

# Information Seeking in the Age of Agentic AI

CHIIR 2026 TUTORIAL

**Preetam Dammu**

PhD Candidate, UW  
preetams@uw.edu

If you found these slides useful, please cite:

Dammu, P. P. S., & Roosta, T. (2026, March). Information Seeking in the Age of Agentic AI: A Half-Day Tutorial. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval* (pp. 429-430).

# Tutorial Structure

- Intro
- Approaches to Information Seeking
- How LLMs fill the gap
- Agentic AI Foundations
- System Complexity Ladder
- Hands-on Activities
  - Activity A: Complexity Ladder
  - Activity B: Structured Lookup & Grounded Closed-Corpus
  - Activity C: Grounded Live-Corpus
  - Activity D: Agentic Navigation

# Tutorial Structure

- Intro
- Approaches to Information Seeking
- How LLMs fill the gap
- Agentic AI Foundations
- System Complexity Ladder
- Hands-on Activities <https://isa-tutorial.github.io/isa-tutorial/index.html>
  - Activity A: Complexity Ladder
  - Activity B: Structured Lookup & Grounded Closed-Corpus
  - Activity C: Grounded Live-Corpus
  - Activity D: Agentic Navigation

# Approaches to Information Seeking

- There are multiple ways to think about information seeking.
- System-first approach: emphasis on architecture, methods, ..etc.
- We take a user-centric approach.



# Approaches to Information Seeking

- Three theoretical frameworks:
  - Belkin's **Anomalous State of Knowledge** approach [1]
  - Bates's **Berrypicking** model [2]
  - Broder's **Information Need** taxonomy [3]
- These frameworks help us describe **what the user is missing**, how **seeking unfolds**, and what **kind of need is being expressed**.

[1] Belkin, Nicholas J. "Anomalous states of knowledge as a basis for information retrieval." *Canadian journal of information science* 5.1 (1980): 133-143.

[2] Bates, Marcia J. "The design of browsing and berrypicking techniques for the online search interface." *Online review* 13.5 (1989): 407-424.

[3] Broder, Andrei. "A taxonomy of web search." *ACM Sigir forum*. Vol. 36. No. 2. New York, NY, USA: ACM, 2002.

# Belkin's *Anomalous State of Knowledge*

- ***“Users seek information because they know something is missing or inconsistent, and they often cannot fully articulate what they need at the outset”***
- ASK argues for the ***non-specifiability of information need***.
- Mismatch is not only about bad indexing or weak retrieval.
- User's request can be incomplete or distorted expression of the underlying need.
- ASK suggests ***different users may need different retrieval strategies***.

# Broder's Taxonomy

- Broder suggests that the need behind search queries is often ***not informational***.
- Informational queries constitute less than 50% of web searches<sup>[3]</sup>.
- Proposes a three-way taxonomy:
  - Navigational
  - Transactional
  - Informational

[3] Broder, Andrei. "A taxonomy of web search." ACM Sigir forum. Vol. 36. No. 2. New York, NY, USA: ACM, 2002.

# Broder's Taxonomy

- Broder suggests that the need behind search queries is often ***not informational***.
- Proposes a three-way taxonomy:
  - **Navigational: The immediate intent is to reach a particular site.**
  - Transactional
  - Informational

**Query:** national car rental

**Probable Target:** <http://www.nationalcar.com>

# Broder's Taxonomy

- Broder suggests that the need behind search queries is often ***not informational***.
- Proposes a three-way taxonomy:
  - Navigational:
  - **Transactional: The intent is to perform some web-mediated activity.**
  - Informational

**Main queries:** main categories for such queries are shopping, finding various web-mediated services, ..etc.

# Broder's Taxonomy

- Broder suggests that the need behind search queries is often ***not informational***.
- Proposes a three-way taxonomy:
  - Navigational:
  - Transactional:
  - **Informational: The intent is to acquire some information assumed to be present on one or more web pages.**

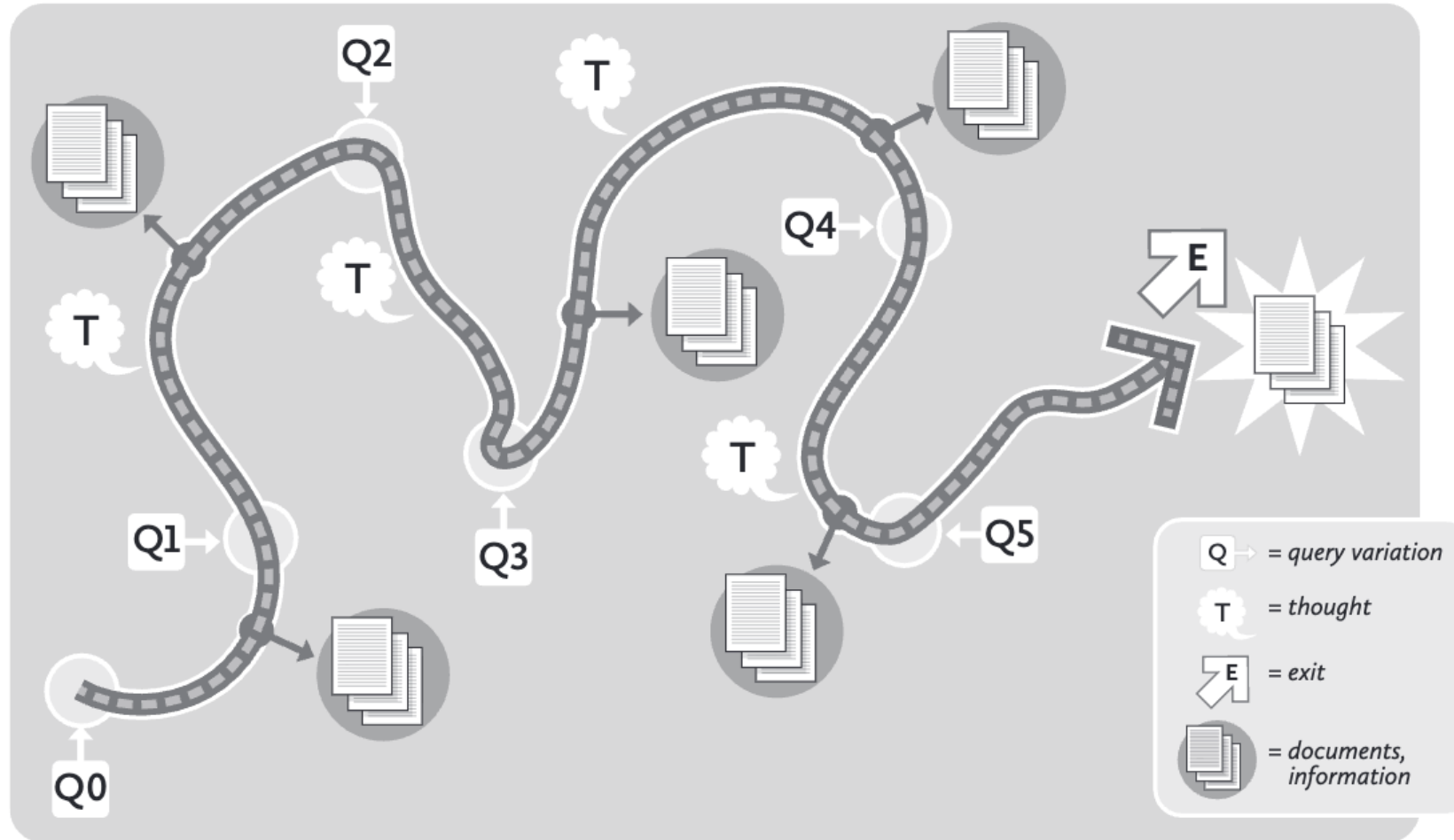
**Query:** normocytic anemia

*(Typically, informational queries are wide / upper funnel)*

# Bates's *Berrypicking* Model

- Bates's **berrypicking** model emphasizes that information seeking evolves over time.
- People reformulate, pivot, and accumulate partial “berries” across multiple sources, rather than executing a single “perfect” query.
- This becomes even more true once the “answer” is not a single fact, but a justification, a comparison, or a synthesis across many documents.

# Bates's *Berrypicking* Model



# Key Takeaways

- ***Belkin's ASK model***

- Expect user queries to be incomplete or imperfect expressions of their information need.
- Different user's may need different retrieval strategies / answers.

- ***Broder's Information Need Taxonomy***

- Less than 50% of queries are non-informational.
- Navigational / Transactional queries need to be treated differently.

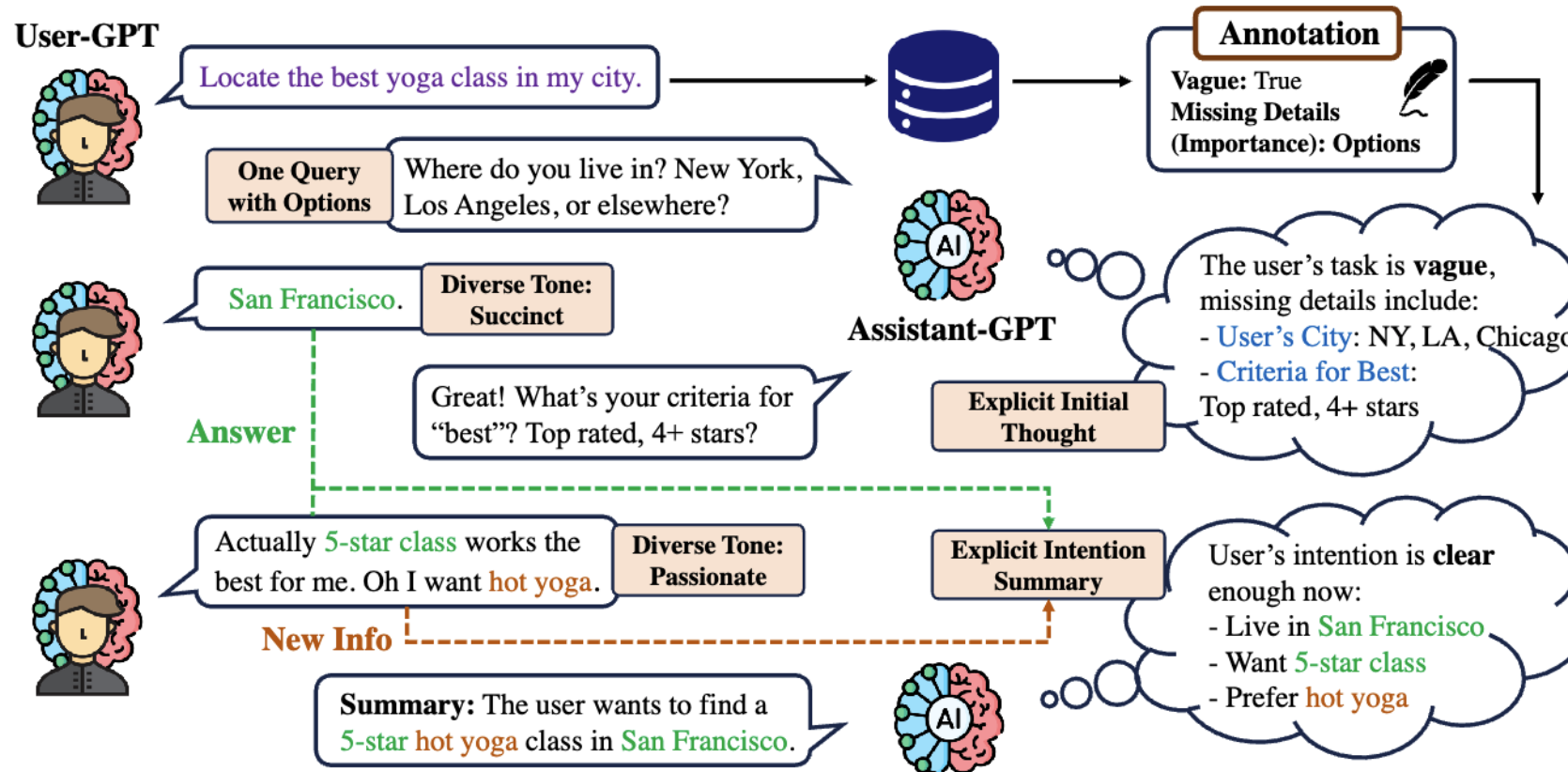
- ***Bates's Berrypicking model***

- Real user information seeking behavior involves exploring, pivoting, and iteratively updating queries.

# How LLMs help

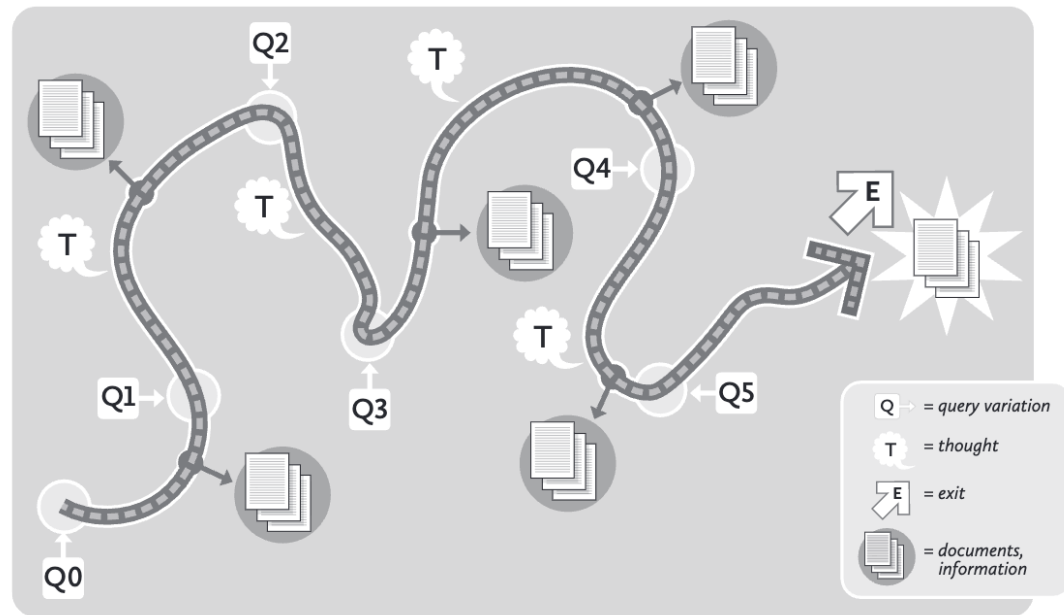
- LLMs unlock capabilities that were previously unattainable.
- For example:
  - Belkin's non-specifiability of information need -> **Follow-up Questions**
  - Bates's behavioral mismatch -> **Agentic Exploration**

# Belkin's *ASK* <> Follow-up Questions



# Bates's *Berrypicking* Model <> ReAct

**Question:** Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?



**Thought 1:** I need to search Apple Remote and find the program it was originally designed to interact with.

**Act 1:** `Search[Apple Remote]`

**Obs 1:** The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the `Front Row media center` program ...

**Thought 2:** Apple Remote was originally designed to control the Front Row media center program. I need to search `Front Row` next and find what other device can control it.

**Act 2:** `Search[Front Row]`

**Obs 2:** Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

**Thought 3:** Front Row is not found. I need to search `Front Row (software)`.

**Act 3:** `Search[Front Row (software)]`

**Obs 3:** Front Row is a discontinued media center software ...

**Thought 4:** `Front Row (software)` is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.

**Act 4:** `Finish[keyboard function keys]`

# Complexity Ladder

LLM-only (follow-up, rephrasing, ..etc.)

**Written Re-expression**

What is the best...?

API Integration (Augmented Generation)

**Structured Lookup**

Has Tesla stock gone up since last year?

Curated RAG (Vector DB, ..etc.)

**Grounded Closed-Corpus**

Who was the president during the great recession?

Live RAG (Search Engine, ..etc.)

**Grounded Live-Corpus**

Do I need snowchains to drive to Mt Rainier?

Agentic RAG (Web Navigation, ..etc.)

**Agentic Navigation**

Are there any courtside seats left on Saturday's Lakers game?

DB-level Reasoning (GraphRAG, ..etc.)

**Corpus Sensemaking**

Across all TIME articles this year, what are the main geopolitical changes?

# Foundations of Agentic AI for IR

Break

# Activity A: Complexity Ladder

- <https://isa-tutorial.github.io/isa-tutorial/activity1.html>

The screenshot shows a web browser window with the address bar containing the URL `isa-tutorial.github.io/isa-tutorial/activity1.html`. The browser interface includes a star icon, a share icon, a print icon, and a refresh icon. Below the browser window, there are two identical activity cards. Each card contains a question, a 'Complexity level' section with five buttons, and a text input field.

What is the current UV index in Seattle right now?

**Complexity level**

Structured Lookup   Grounded Closed-Corpus   Grounded Live-Corpus   Agentic Navigation   Corpus Sensemaking

Why is this level most suitable?

What time does the coffee break start in the provided Tutorial Schedule PDF?

**Complexity level**

Structured Lookup   Grounded Closed-Corpus   Grounded Live-Corpus   Agentic Navigation   Corpus Sensemaking

Why is this level most suitable?

# Complexity Ladder

LLM-only (follow-up, rephrasing, ..etc.)

**Written Re-expression**

What is the best...?

API Integration (Augmented Generation)

**Structured Lookup**

Has Tesla stock gone up since last year?

Curated RAG (Vector DB, ..etc.)

**Grounded Closed-Corpus**

Who was the president during the great recession?

Live RAG (Search Engine, ..etc.)

**Grounded Live-Corpus**

Do I need snowchains to drive to Mt Rainier?

Agentic RAG (Web Navigation, ..etc.)

**Agentic Navigation**

Are there any courtside seats left on Saturday's Lakers game?

DB-level Reasoning (GraphRAG, ..etc.)

**Corpus Sensemaking**

Across all TIME articles this year, what are the main geopolitical changes?

# Properties Determining Position on the Complexity Ladder

- 1. External Information Channel**
- 2. Freshness and temporal sensitivity**
- 3. Synthesis requirement**
- 4. Multi-hop dependency**
- 5. Scope of evidence**
- 6. Interaction pattern**

# Properties Determining Position on the Complexity Ladder

## 1. External Information Channel

- Is it an API that returns clean JSONs? Is it an indexed DB? Open Web?

## 2. Freshness and temporal sensitivity

## 3. Synthesis requirement

## 4. Multi-hop dependency

## 5. Scope of evidence

## 6. Interaction pattern

# Properties Determining Position on the Complexity Ladder

## 1. External Information Channel

- Is it an API that returns clean JSONs? Is it an indexed DB? Open Web?

## 2. Freshness and temporal sensitivity

- Is the answer stable, or does it change daily/hourly?

## 3. Synthesis requirement

## 4. Multi-hop dependency

## 5. Scope of evidence

## 6. Interaction pattern

# Properties Determining Position on the Complexity Ladder

## 1. External Information Channel

- Is it an API that returns clean JSONs? Is it an indexed DB? Open Web?

## 2. Freshness and temporal sensitivity

- Is the answer stable, or does it change daily/hourly?

## 3. Synthesis requirement

- Does it require summarizing, comparing, or reasoning across multiple sources?

## 4. Multi-hop dependency

## 5. Scope of evidence

## 6. Interaction pattern

# Properties Determining Position on the Complexity Ladder

## 1. External Information Channel

- Is it an API that returns clean JSONs? Is it an indexed DB? Open Web?

## 2. Freshness and temporal sensitivity

- Is the answer stable, or does it change daily/hourly?

## 3. Synthesis requirement

- Does it require summarizing, comparing, or reasoning across multiple sources?

## 4. Multi-hop dependency

- Do you need intermediate sub-questions or tool calls to get the final answer?

## 5. Scope of evidence

## 6. Interaction pattern

# Properties Determining Position on the Complexity Ladder

## 1. External Information Channel

- Is it an API that returns clean JSONs? Is it an indexed DB? Open Web?

## 2. Freshness and temporal sensitivity

- Is the answer stable, or does it change daily/hourly?

## 3. Synthesis requirement

- Does it require summarizing, comparing, or reasoning across multiple sources?

## 4. Multi-hop dependency

- Do you need intermediate sub-questions or tool calls to get the final answer?

## 5. Scope of evidence

- Is it enough to cite one source, or do we need broad coverage across a corpus?

## 6. Interaction pattern

# Properties Determining Position on the Complexity Ladder

## 1. External Information Channel

- Is it an API that returns clean JSONs? Is it an indexed DB? Open Web?

## 2. Freshness and temporal sensitivity

- Is the answer stable, or does it change daily/hourly?

## 3. Synthesis requirement

- Does it require summarizing, comparing, or reasoning across multiple sources?

## 4. Multi-hop dependency

- Do you need intermediate sub-questions or tool calls to get the final answer?

## 5. Scope of evidence

- Is it enough to cite one source, or do we need broad coverage across a corpus?

## 6. Interaction pattern

- One-shot lookup vs iterative clarification, mixed-initiative dialogue, exploration.

# Complexity Ladder

LLM-only (follow-up, rephrasing, ..etc.)

**Written Re-expression**

What is the best...?

API Integration (Augmented Generation)

**Structured Lookup**

Has Tesla stock gone up since last year?

Curated RAG (Vector DB, ..etc.)

**Grounded Closed-Corpus**

Who was the president during the great recession?

Live RAG (Search Engine, ..etc.)

**Grounded Live-Corpus**

Do I need snowchains to drive to Mt Rainier?

Agentic RAG (Web Navigation, ..etc.)

**Agentic Navigation**

Are there any courtside seats left on Saturday's Lakers game?

DB-level Reasoning (GraphRAG, ..etc.)

**Corpus Sensemaking**

Across all TIME articles this year, what are the main geopolitical changes?

# Written Re-expression

- **Need archetype**
  - “Creative gift”
  - “What is the best...?”
- **Properties:**
  - Vague or underspecified queries.
  - Subjective Queries

What is the best gift for my grandson?

# Structured Lookup

- **Need archetype**
  - “Give me the current weather in Seattle,”
  - “Apple stock price”
  - “Directions from A to B.”
- **Properties:**
  - Highly structured target (APIs, ..etc.)
  - Little to no “retrieval” algo/work required.

What was the stock price of Apple last week before closing?

# Grounded Closed-Corpus

- **Need archetype**

- “Answer based on my company docs,”
- “Use the course notes,”
- “Use a known database.”

- **Properties:**

- user needs factuality and provenance,
- the information may be long-tail or not in model weights.

In the provided insurance plan summary PDF, what is the in-network deductible for individual coverage?

# Grounded Live-Corpus

- **Need archetype**

- “Have layovers via Airport XYZ resumed?”
- “What is the latest announcement about Anthropic?”

- **Properties:**

- high temporal sensitivity,
- potentially adversarial misinformation,
- potential contradictions.

Do I need tire chains to drive to Mt. Rainier tomorrow?

# Agentic Navigation

- **Need archetype**

- “Are there any upper deck corners seats left for FIFA game 36?”
- “What is the latest latest commit to Transformers repo?”

- **Properties:**

- Requires clicking on buttons / navigation,
- Information inaccessible using search APIs.

Are any reservable sites available at Cougar Rock Campground (Mt. Rainier) for August 10-12?

# Corpus Sensemaking

- **Need archetype**

- “Are there any upper deck corners seats left for FIFA game 36?”
- “What is the latest latest commit to Transformers repo?”

- **Properties:**

- Requires knowledge obtained by consuming lots of info,
- Does not fit context length,

From the provided folder of 50 news articles about ‘AI regulation,’ what are the main policy themes and key disagreements?

# Activity A: Complexity Ladder

- <https://isa-tutorial.github.io/isa-tutorial/activity1.html>

# LLMs

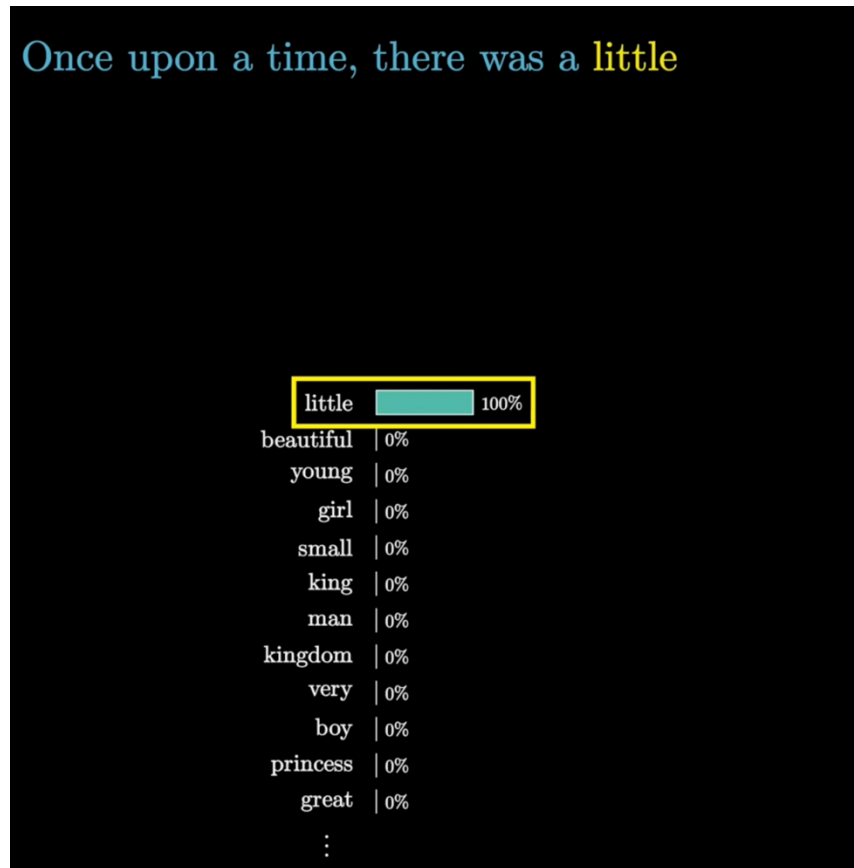
- LLMs are still surprisingly simple.

LLMs are just autocomplete on steroids

- Linus Towalds (creator of Linux)

# LLMs

- LLMs are autoregressive models – they only predict one token at a time!



Credits: 3Blue1Brown

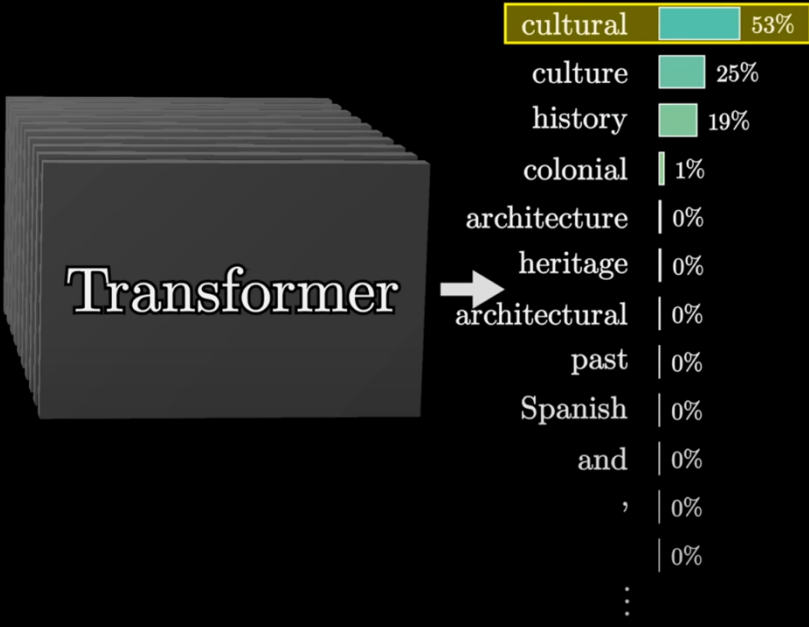
# LLMs

- LLMs are autoregressive models – they only predict one token at a time!

What follows is a conversation between a user and a helpful, very knowledgeable AI assistant.

User: Give me some ideas for what to do when visiting Santiago.

AI Assistant: Sure, there are plenty of things to do in Santiago! One option could be to take a walking tour of the city's historic center and learn about its rich **cultural** \_\_\_\_\_



<b>cultural</b>	53%
culture	25%
history	19%
colonial	1%
architecture	0%
heritage	0%
architectural	0%
past	0%
Spanish	0%
and	0%
,	0%
	0%
⋮	

Credits: 3Blue1Brown

# Typical Base LLM I/O Tokens

**System Prompt:**

You are a helpful  
assistant...

**User Profile:**

Prefers concise/detailed  
responses...

**User Query:**

Who won the Superbowl 2026?

**System Response:**

I don't have info after  
2024 (training cutoff).

# RAG (LLM + External Info)

**System Prompt:**  
You are a helpful assistant...

**User Profile:**  
Prefers concise/detailed responses...

**User Query:**  
Who won the Superbowl 2026?

**System Response:**  
I don't have info after 2024 (training cutoff).

**System Prompt:**  
You are a helpful assistant...

**User Profile:**  
Prefers concise/detailed responses...

**User Query:**  
Who won the Superbowl 2026?

**Retriever Response:**  
[1] Doc from espn.com  
[2] Doc from nfl.com, ...

**System Response:**  
Seattle Seahawks

# RAG (LLM + External Info)

**System Prompt:**  
You are a helpful assistant...

**User Profile:**  
Prefers concise/detailed responses...

**User Query:**  
Who won the Superbowl 2026?

**System Response:**  
I don't have info after 2024 (training cutoff).

**System Prompt:**  
You are a helpful assistant...

**User Profile:**  
Prefers concise/detailed responses...

**User Query:**  
Who won the Superbowl 2026?

**Retriever Response:**  
[1] Doc from espn.com  
[2] Doc from nfl.com, ...

**System Response:**  
Seattle Seahawks

## Builds on IR Machinery

- BM25
- Semantic Search
- SQL
- Knowledge Graphs
- Search Engine
- Web Browsers
- ....etc.

# Activity B: Structured Lookup & Grounded Closed-Corpus

- <https://isa-tutorial.github.io/isa-tutorial/activity2.html>

# Activity C: Grounded Live-Corpus

- <https://isa-tutorial.github.io/isa-tutorial/activity3.html>

# Activity D: Agentic Navigation

- <https://isa-tutorial.github.io/isa-tutorial/activity4.html>

# Evaluating Agentic Information Seeking

Background & Challenges

# Background

- Benchmarking, in the context of AI, refers to the process of ***capturing the performance of different AI models or systems*** using a predefined set of metrics.
- Test sets of popular ML datasets are used for benchmarking.

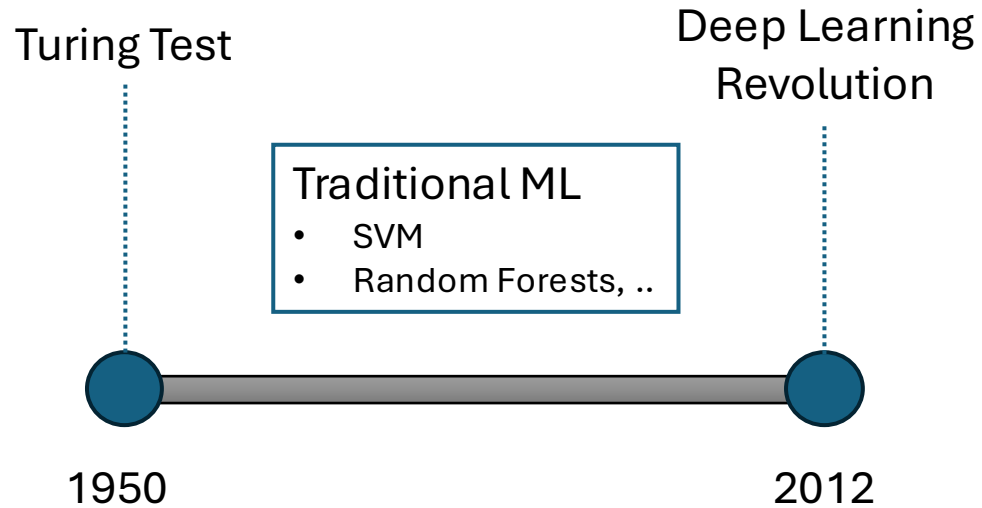
# Popular Benchmarks

- Web Navigation
  - WebShop
  - WebArena
  - Mind2Web
- Question Answering
  - HotPotQA
  - Natural Questions
  - TriviaQA
- Fact Verification
  - FEVER
  - CREAK
  - FaVIQ
- Generalist
  - AgentBench
  - GAIA
  - AGIEval
- Tool Calling
  - ToolBench
  - Toolformer
  - $\tau$ -Bench
- Coding
  - SWE-bench
  - HumanEval
  - RepoBench

## Common Traits:

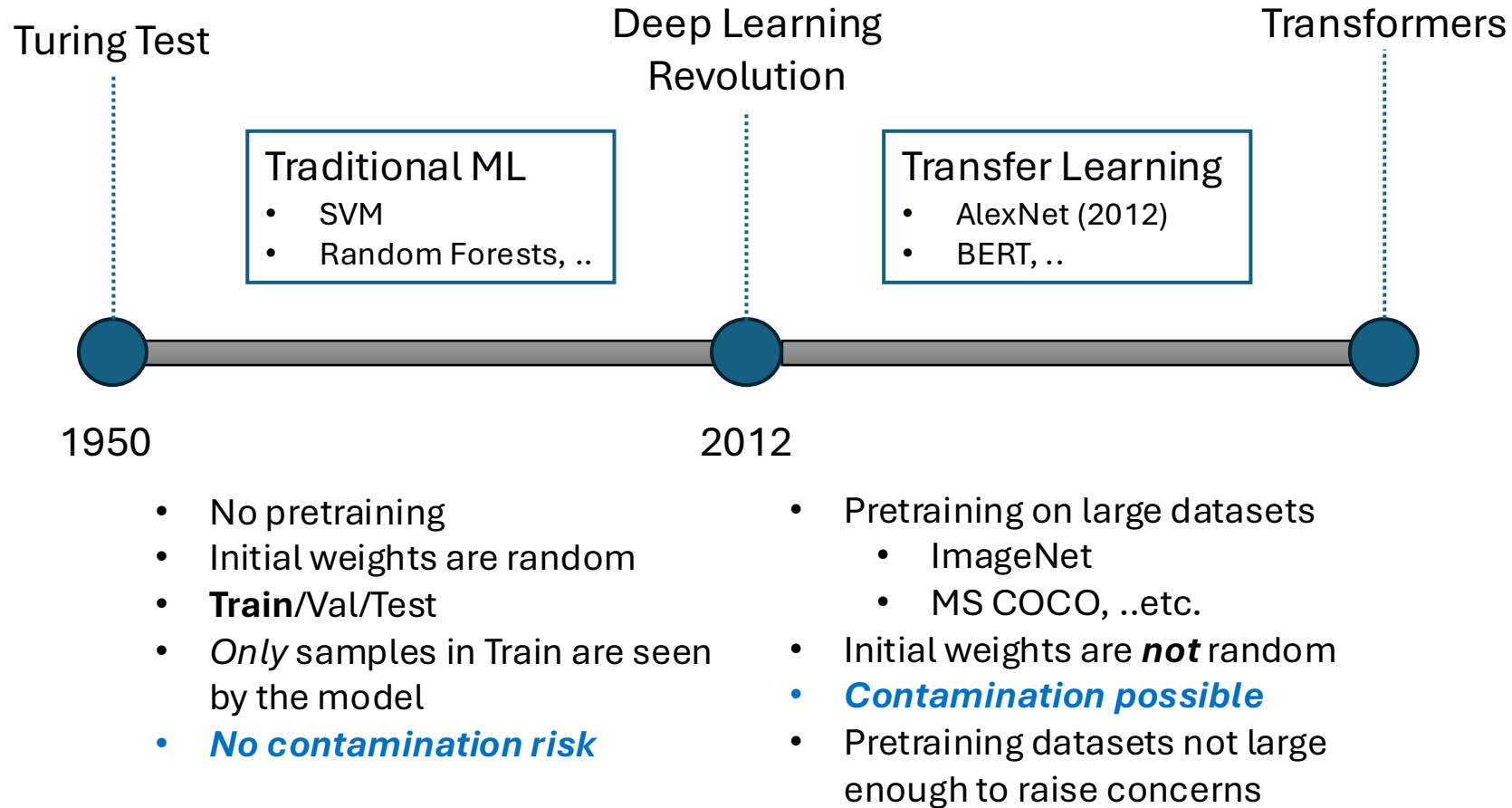
- Static datasets
- Snapshots of web env
- Most are task-specific

# Background

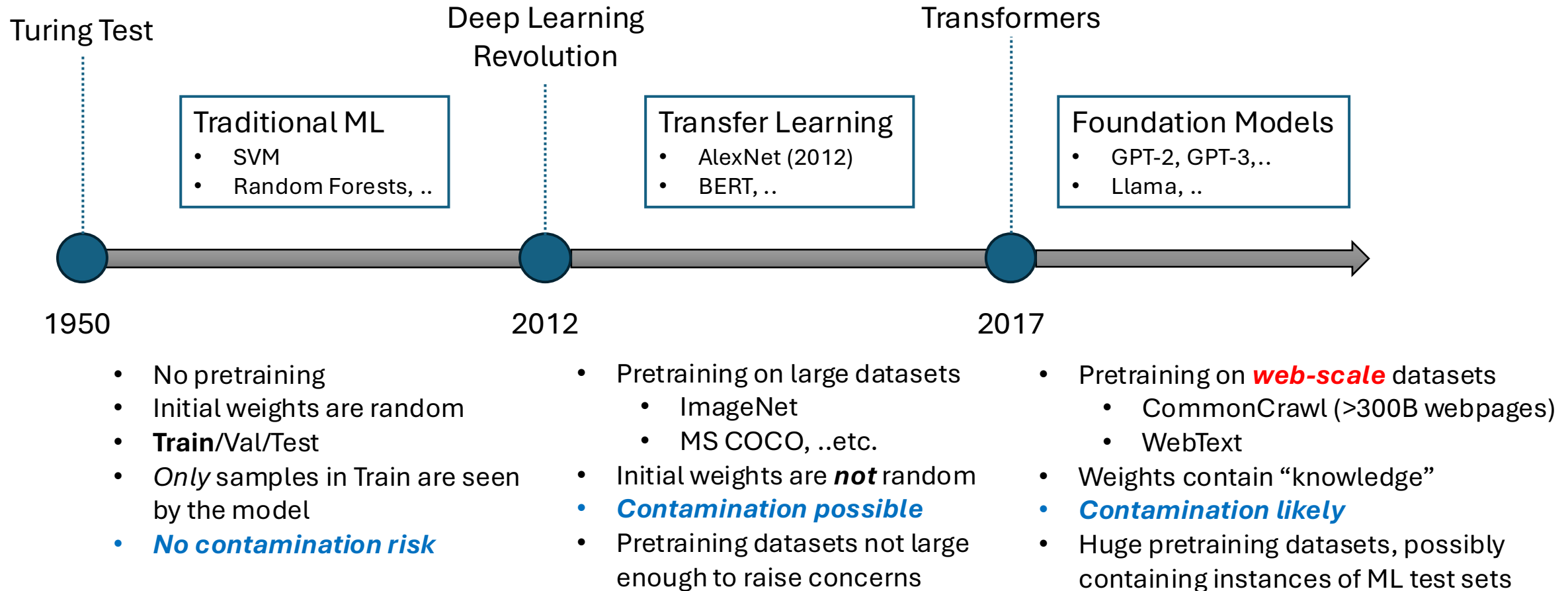


- **No pretraining**
- Initial weights are random noise
- **Train/Val/Test**
- *Only* samples in Train are seen by the model
- ***No contamination risk***

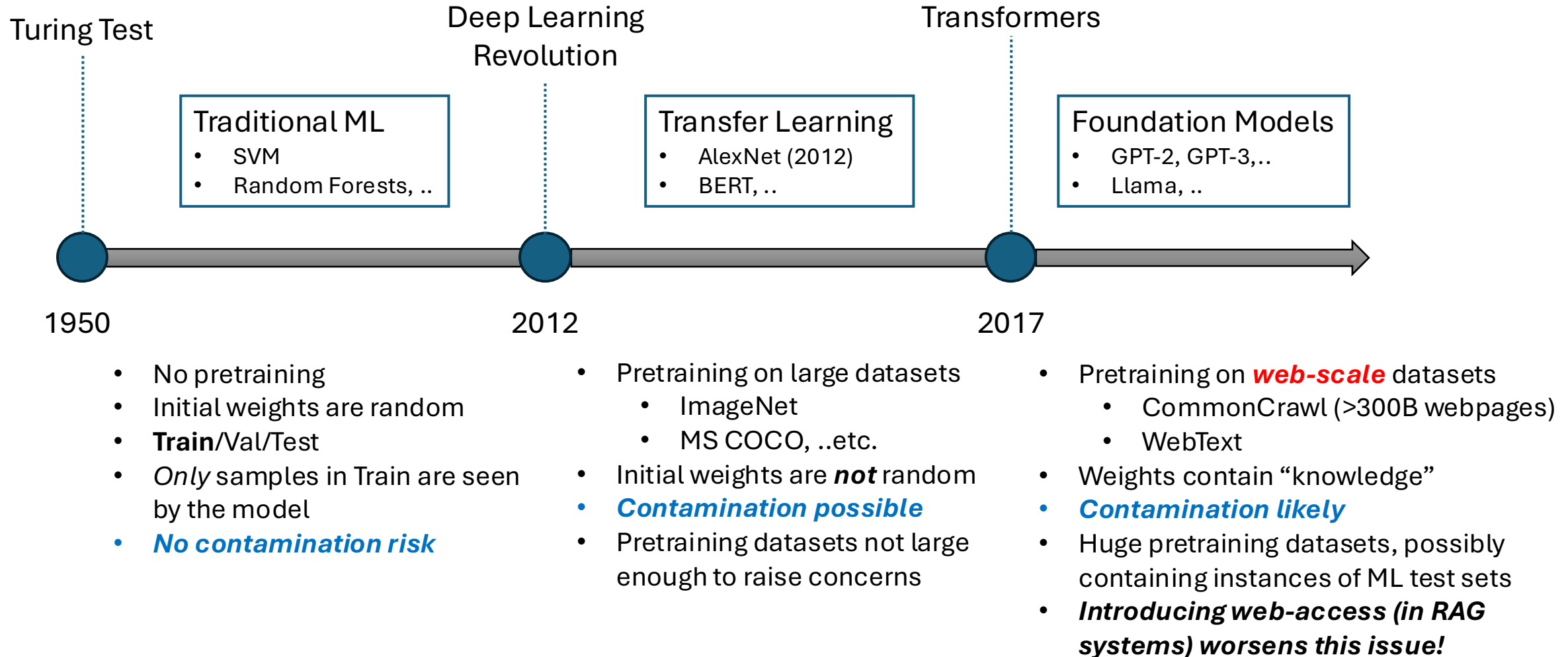
# Background



# Background



# Background



# Contamination Definition

- Following the formal definition of contamination in [1], we say that the dataset  $D_{eval}$  is **contaminated** by the dataset  $D_M$  if:

$$\exists x \in D_{eval}, \exists x' \in D_M, f(x, x') \geq \tau$$

- where  $f$  is a similarity function, and  $\tau$  a similarity threshold.  $\tau = 1$  corresponds to **verbatim** or **strict** contamination.
- $D_M$  is the dataset that model  $M$  has pre-trained on **or has access to**.

# Contamination in Web-enabled QA Agents

- In **RAG-like QA systems**, powered by LLMs, equipped with ***web access***, contamination can happen at two stages:
  1. **Pretraining Contamination:** occurs when the test sample is included in the pretraining corpus of the base LLM.
  2. **Retrieval Contamination:** occurs when the test sample is present in the retrieved results (from web engine).

# Examples from our findings

Datasets: bojanbabic/freshqa\_08182025 like 0

Split (2)  
test · 500 rows

what was the shortest

id	split	question
int64	string · classes	string · lengths
152	TEST	What was the shortest war in history?

## Pretraining Leak

## Leaked Source Indexed in C4 pretraining dataset: <https://www.britannica.com/event/Anglo-Zanzibar-War>

britannica.com/event/Anglo-Zanzibar-War

men Britannica Search Britannica... What was the shortest war in h 0/2

Anglo-Zanzibar War keyboard\_arrow\_left Question

Introduction  
Context  
Battle  
Aftermath  
References & Edit History  
Quick Facts & Related Topics

Images & Videos

Quizzes  
Pick Your Battles Quiz  
World Wars

and partition his land among the British and the Germans; later sultans followed suit. By 1890 Britain had convinced the sultanate to cede its mainland territories to [Germany](#) while allowing the island itself to become a [protectorate](#) of the British Empire. Britain agreed to preserve the sultanate as an institution, but it stressed that the power of the sultan was [nominal](#).

Upon becoming Zanzibar's [protector](#), Britain outlined two key policy goals: the abolition of [slavery](#) and the restoration of the island's strong mercantile economy. Consul General Gerald Portal's [economic policy](#) angered the Zanzibari merchants, but what angered them more was the policy on slavery, which had become [integral](#) to their way of life as spice plantations surged in value throughout the 19th century. Sayyid Ali ibn Sa'id resisted Britain's directives until his death in 1893, and, when it came time to choose his successor, Britain voiced its expectation that the next sultan be more submissive to imperial authority. Out of the several claimants to the throne, the British backed Ḥamad ibn Thuwayn. However, a defiant Prince Khālid ibn Barghash occupied the palace in response; he based his own claim on being the only son of the late Barghash and on

**What was the shortest war in history?**  
An episode of *Breaking Bad* has a longer run time than the Anglo-Zanzibar War.  
[See all videos for this article](#)

# Examples from our findings

## Question from HotPotQA:

“Which Hells Angel member stabbed and beat an attendant of the 1969 counterculture-era rock concert in the United States?”

## Retrieval Leak

## H Case Studies on LLMs Trained with Different Search Engines

Table 9: Case study A of LLM-based search agent trained with different search engines (*i.e.*, random noise, BM25, and E5).

---

**Question:** Which Hells Angel member stabbed and beat an attendant of the 1969 counterculture-era rock concert in the United States?

---

**Ground Truth:** Alan Passaro

---

**w. random noise:**

**<think>** To answer this question, we need to recall the incident involving the 1969 counterculture-era rock concert known as Woodstock. The incident in question is the killing of festival security guard Thomas Benson. The Hells Angel member responsible for the stabbing and subsequent death of Benson is CompareTag Stevenson. **</think>**

**<answer>** Manson **</answer>**

---

**w. BM25:**

**<think>** To answer this question, I'll need to identify the Hells Angel member who was involved in the stabbing and beating of an attendant at a 1969 counterculture-era rock concert in the United States. I'll use a search engine to gather information about this event and the responsible Hells Angel member. **</think>**

**<search>** Hells Angel member at 1969 counterculture-era rock concert in the United States **</search>**

**<information>** Doc 1(Title: "Altamont Free Concert") and for the negative portrayal of the Angels in the "'Gimme Shelter'" film. The conspirators reportedly used a boat to approach a residence where Jagger was

***Leaked Source Indexed in Search Engines:***

<https://arxiv.org/pdf/2505.15117>

# Why is this a problem?

- Contamination ***inflates performance metrics***.
- In extreme cases, it becomes synonymous to “training on test data”.
- Instead of evaluating generalization or “true capabilities”, it becomes a test of memorization or lookup capabilities<sup>[1]</sup>.
- Even simple paraphrasing can drop performance <sup>[2]</sup>.

[1] Morris, John X., et al. "How much do language models memorize?." arXiv preprint arXiv:2505.24832 (2025).

[2] Zhu, Kaijie, et al. "Dynamic evaluation of large language models by meta probing agents." arXiv preprint arXiv:2402.14865 (2024).

# Potential Solutions

- Hiding labels/groundtruth for test sets
  - Examples: GAIA, SQUAD 2.0, ..etc.
- Evaluation Sandbox Environments
  - AgentSims, MLGym-bench , ..etc.
- Dynamic Benchmarks
  - DARG, DyVal, DynamicKGQA, iAgentBench, ..etc.

# Dynamic Benchmarking

- Dynamic benchmarking is an evaluation paradigm that **generates or adapts test samples at evaluation time**, so models are scored on fresh, unseen data and test-set contamination is avoided [1,2,3].

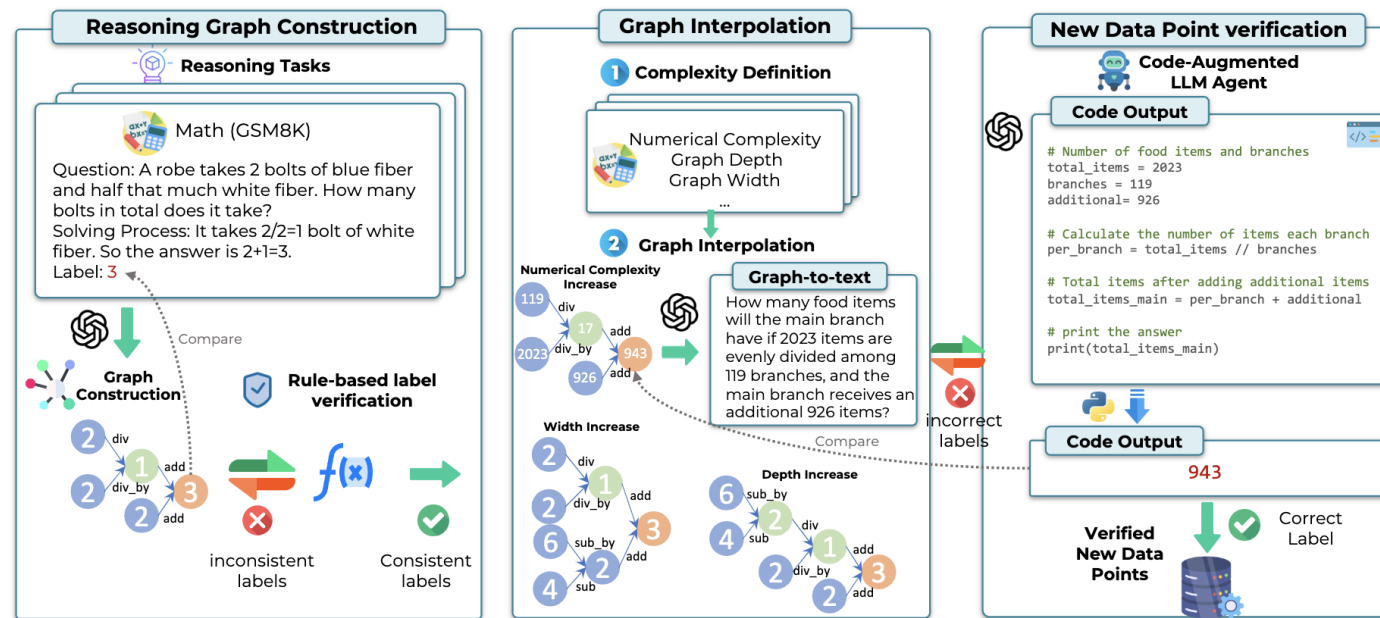


Fig: Representative Example for Dynamic Benchmarking [2]

[1] Zhu, Kaijie, et al. "Dyval: Dynamic evaluation of large language models for reasoning tasks." *arXiv preprint arXiv:2309.17167* (2023).

[2] Zhang, Zhehao, Jiaao Chen, and Diyi Yang. "Darg: Dynamic evaluation of large language models via adaptive reasoning graph." *arXiv preprint arXiv:2406.17271* (2024).

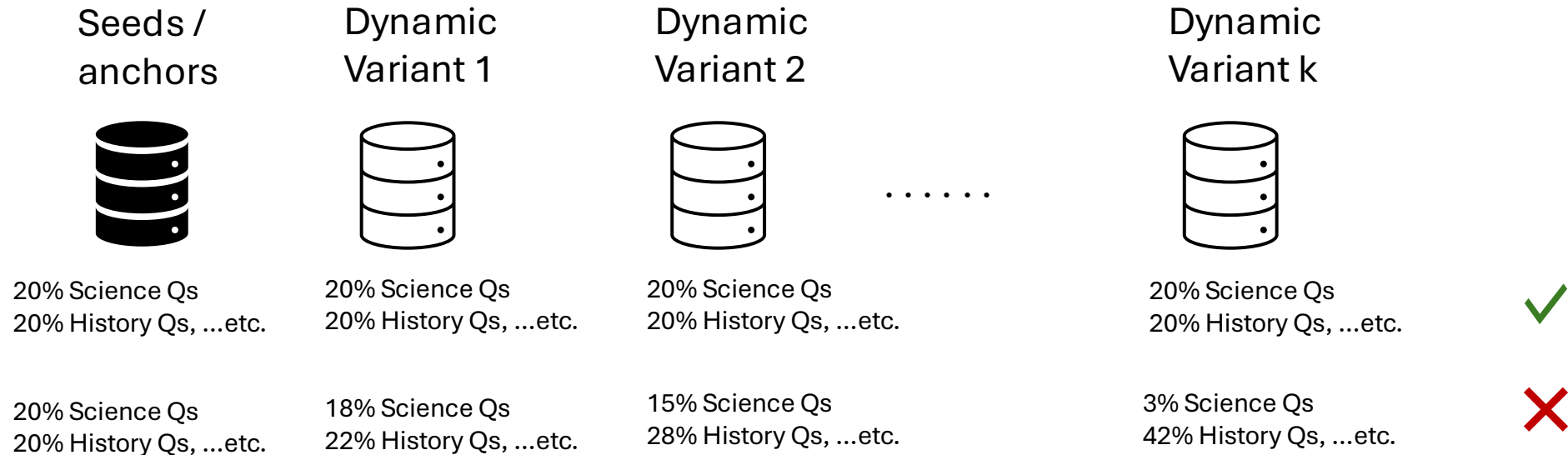
[3] Dammu, Preetam Prabhu Srikar, et al. "Dynamic-KGQA: A Scalable Framework for Generating Adaptive Question Answering Datasets." *arXiv preprint arXiv:2503.05049* (2025).

# Dynamic Benchmarking

- Desirable Characteristics
  - Distributional consistency across rounds
  - Stable or controllable complexity

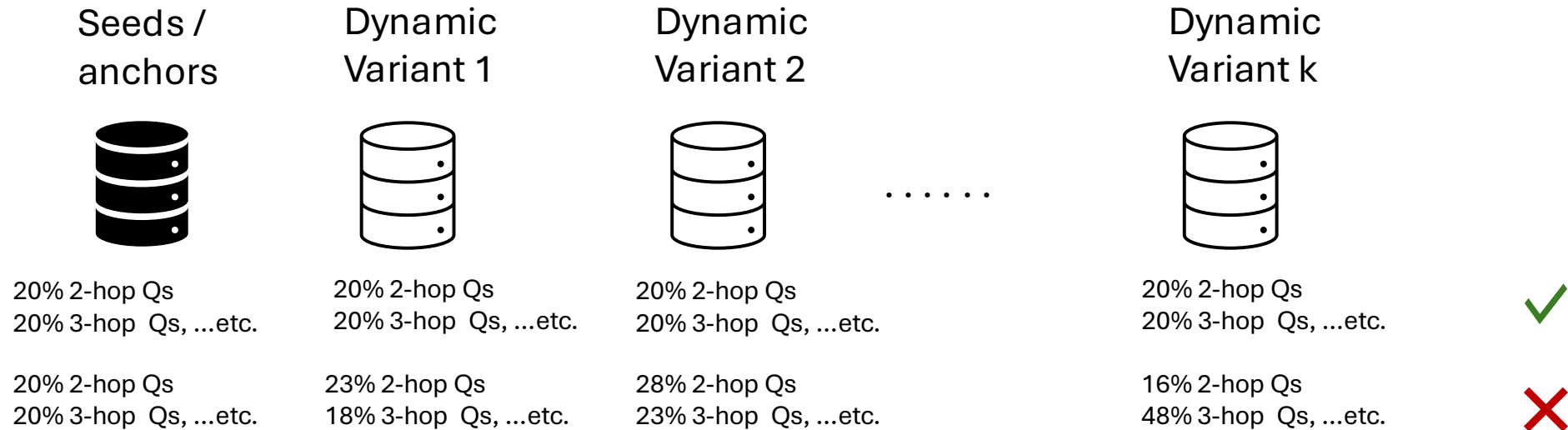
# Dynamic Benchmarking

- Desirable Characteristics
  - **Distributional consistency across rounds**
  - Stable or controllable complexity



# Dynamic Benchmarking

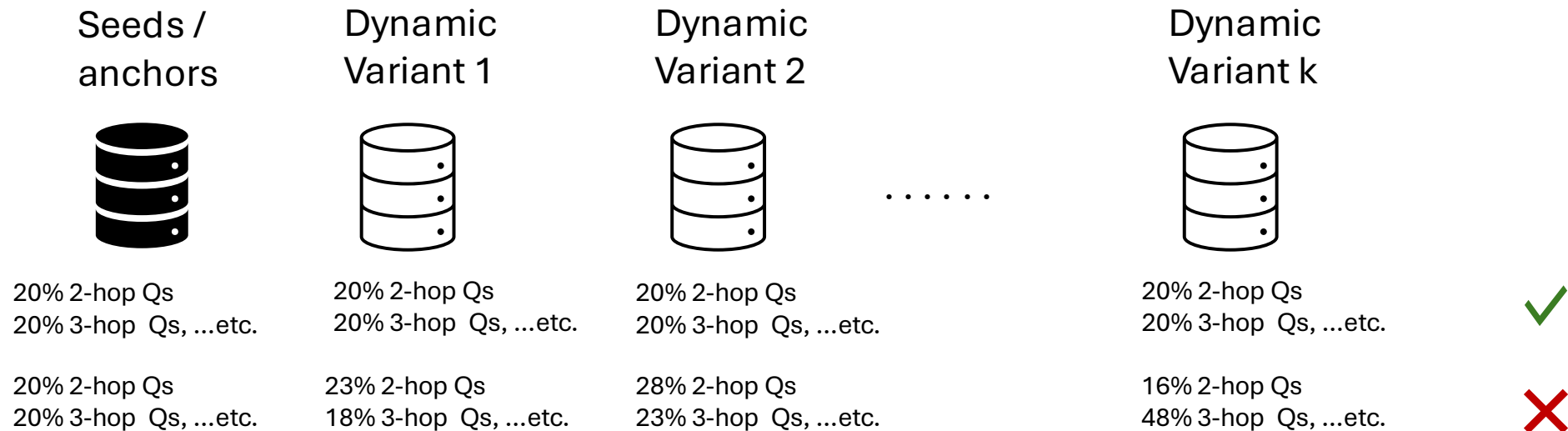
- Desirable Characteristics
  - Distributional consistency across rounds
  - **Stable or controllable complexity**



# Dynamic Benchmarking

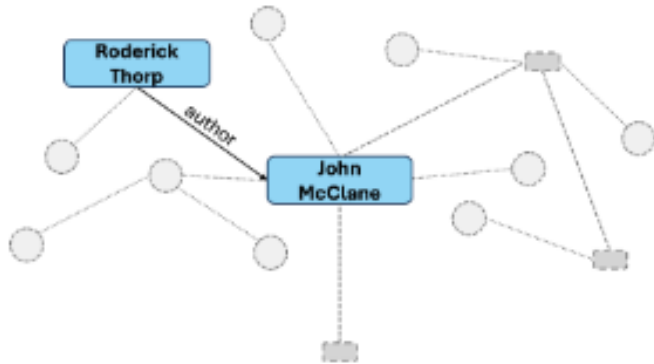
- Desirable Characteristics

- Distributional consistency across rounds: **achieved through seeds/anchors**
- Stable or controllable complexity: **achieved through generation parameters**



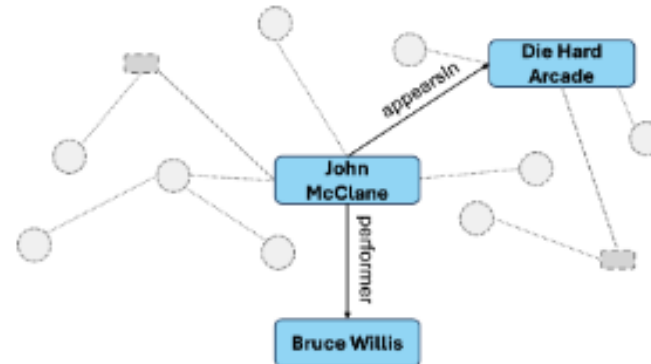
# Dynamic-KGQA

Q: Who is the author of the character John McClane?  
A: Roderick Thorp



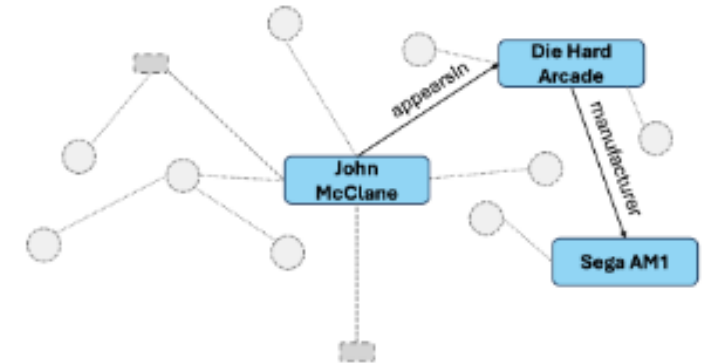
(a) Answer Path 1

Q: Which video game features Bruce Willis's famous character?  
A: Die Hard Arcade



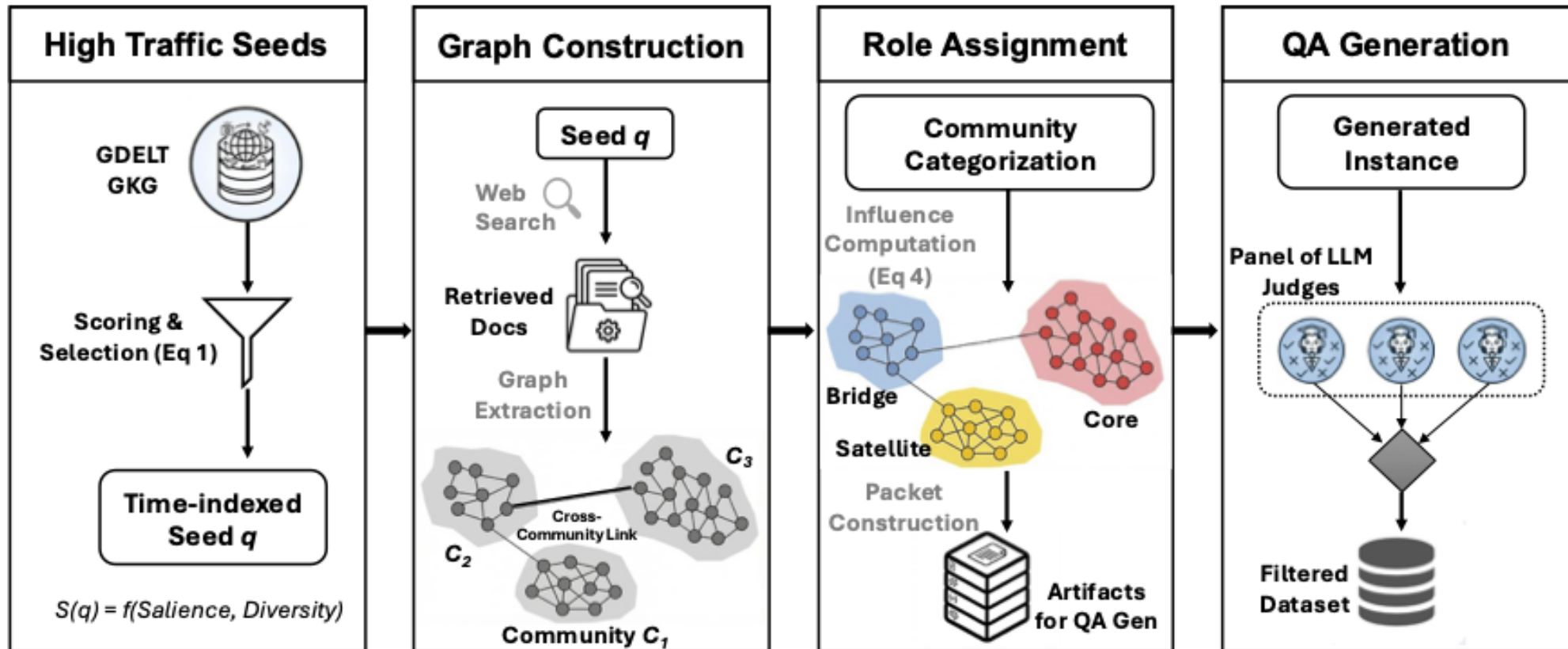
(b) Answer Path 2

Q: Which company manufactured the game in which John McClane appears?  
A: Sega AM1



(c) Answer Path 3

# iAgentBench



Dammu, P. P. S., Palkhiwala, A., Roosta, T., & Shah, C. (2026). iAgentBench: Benchmarking Sensemaking Capabilities of Information-Seeking Agents on High-Traffic Topics. *arXiv preprint arXiv:2603.04656*.

# Closing

- Please reach out to [preetams@uw.edu](mailto:preetams@uw.edu) with any questions.
- If you found these slides useful, please consider citing:

Dammu, P. P. S., & Roosta, T. (2026, March). Information Seeking in the Age of Agentic AI: A Half-Day Tutorial. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval* (pp. 429-430).

THANK YOU!