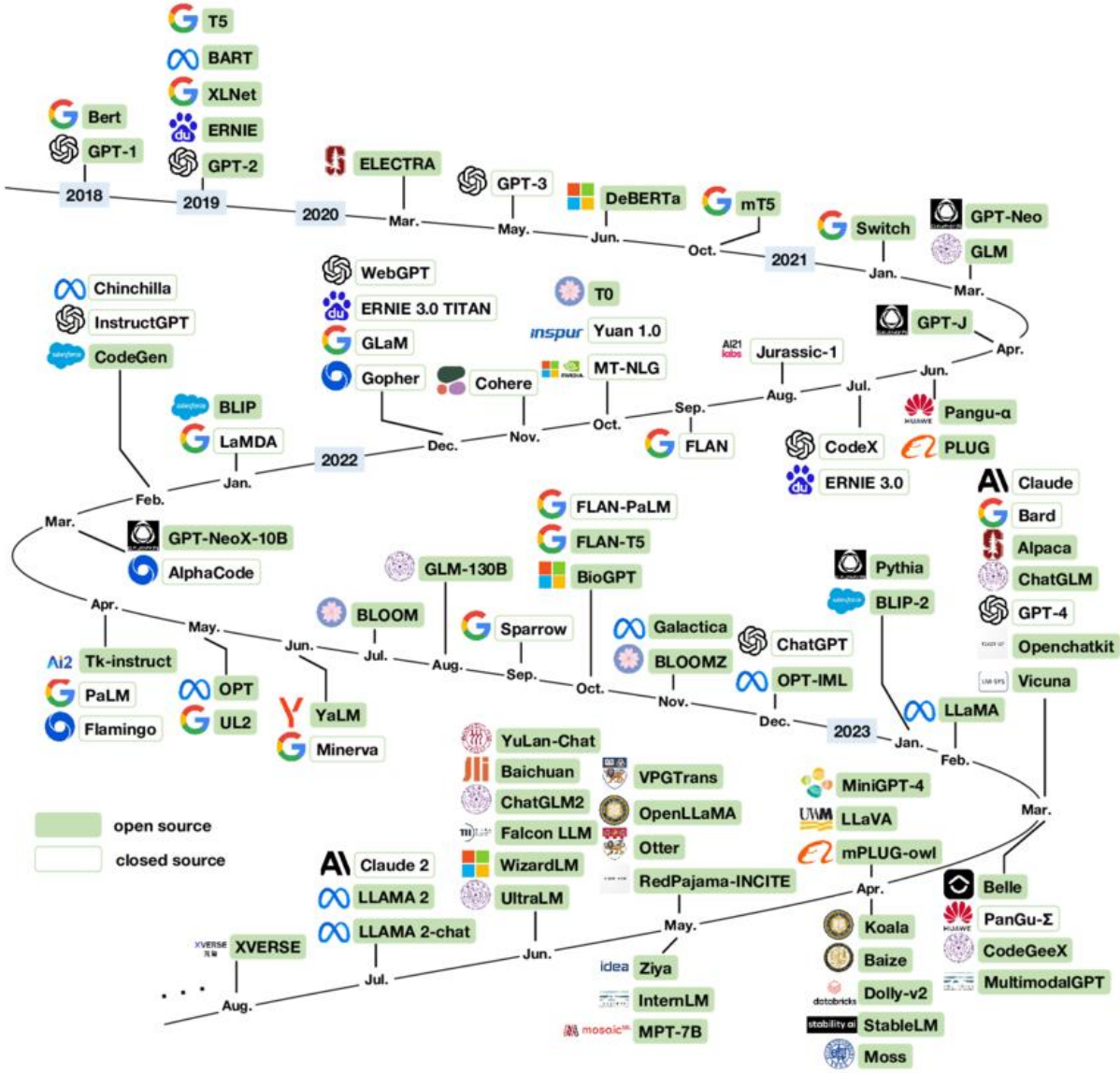
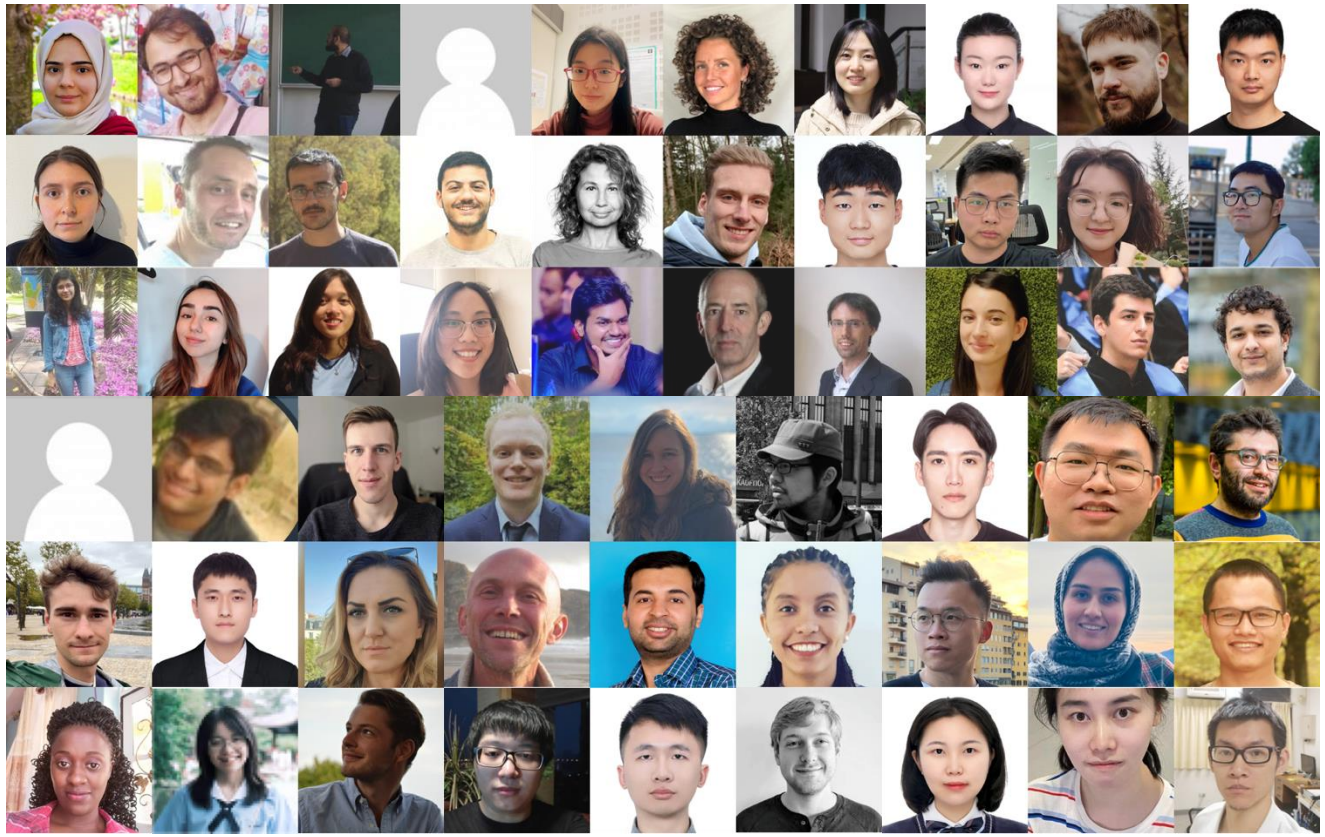
Agentic Information Access

Evangelos Kanoulas, UvA



Research at IRLab

- Neural and Generative Information Retrieval
  - Multimodal representation learning
  - Retrieval Augmented Generation
  - Online/counterfactual learning to rank
- Robust Retrieval and Language Models
  - Adversarial attacks/learning
  - Bias/noise in the input
  - Uncertainty estimation
- Explainability/Interpretability
  - Mechanistic interpretability
  - Machine unlearning
- User behaviour modelling
- Recommender systems
  - Interactive/sequential recsys
  - Bias/fairness/diversity
  - Learning



The Future of Information Access

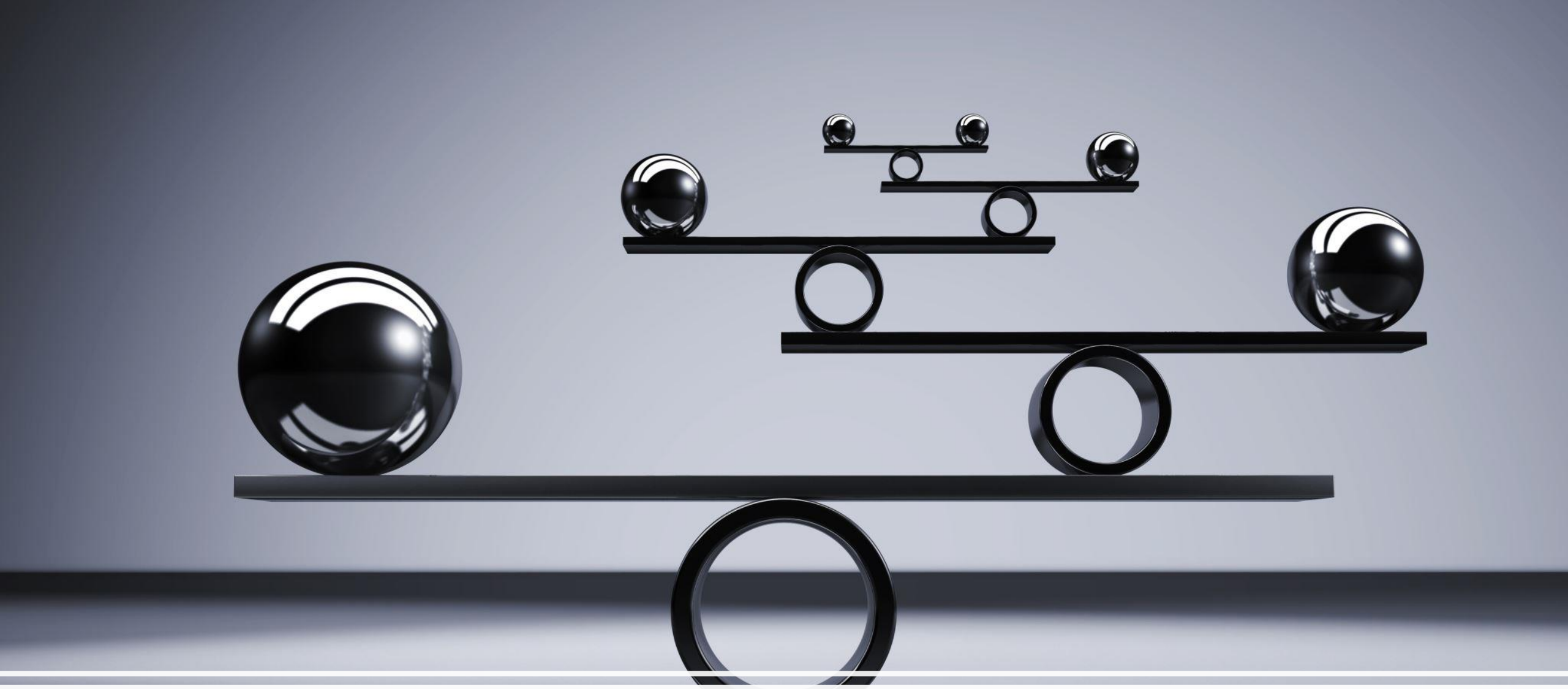
[Submitted on 26 Feb 2025]

Agent-centric Information Access

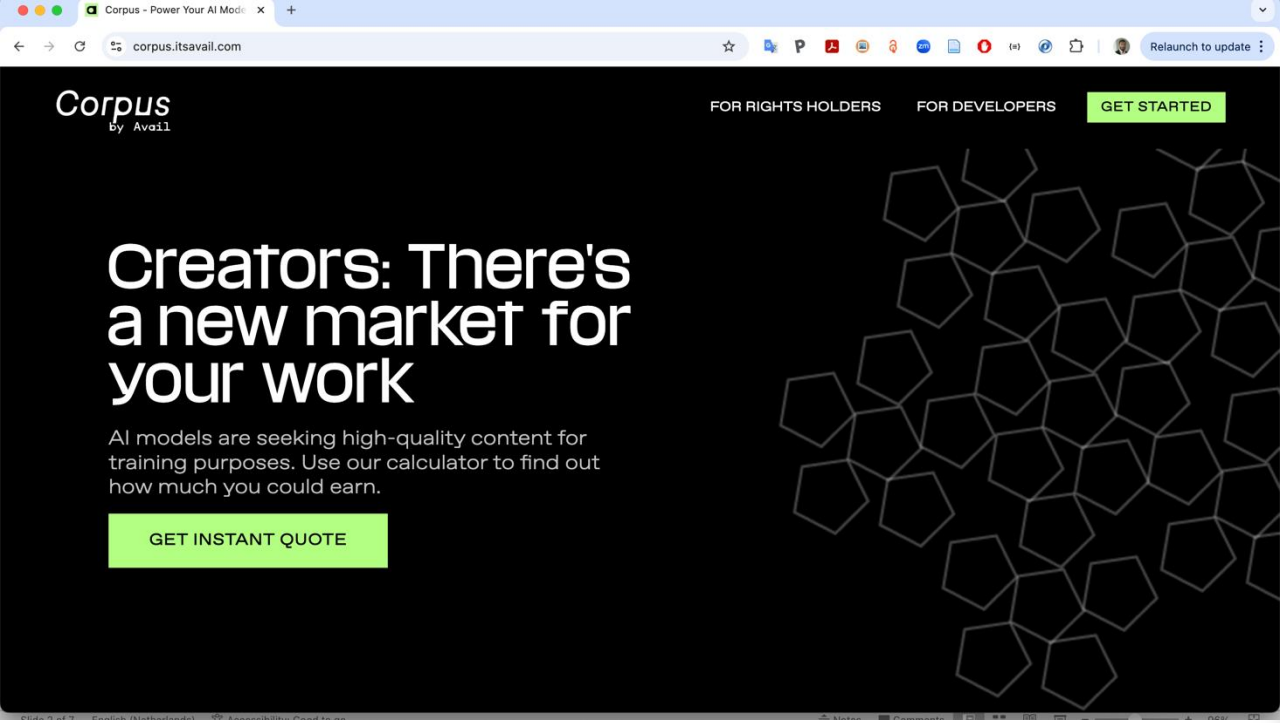
Evangelos Kanoulas, Panagiotis Eustratiadis, Yongkang Li, Yougang Lyu, Vaishali Pal, Gabrielle Poerwawinata, Jingfen Qiao, Zihan Wang

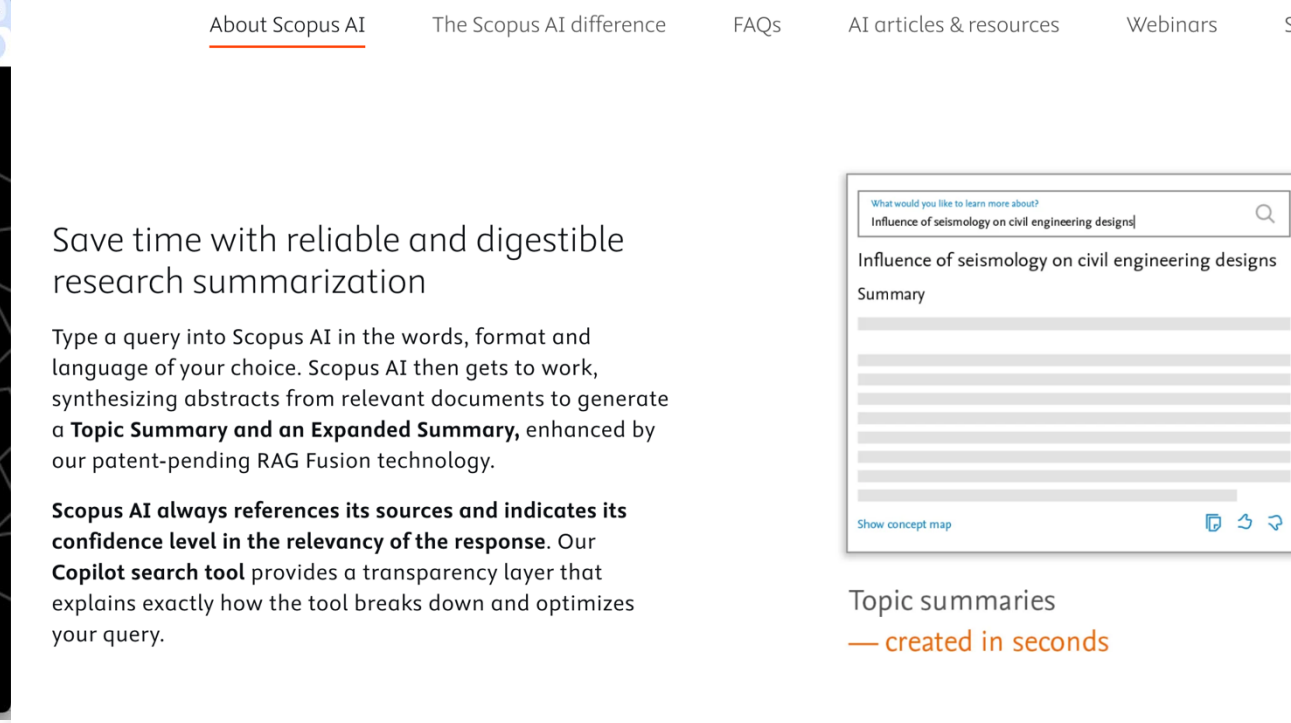
As large language models (LLMs) become more specialized, we envision a future where millions of expert LLMs exist, each trained on proprietary data and excelling in specific domains. In such a system, answering a query requires selecting a small subset of relevant models, querying them efficiently, and synthesizing their responses. This paper introduces a framework for agent-centric information access, where LLMs function as knowledge agents that are dynamically ranked and queried based on their demonstrated expertise. Unlike traditional document retrieval, this approach requires inferring expertise on the fly, rather than relying on static metadata or predefined model descriptions. This shift introduces several challenges, including efficient expert selection, cost-effective querying, response aggregation across multiple models, and robustness against adversarial manipulation. To address these issues, we propose a scalable evaluation framework that leverages retrieval-augmented generation and clustering techniques to construct and assess thousands of specialized models, with the potential to scale toward millions.

Subjects: **Information Retrieval (cs.IR)**
Cite as: arXiv:2502.19298 **[cs.IR]**
(or arXiv:2502.19298v1 **[cs.IR]** for this version)
https://doi.org/10.48550/arXiv.2502.19298

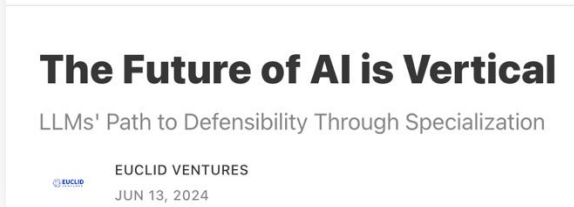


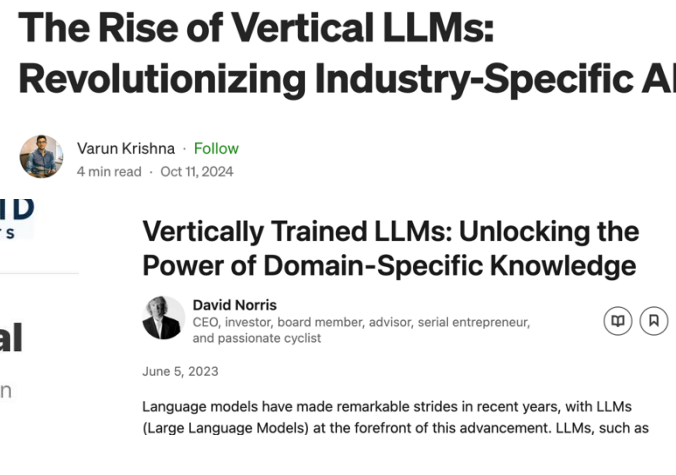
General purpose versus specialized LLMs



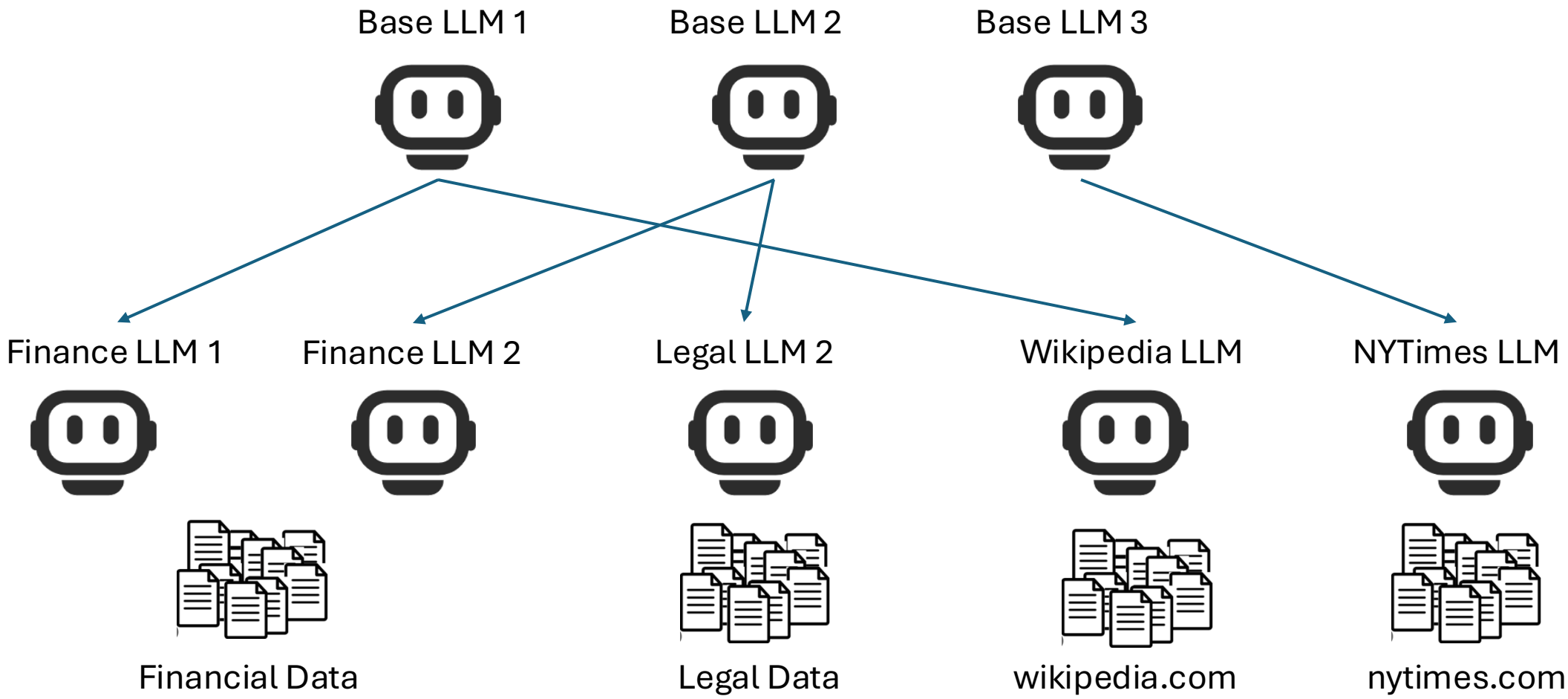


Content offered by vertical LLMs





A million LLMs



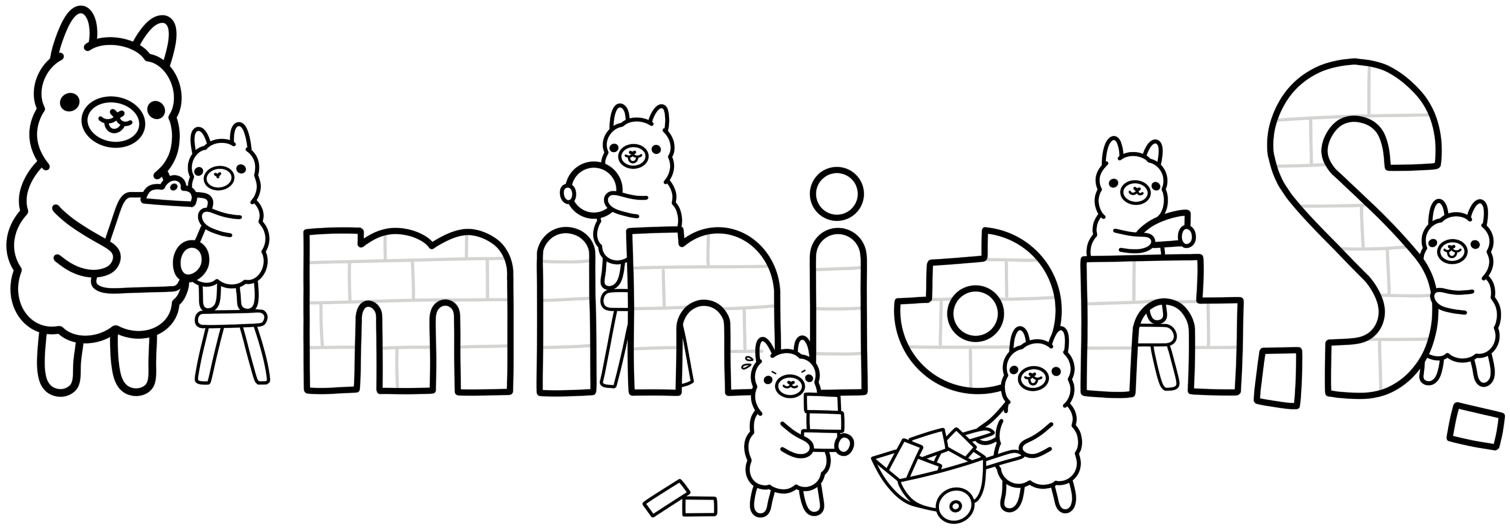
Collaboration between large and small models

Feb 24, 2025 · 7 min read

Minions: the rise of small, on-device LMs

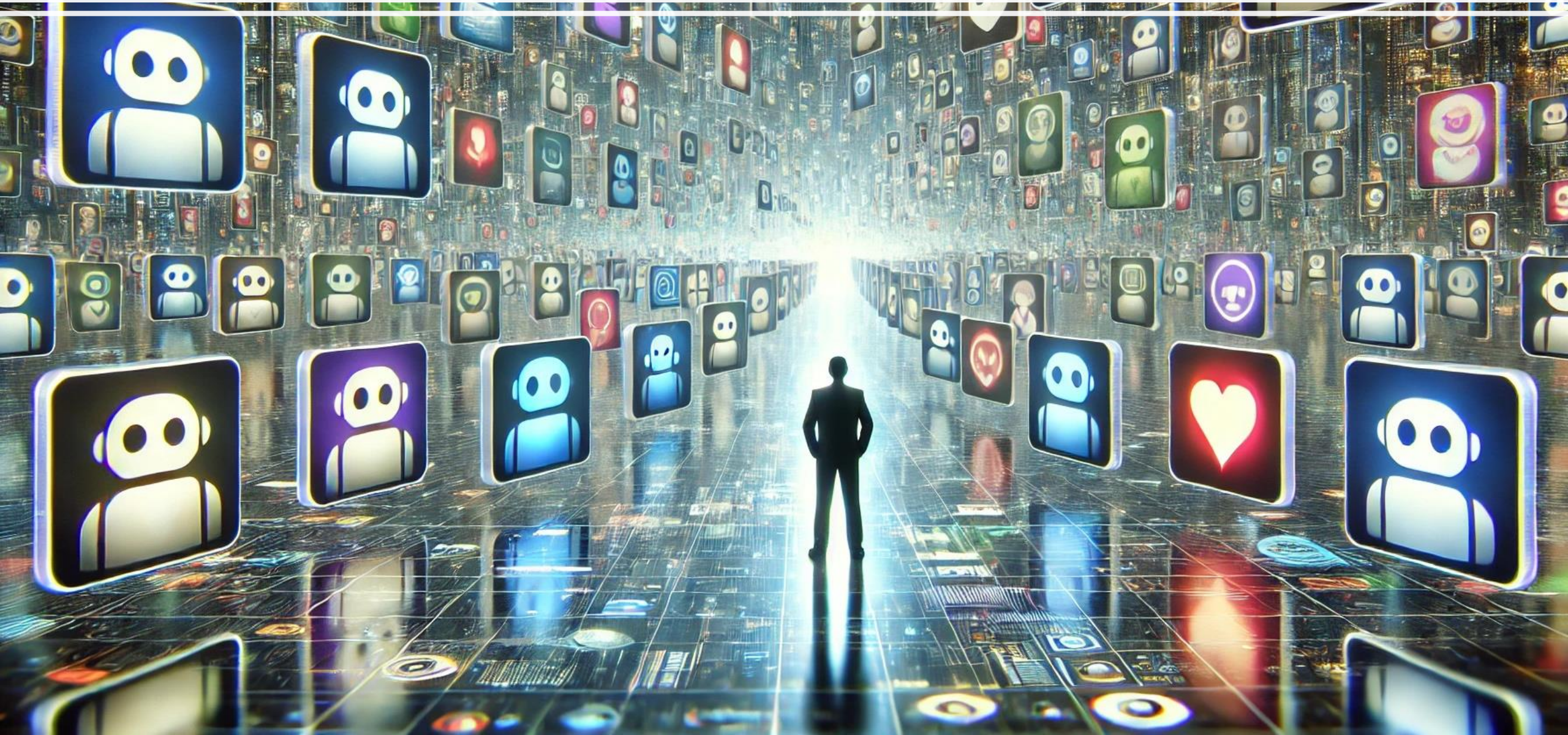
Sabri Eyuboglu*, Dan Biderman*, and Avanika Narayan*.

Embracing small LMs, shifting compute on-device, and cutting cloud costs in the process



🌬️ The Tailwinds: Frontier Models Cost $$$, While Local Hardware Collects Dust

A million LLMs



README | Code of conduct | Apache-2.0 license | Security

Agent2Agent (A2A) Protocol

pypi v0.2.4 License Apache 2.0

An open protocol enabling communication and interoperability between opaque agentic applications.

The Agent2Agent (A2A) protocol addresses a critical challenge in the AI landscape: enabling gen AI agents, built on diverse frameworks by different companies running on separate servers, to communicate and collaborate effectively - as agents, not just as tools. A2A aims to provide a common language for agents, fostering a more interconnected, powerful, and innovative AI ecosystem.

With A2A, agents can:

- Discover each other's capabilities.
- Negotiate interaction modalities (text, forms, media).
- Securely collaborate on long running tasks.
- Operate without exposing their internal state, memory, or tools.

Agent Discovery in A2A

For AI agents to collaborate using the Agent2Agent (A2A) protocol, they first need to find each other and understand what capabilities the other agents offer. A2A standardizes the format of an agent's self-description through the **Agent Card**. However, the methods for discovering these Agent Cards can vary depending on the environment and requirements.

The Role of the Agent Card

The Agent Card is a JSON document that serves as a digital "business card" for an A2A Server (the remote agent). It is crucial for discovery and initiating interaction. Key information typically included in an Agent Card:

- **Identity:** `name`, `description`, `provider` information.
- **Service Endpoint:** The `url` where the A2A service can be reached.
- **A2A Capabilities:** Supported protocol features like `streaming` or `pushNotifications`.
- **Authentication:** Required authentication `schemes` (e.g., "Bearer", "OAuth2") to interact with the agent.
- **Skills:** A list of specific tasks or functions the agent can perform (`AgentSkill` objects), including their `id`, `name`, `description`, `inputModes`, `outputModes`, and `examples`.

Client agents parse the Agent Card to determine if a remote agent is suitable for a given task, how to structure requests for its skills, and how to communicate with it securely.

HuggingGPT

Please generate an image where a girl is reading a book, and her pose is the same as the boy in the image example.jpg, then please describe the new image with your voice.

Request

Stage #1: Task Planning

[T1] task: pose-det. args: image: example.jpg

[T3] task: image-class. args: image: <resource-2>

[T5] task: image-to-text args: image: <resource-2>

task dependency

[T2] task: pose-to-image args: text: a reading girl image: <resource-1>

[T4] task: object-det. args: image: <resource-2>

[T6] task: text-to-speech args: text: <resource-5>

Stage #2: Model Selection

task: pose-det.

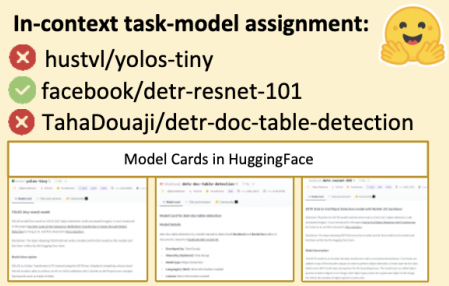
Query

task: object-det. args: image: <resource-2>

task: image-class.

In-context task-model assignment:

- hustvl/yolos-tiny
- facebook/detr-resnet-101
- TahaDouaji/detr-doc-table-detection



Stage #3: Task Execution

task, args, model

obj-det. <resource-2>

facebook/detr-resnet-101

task, args, model

Hybrid Endpoints

HuggingFace Endpoint (facebook/detr-resnet-101)

Local Endpoint (facebook/detr-resnet-101)

Bounding boxes with probabilities

Predictions

Stage #4: Response Generation

[T1] pose-det. openpose [image-2]

[T2] pose-to-image lym/sd-ctl-pose [image-3]

[T3] image-class. google/vit [classes with prob]

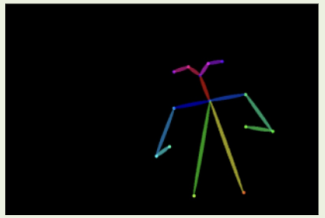
[T4] object-det. facebook/detr-resnet [bounding box]

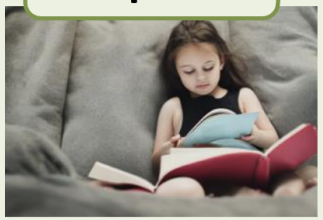
[T5] image-to-text nlpconnect/vit-gpt2 "a girl sitting on ..."

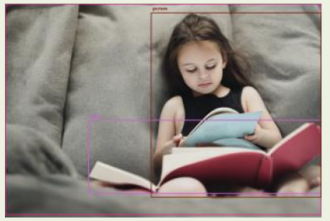
[T6] text-to-speech facebook/fastspeech [audio-1]

Summary

Response







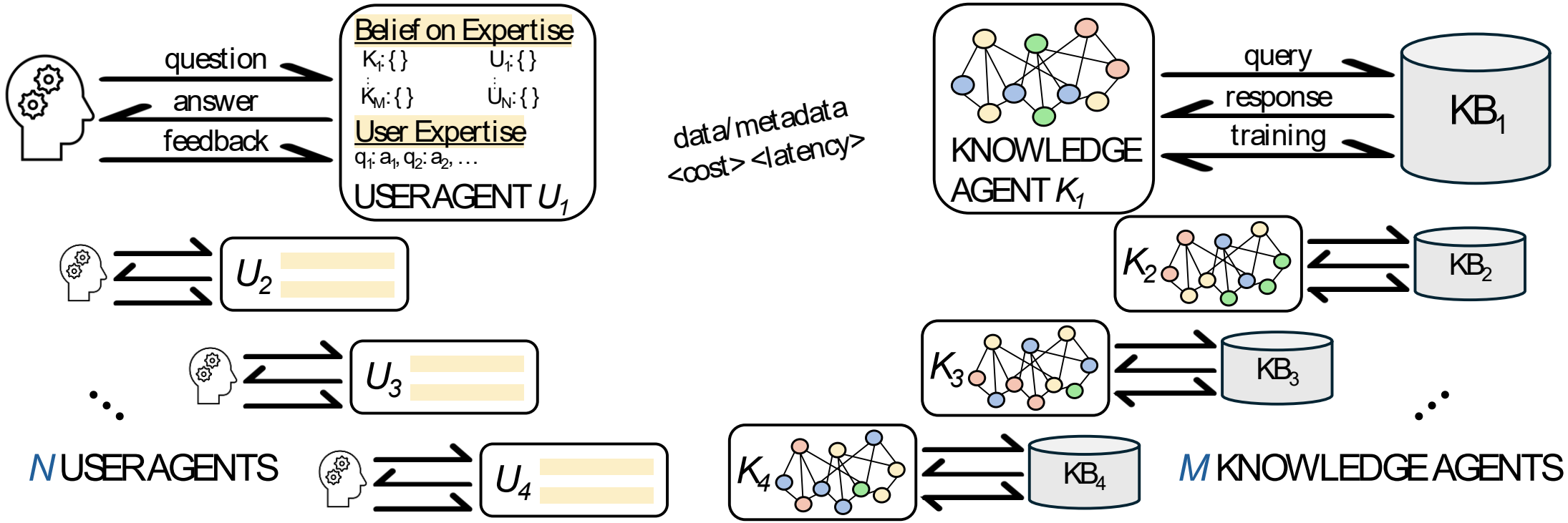
a girl sitting on a bed reading a book

[Text-1] = <resource-5>

[Image-1] = example.jpg [Image-2] = <resource-1> [Image-3] = <resource-2> [Image-4] [Audio-1]

The image you gave me is of "boy". The first thing I did was use OpenCV's openpose model to analyze the pose of the boy in the image. I fed that image, specifically located at [Image-1], into the model to get an output image of the pose, located at [Image-2]. Then I take that pose and generate a new image, located at [Image-3], using lllyasviel/sd-controlnet-openpose and the instruction "a girl is reading". Next, I

A million LLMs ecosystem



Which LLMs should be queried for a task?

How can we define and measure the expertise of an LLM?

How many LLMs should be queried once we identify a ranking of expertise for a given query?

How can we minimize the computational overhead of querying LLMs?

How can we update the expertise of LLMs in a dynamic fashion?

How should LLM responses be aggregated?

How can we develop methods that optimally aggregate different answers account for the query and user needs?

How do we present conflicting answers to users?

How do we quantify the reliability of an answer in a way that we can compare multiple heterogeneous models?

How do we attribute answers to the underlying models that provided them?

How do we ensure safety?

How do we ensure that expert selection is diverse and representative?

What is an effective adversarial attack in a model-centric retrieval?

How can we robustify ranking algorithms against adversarial attacks?

How do we evaluate agentic info access?

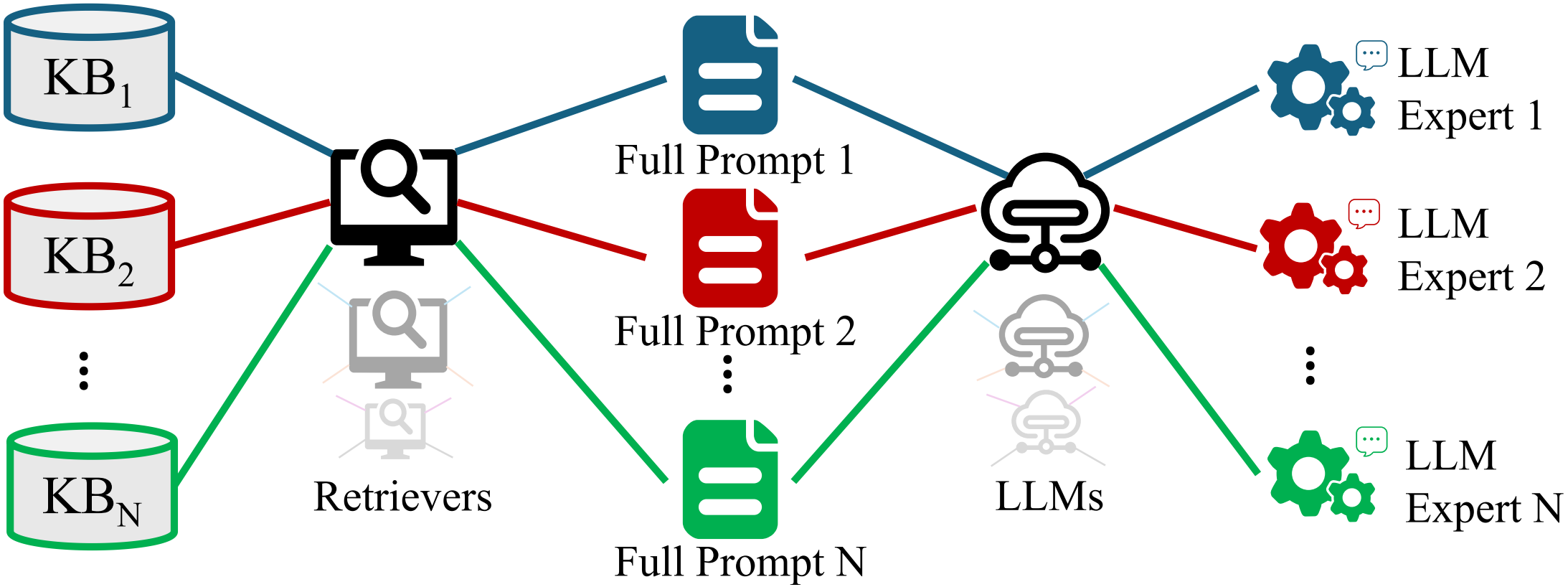
How can we build the necessary infrastructure to support a large number of specialized LLMs?

How can we build static and reusable test collections to assess the LLM expertise and the quality of the response aggregation?

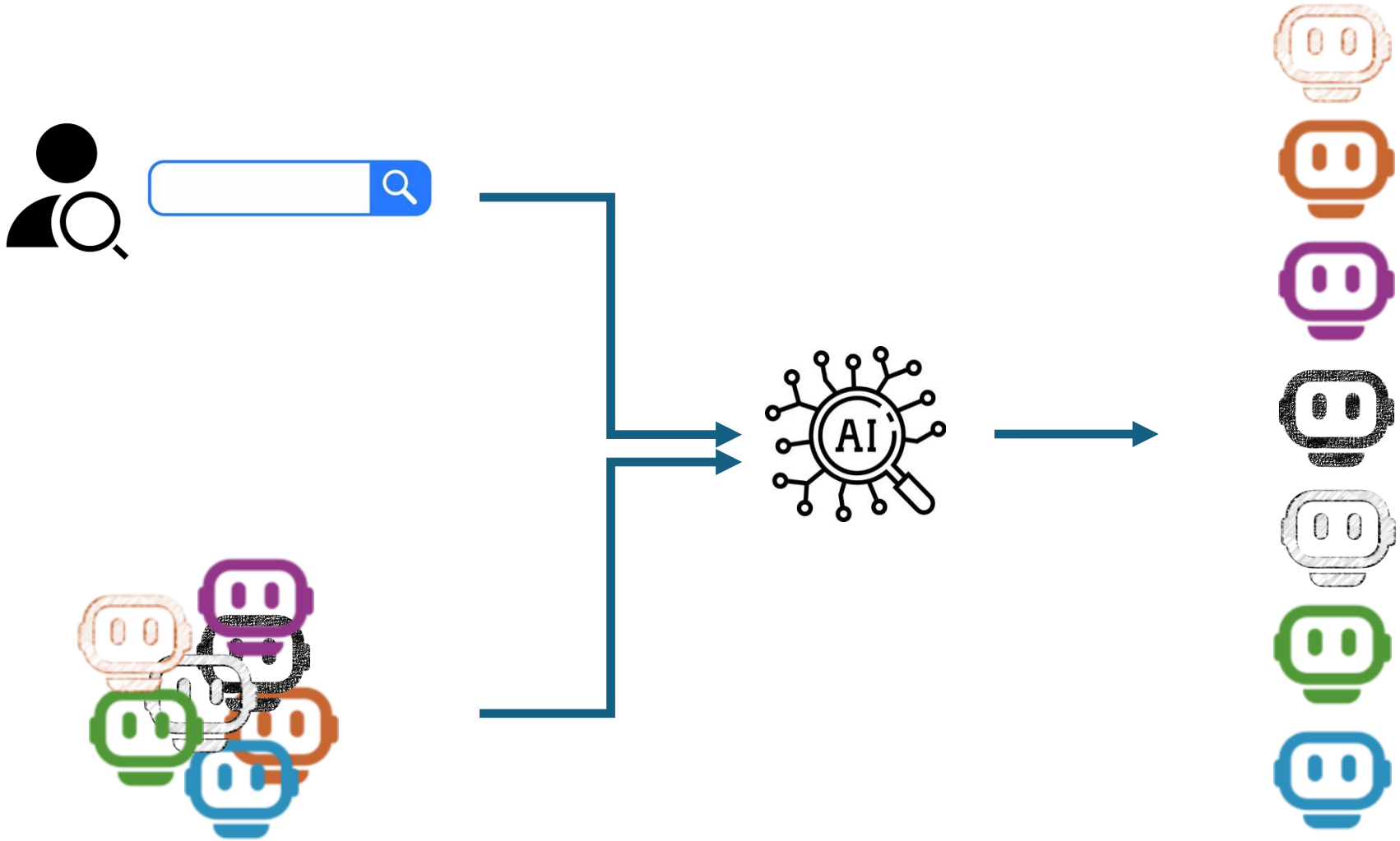
What metrics allow us to compare different user agents that search, retrieve, query and aggregate answers against multiple LLMs?

How can we get a million LLMs?

RAG instead of fine-tuning

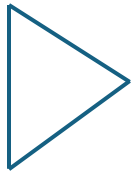


Problem Definition



Problem Definition

- Input:
  - A set of training queries and the answers from each one of the LLMs along with potential metadata
  - A set of test queries
- Output:
  - A ranked list of LLMs



Can you discover the expertise of each LLM?

$$
\begin{array}{lllll}
q_1 & a_1^1 & a_1^2 & \ldots & a_1^N \\
q_2 & a_2^1 & a_2^2 & \ldots & a_2^N \\
\ldots & & & & \\
\ldots & & & & \\
q_M & a_M^1 & a_M^2 & \ldots & a_M^N
\end{array}
$$

Implementation

- Document Collection
  - MS MARCO Web Search subset (ClueWeb22):
    - 80 million English documents out of 200M
    - Positively clicked Bing results (q, d+)
    - Reflects real-world search behavior and broad topical coverage
  - Researchy Questions Dataset:
    - 90,000 complex Bing queries
    - Chosen based on long dwell times
    - Yielded ~0.5 million clicked documents (q, d_1+, d_2+, ...)
    - Captures exploratory, multi-faceted information needs

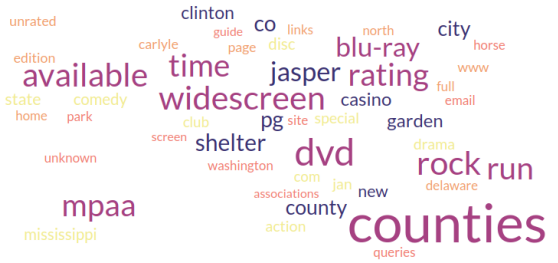
Document Collection

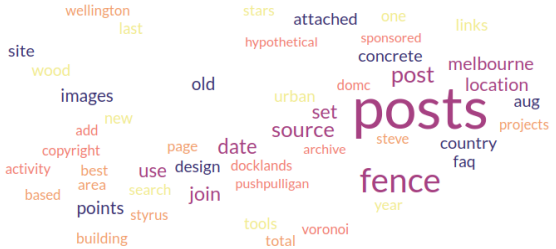
- MS MARCO Web Search subset (ClueWeb22):
  - 80 million English documents out of 200M
  - Positively clicked Bing results (q, d+)
  - Reflects real-world search behavior and broad topical coverage

- Researchy Questions Dataset:
  - 90,000 complex Bing queries
  - Chosen based on long dwell times
  - Yielded ~0.5 million clicked documents (q, d_1+, d_2+, …)
  - Captures exploratory, multi-faceted information needs

- Filtering for Metadata:
  - ClueWeb22-A provides semantic tags (topics, language, quality)
  - Only documents from the intersection with ClueWeb22-A were kept
  - Final corpus: ~80.5 million tagged English documents

Clustering





- Used k-means on tag-based features from ClueWeb22-A
- Formed 1,000 clusters, each ≥10,000 docs
- Each cluster = a domain of expertise

Retrieval Strategy

- Initial approach: Build 1,000 RAG systems (1 index per cluster)
  - Too expensive and inefficient
- Improved approach:
  - One global index over all 80.5M docs
  - For each query:
    - Retrieve top 2,000 documents
    - For each of the 1,000 clusters:
      - Select top-3 docs (if any) from the cluster
      - Use them to simulate the cluster's response
      - If no docs: return a default/template response

Query Collection

- Total: 20,000 queries
  - 10,000 Web Search queries (MS MARCO):
    - Short, keyword-style
    - Matched to clicked documents
    - Filtering step: Ensured answer quality depended on retrieval
      - Compared LLM answer with and without the clicked doc
      - Retained only queries with substantive differences
      - Used GPT-4o as LLM-as-a-Judge
  - 10,000 Researchy Queries:
    - Fully-formed natural language questions
    - Selected based on user engagement
    - Assumed to need retrieval, so used without filtering

Examples

Query	LLM Response
zeno gym review	I'm not aware of any reviews specifically for Zeno Gym. You might want to check recent fitness forums or consumer review sites.
zeno gym review	I'm not familiar with Zeno Gym. It could be a niche or less-reviewed brand.
riverside mental health	Riverside County offers various mental health services, including counseling and crisis support.
riverside mental health	Many schools in Riverside provide mental health resources to support students.
riverside mental health	Riverside Community Care is one provider offering mental health services and community support.
riverside mental health	Riverside has a range of mental health facilities, including outpatient clinics and community programs.
riverside mental health	If you're in the Riverside area, there are public and private organizations offering mental health assistance.

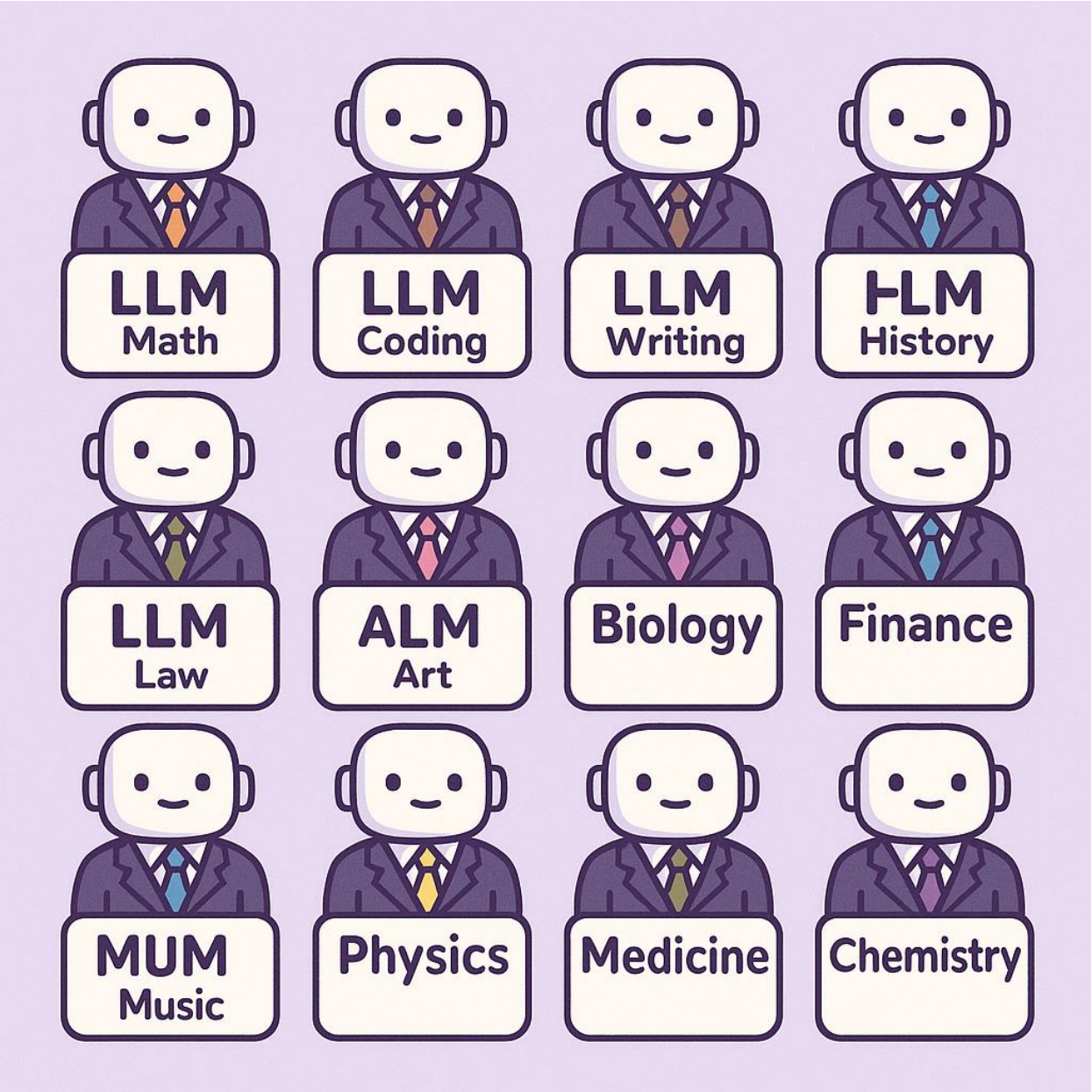
How do we evaluate an LLM ranking

Evaluation

- Input: test query; golden answer; ranked list of LLMs

- Method
  - Run the test query against all LLMs and produce a set of answers
    - Option 1:
      - Compute metric between golden answer and each of the LLM answers (e.g. ROUGE)
      - Use “Offline Evaluation by Maximum Similarity to an Ideal Ranking”
    - Option 2:
      - Consider the top-2000 documents and the gold answer and run Nuggetizer + UMBRELLA to identify the nuggets and which documents carry these nuggets.
      - Label documents by relevance grade based on coverage of the nuggets
      - Transfer these grades to the LLMs that have access to them.

How can we define and measure the expertise of an LLM?



How can we define and measure the expertise of an LLM?

- Simple baselines based on answers to the 20,000 questions
  - E.g. represent each LLM with the set of answers
  - Upon a query find the similarity between the query and LLM representation

How can we define and measure the expertise of an LLM?

- More sensible representations of expertise could try to explore metadata
  - E.g. the confidence of the produced answered by LLMs

[Submitted on 12 May 2025]

Why Uncertainty Estimation Methods Fall Short in RAG: An Axiomatic Analysis

Heydar Soudani, Evangelos Kanoulas, Faegheh Hasibi

Large Language Models (LLMs) are valued for their strong performance across various tasks, but they also produce inaccurate or misleading outputs. Uncertainty Estimation (UE) quantifies the model's confidence and helps users assess response reliability. However, existing UE methods have not been thoroughly examined in scenarios like Retrieval-Augmented Generation (RAG), where the input prompt includes non-parametric knowledge. This paper shows that current UE methods cannot reliably assess correctness in the RAG setting. We further propose an axiomatic framework to identify deficiencies in existing methods and guide the development of improved approaches. Our framework introduces five constraints that an effective UE method should meet after incorporating retrieved documents into the LLM's prompt. Experimental results reveal that no existing UE method fully satisfies all the axioms, explaining their suboptimal performance in RAG. We further introduce a simple yet effective calibration function based on our framework, which not only satisfies more axioms than baseline methods but also improves the correlation between uncertainty estimates and correctness.

LLM Uncertainty

- Uncertainty Estimation is a widely studied for assessing the reliability of LLM outputs
- Assigns an uncertainty score to each (input, output) pair, reflecting its truthfulness
- Ideally, a perfect UE method would assign lower uncertainty to correct samples and higher uncertainty to incorrect ones

LLM Uncertainty

- White-box methods
  - Predictive Entropy (PE)
  - Semantic Entropy (SE) (Kuhn et al., 2023)
  - Length Normalization and Semantic Awareness (Malinin and Gales, 2021)
  - MARS (Bakman et al., 2024)
  - SAR (Duan et al., 2024)
- Black-box methods
  - Sum of Eigenvalues (EigV) (Lin et al., 2024)
  - Degree Matrix (Deg)
  - Eccentricity (ECC)

Axiomatic Framework

- Theorem 1: Positively Consistent

  If the LLM gives the same answer before and after seeing a document, and that document supports the answer, then the LLM should be more confident after seeing the document.

  🔍 *Example:*
  Query: "What is the stadium of Manchester United?"
  Answer: "Old Trafford"
  The document says: "Manchester United plays at Old Trafford”
  Uncertainty drops.

UE	PopQA		
	BM25	**Contriever**	**Doc$^+$**
Axiom 1: Positively Consistent ↓			
PE	0.735 → 0.419 *	0.735 → 0.408 *	1.242 → 0.340 *
SE	3.781 → 3.205 *	3.791 → 3.158 *	4.682 → 3.113 *
PE+M	0.896 → 0.483 *	0.881 → 0.458 *	1.530 → 0.406 *
SE+M	4.102 → 3.286 *	4.091 → 3.248 *	5.146 → 3.173 *
EigV	1.951 → 1.166 *	2.025 → 1.143 *	4.074 → 1.078 *
ECC	0.417 → 0.110 *	0.426 → 0.094 *	0.710 → 0.055 *
Deg	0.220 → 0.048 *	0.230 → 0.043 *	0.496 → 0.022 *

Axiomatic Framework

- Theorem 2: Negatively Consistent

  If the LLM gives the same answer before and after seeing a document, but the document disagrees with that answer, the LLM should become less confident.

  🔍 *Example:*
  Query: "What is the stadium of Manchester United?"
  LLM says: "Wembley Stadium" (both before and after)
  Document says: "Manchester United plays at Old Trafford”
  Uncertainty should increase.

Axiom 2: Negatively Consistent ↑			
PE	1.068 → 0.746	0.820 → 0.593	1.083 → 0.597
SE	4.163 → 3.548 *	4.104 → 3.381 *	4.388 → 4.107
PE+M	1.309 → 0.844	1.016 → 0.782	1.328 → 0.684
SE+M	4.599 → 3.700 *	4.481 → 3.610 *	4.764 → 4.221
EigV	2.453 → 1.338 *	2.088 → 1.274 *	2.758 → 1.910
ECC	0.541 → 0.197 *	0.477 → 0.152 *	0.503 → 0.443
Deg	0.286 → 0.101 *	0.228 → 0.073 *	0.343 → 0.254

Axiomatic Framework

- Theorem 3: Positively Changed

  If the LLM changes its answer after seeing the document, and the new answer is correct while the old one was wrong, then the LLM should now be more confident.

  🔍 *Example:*
  Before: "Wembley Stadium"
  After seeing context: "Old Trafford"
  Uncertainty in the new answer should be lower.

Axiom 3: Positively Changed ↓			
PE	1.375 → 0.347 *	1.416 → 0.298 *	1.342 → 0.268 *
SE	4.889 → 3.015 *	5.091 → 3.013 *	4.884 → 3.051 *
PE+M	1.708 → 0.398 *	1.735 → 0.374 *	1.604 → 0.340 *
SE+M	5.514 → 3.072 *	5.681 → 3.082 *	5.379 → 3.099 *
EigV	4.131 → 1.139 *	4.733 → 1.114 *	4.449 → 1.102 *
ECC	0.790 → 0.085 *	0.823 → 0.081 *	0.780 → 0.072 *
Deg	0.547 → 0.044 *	0.588 → 0.035 *	0.544 → 0.032 *

Axiomatic Framework

- Theorem 4: Negatively Changed

  If the LLM initially gave the correct answer but changes it after seeing the document—and the new answer is incorrect—then the confidence should drop.

  🔍 *Example:*
  Before: "Old Trafford" (correct)
  After: "Wembley Stadium" (wrong)

  Uncertainty should now be higher than before.

Axiom 4: Negatively Changed ↑			
PE	0.933 → 0.636	1.006 → 0.558	1.252 → 0.463
SE	4.152 → 3.552 *	4.192 → 3.409 *	4.830 → 3.690 *
PE+M	1.164 → 0.714 *	1.298 → 0.748 *	1.689 → 0.747
SE+M	4.553 → 3.690 *	4.653 → 3.608 *	5.381 → 4.007 *
EigV	2.593 → 1.449 *	2.557 → 1.412 *	3.567 → 1.449 *
ECC	0.540 → 0.262 *	0.548 → 0.220 *	0.707 → 0.237 *
Deg	0.320 → 0.128 *	0.320 → 0.115 *	0.463 → 0.140 *

Axiomatic Framework

- Theorem 5: Neutrally Consistent

  If the document has nothing to do with the query or answer, and the LLM's answer doesn't change, its confidence should stay the same.

  🔍 *Example:*
  Query: "Who wrote The Origin of Species?"
  LLM says: "Charles Darwin"
  Document is about cooking recipes
  Uncertainty should stay the same.

Unc.	NQ-open	TriviaQA	PopQA
PE	2.072 → 2.248 *	0.872 → 1.155 *	0.897 → 0.909 *
SE	5.253 → 5.471 *	3.863 → 4.158 *	3.897 → 4.319 *
PE+M	4.791 → 4.805	1.415 → 1.699 *	1.031 → 1.130 *
SE+M	7.993 → 7.933	4.540 → 4.817 *	4.297 → 4.591
EigV	2.211 → 2.446 *	1.757 → 1.870 *	2.270 → 2.218
ECC	0.512 → 0.625 *	0.382 → 0.448 *	0.490 → 0.507
Deg	0.265 → 0.333 *	0.171 → 0.211 *	0.256 → 0.309

Uncertainty Estimation

Our results showed that none of the existing UE methods pass all the axiom, pin-pointing the problem in these methods.

Questions?

Image generated by ChatGPT using the prompt: "Create an image in the style of Van Gogh depicting a million LLMs in humanoid form (as a metaphor), working together to respond to user queries."

TREC 2025: Million LLMs Track

Defining "Search" in a world where the front-facing entity for every organization is a Conversational AI

Overview

In the era of Large Language Models (LLMs), we envision a transformation in how information is served to users. Imagine a world where:

- Content providers like Wikipedia, NY Times, Reddit, and Elsevier offer information through specialized LLMs
- Some LLMs provide cited responses with original source materials, while others offer direct answers
- Different levels of cooperation exist among LLMs, from those sharing embeddings and internal weights to those providing only text outputs
- Various access models coexist, from free-to-query to cost-based services
- Millions of LLMs operate with distinct, sometimes overlapping expertise domains