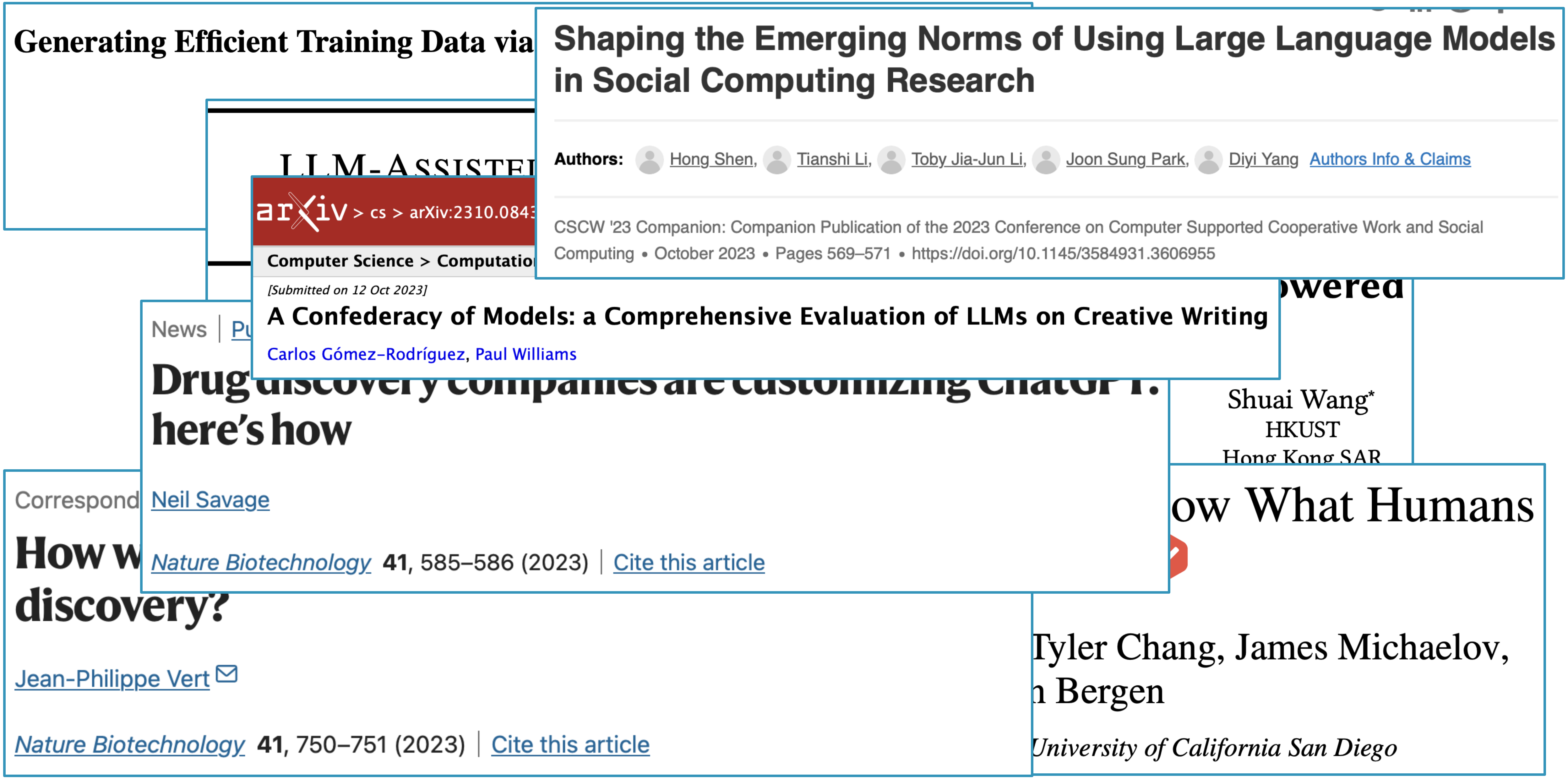
From Relevance to Reality:
Scaling Human-Centered Evaluation in the LLM Era

Chirag Shah

RAISE
Responsibility in AI Systems & Experiences

InfoSeeking

University of Washington



Generating Efficient Training Data via

LLM-Assistei

owered

Shuai Wang*
HKUST
Hong Kong SAR

Shaping the Emerging Norms of Using Large Language Models in Social Computing Research

Authors: Hong Shen, Tianshi Li, Toby Jia-Jun Li, Joon Sung Park, Diyi Yang Authors Info & Claims

CSCW '23 Companion: Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing • October 2023 • Pages 569–571 • https://doi.org/10.1145/3584931.3606955

arXiv > cs > arXiv:2310.084

Computer Science > Computatio

[Submitted on 12 Oct 2023]

A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing

Carlos Gómez-Rodríguez, Paul Williams

News | Pu

Drug discovery companies are customizing ChatGPT: here's how

Neil Savage

Nature Biotechnology 41, 585–586 (2023) | Cite this article

Correspond

How w discovery?

Jean-Philippe Vert

Nature Biotechnology 41, 750–751 (2023) | Cite this article

ow What Humans

Tyler Chang, James Michaelov,
n Bergen

University of California San Diego

Large Language Models can Accurately Predict Searcher Preferences

Paul Thomas
Microsoft
Adelaide, Australia
pathom@microsoft.com

Nick Craswell
Microsoft
Seattle, USA
nickcr@microsoft.com

LLM-Evaluation Tropes: Perspectives on the Validity of LLM-Evaluations

Laura Dietz
University of New Hampshire
USA

Oleg Zendel
RMIT University
Australia

Peter Bailey
Canva
Australia

Charles Clarke
University of Waterloo
Canada

Ellese Cotterill
Canva
Australia

Jeff Dalton
University of Edinburgh
United Kingdom

Faegheh Hasibi
Radboud University
Netherlands

Mark Sanderson
RMIT University
Australia

Nick Craswell
Microsoft
USA

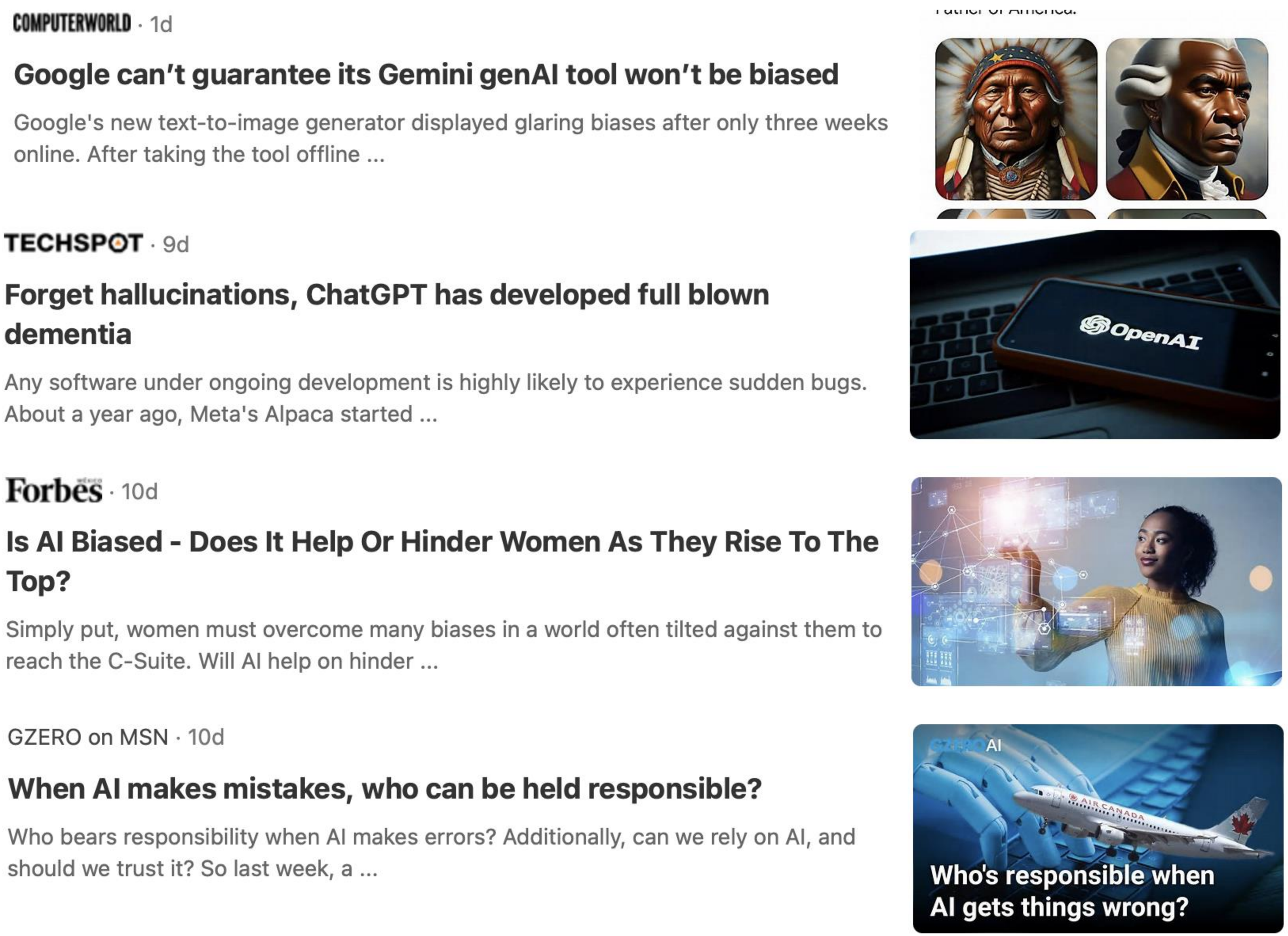
LLM-Driven Usefulness Labeling for IR Evaluation

Mouly Dewan
University of Washington
Seattle, WA, United States
mdewan@uw.edu

Jiqun Liu
The University of Oklahoma
Norman, OK, United States
jiqunliu@ou.edu

Chirag Shah
University of Washington
Seattle, WA, United States
chirags@uw.edu

But…



COMPUTERWORLD · 1d

Google can't guarantee its Gemini genAI tool won't be biased

Google's new text-to-image generator displayed glaring biases after only three weeks online. After taking the tool offline ...

TECHSPOT · 9d

Forget hallucinations, ChatGPT has developed full blown dementia

Any software under ongoing development is highly likely to experience sudden bugs. About a year ago, Meta's Alpaca started ...

Forbes México · 10d

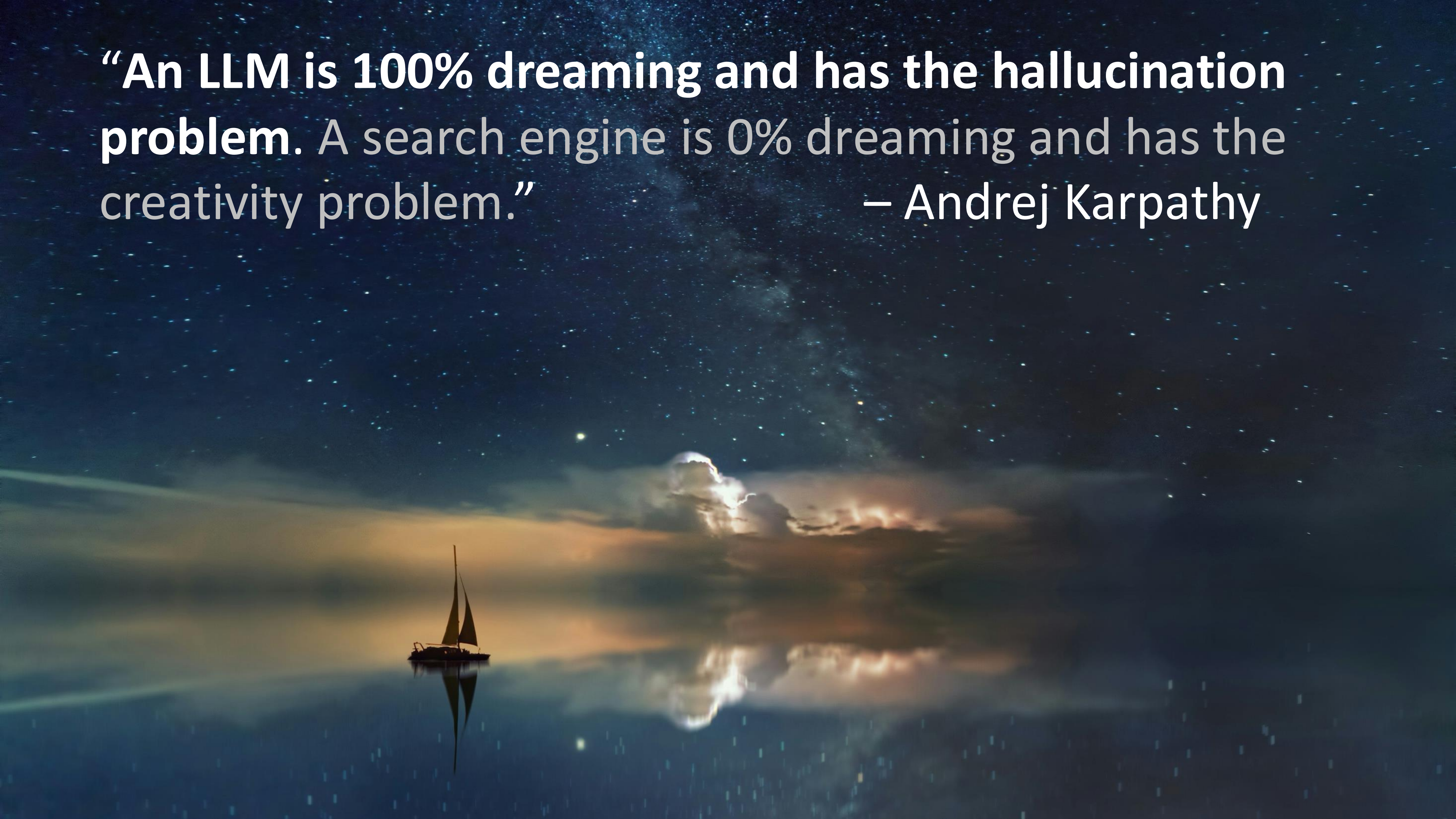
Is AI Biased - Does It Help Or Hinder Women As They Rise To The Top?

Simply put, women must overcome many biases in a world often tilted against them to reach the C-Suite. Will AI help on hinder ...

GZERO on MSN · 10d

When AI makes mistakes, who can be held responsible?

Who bears responsibility when AI makes errors? Additionally, can we rely on AI, and should we trust it? So last week, a ...



“**An LLM is 100% dreaming and has the hallucination problem**. A search engine is 0% dreaming and has the creativity problem.” – Andrej Karpathy

What Do We Want?

We want that **creativity**. Can we do something about its **~~hallucinations~~**?

reliability

trustworthiness

How do we turn **engineering** to **science**?

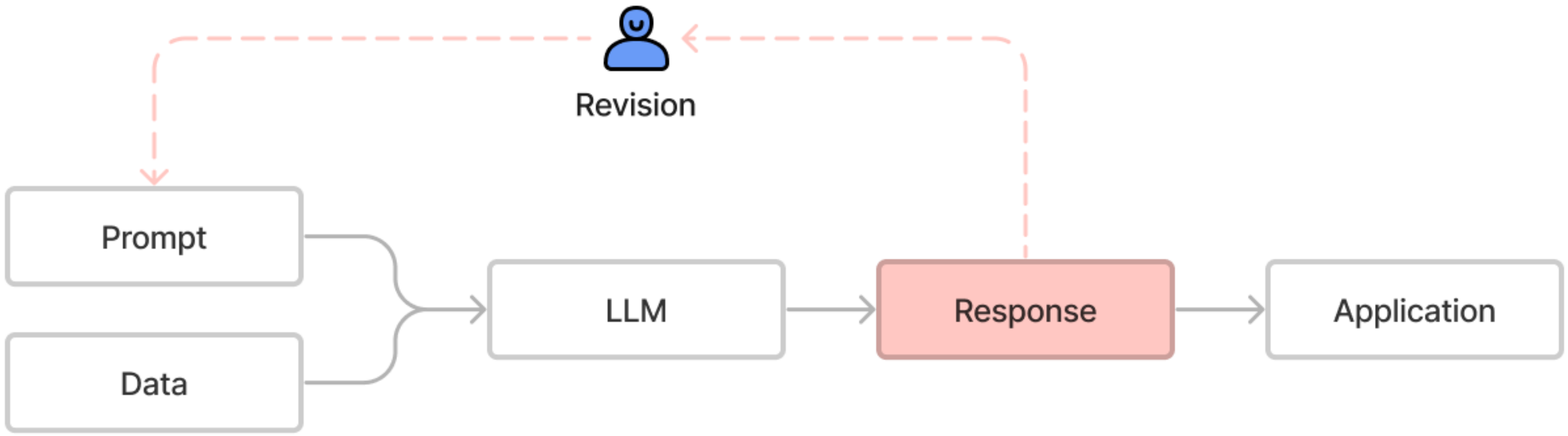
↓

Applying to same level of scrutiny we apply to any scientific instrument.

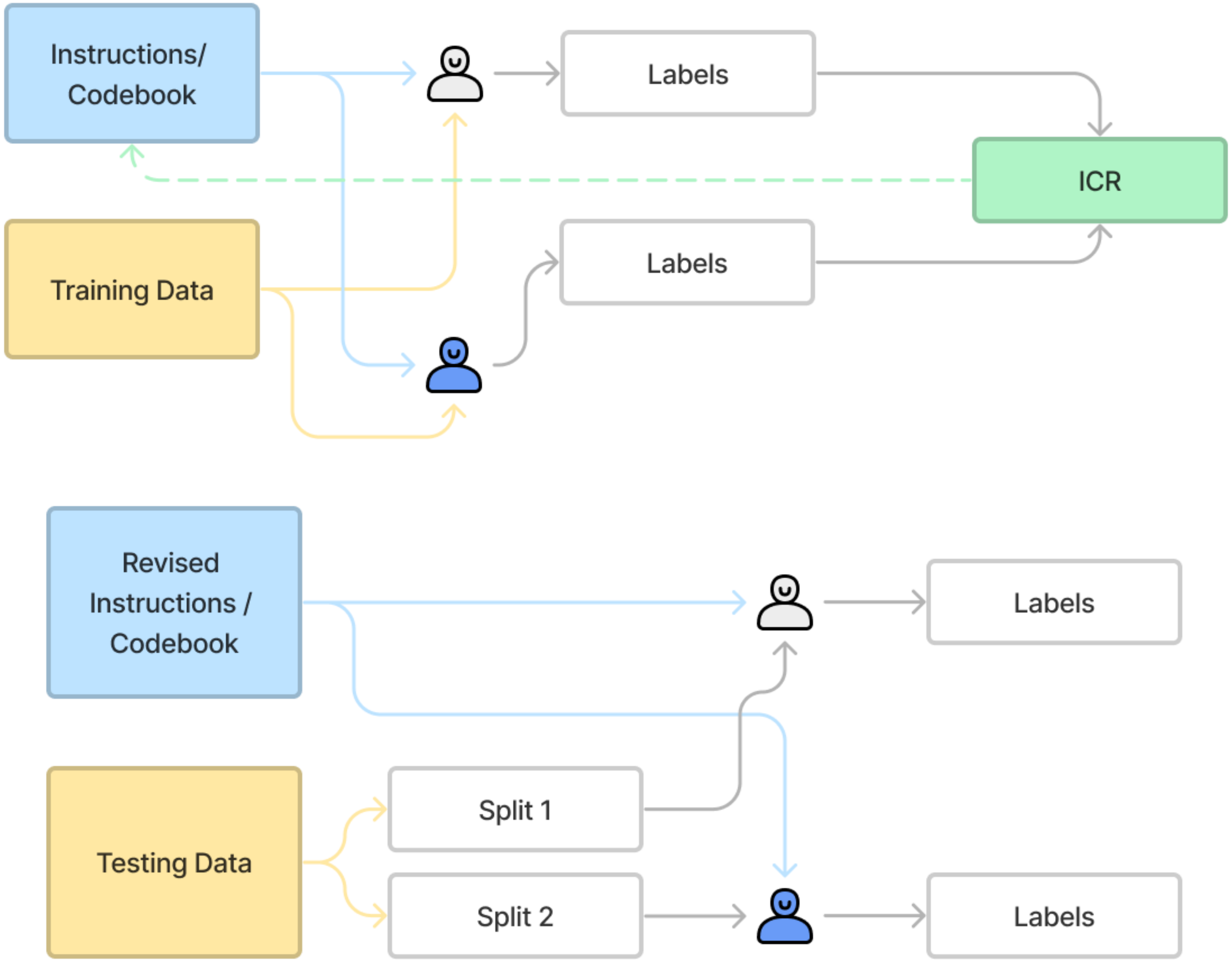
PROMPT SCIENCE

Shah (2025). From Prompt Engineering to Prompt Science. *Communications of the ACM (CACM).*

The Ad-hocness of Prompt Engineering



Qualitative Coding



Instructions/ Codebook

Labels

ICR

Training Data

Labels

Revised Instructions / Codebook

Labels

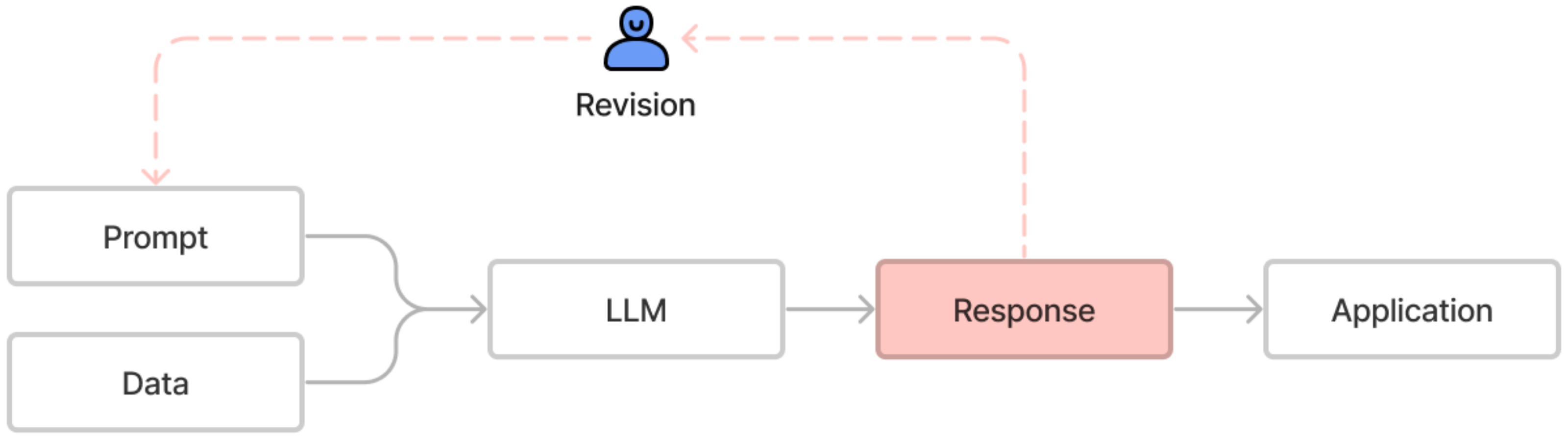
Testing Data

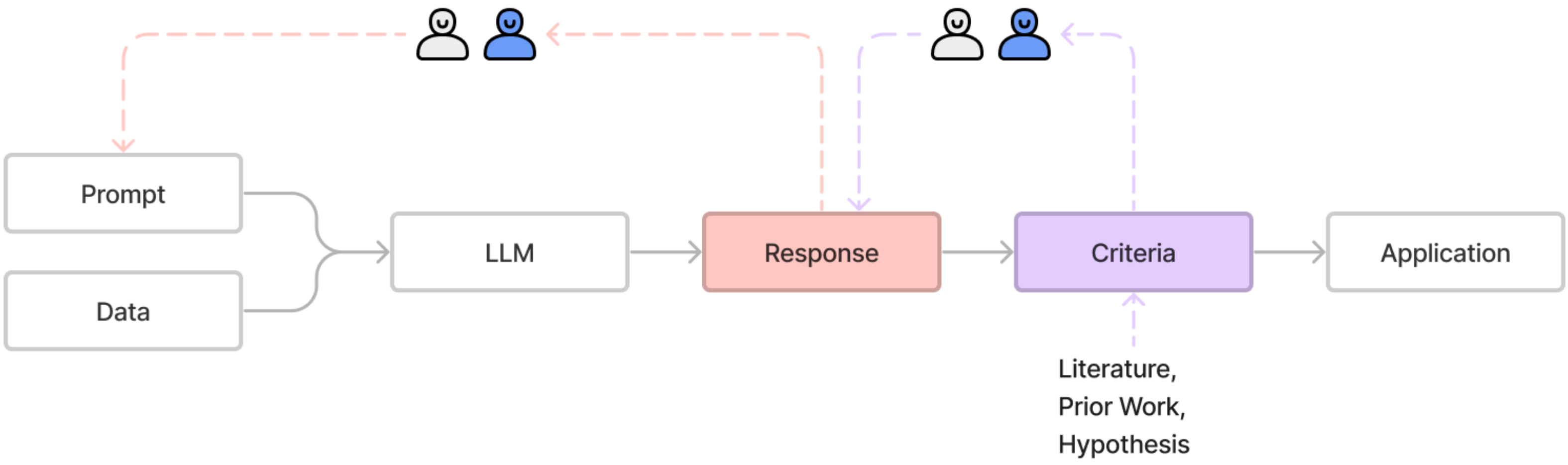
Split 1

Split 2

Labels

Turning Prompt Engineering to Prompt Science



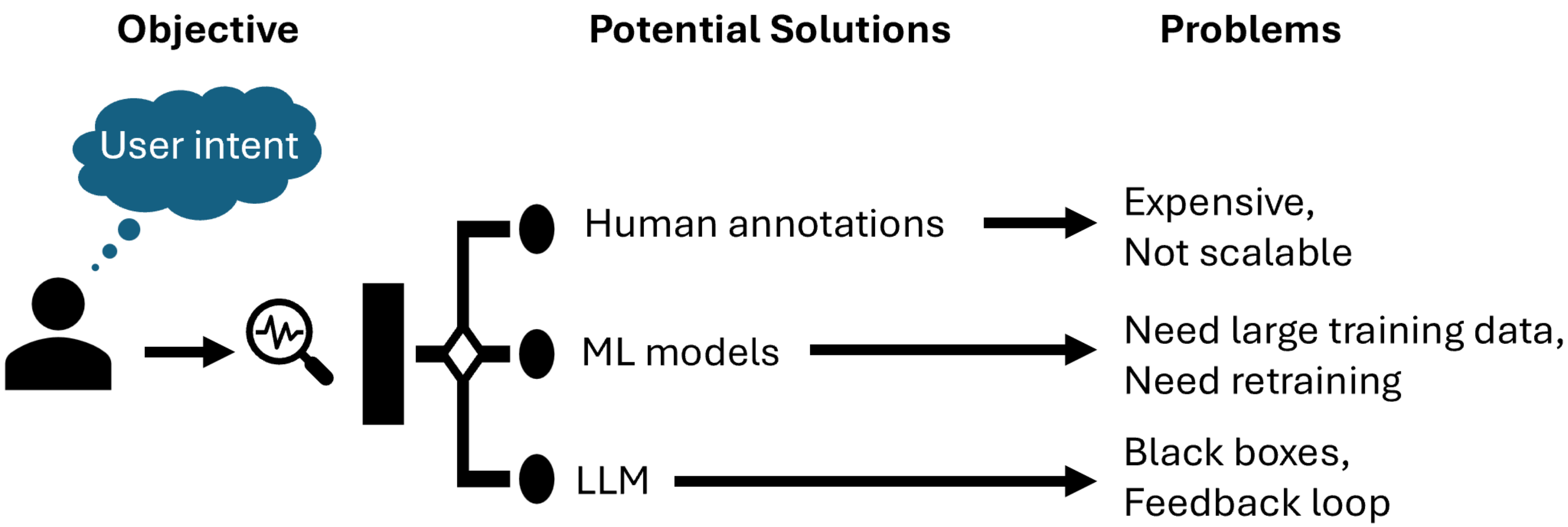


Case Study 1

SEARCH VS. CHAT

Shah et al. (2025). Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies. *ACM Transactions of the Web (TWeb).*

Taxonomy Generation and Log Analysis



Objective

User intent

Potential Solutions

Human annotations

ML models

LLM

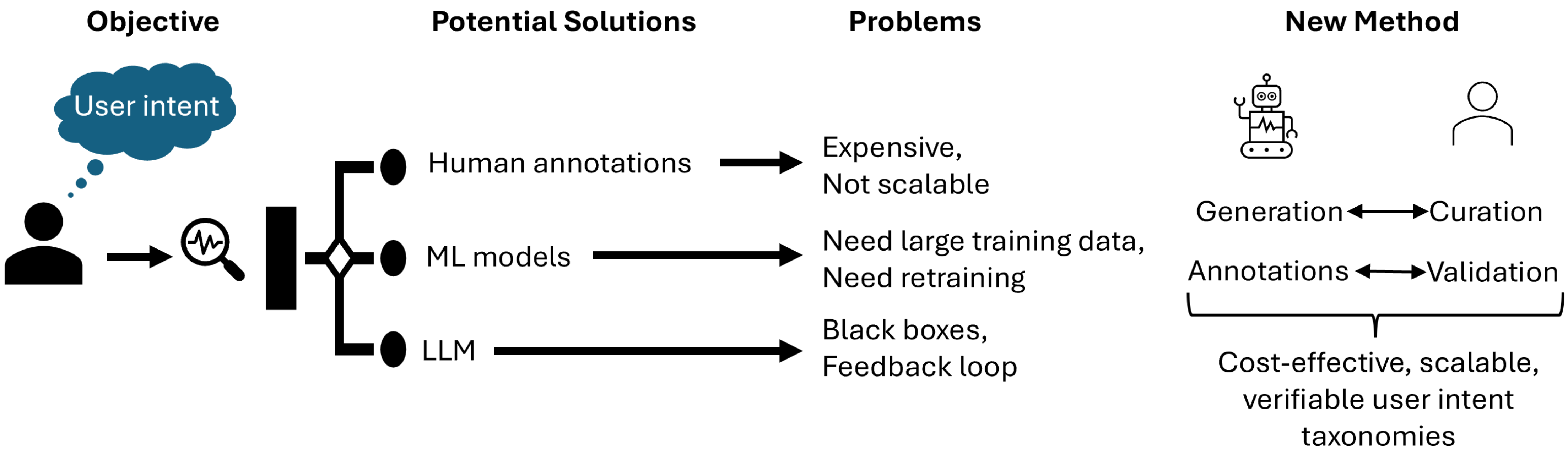
Problems

Expensive,
Not scalable

Need large training data,
Need retraining

Black boxes,
Feedback loop

Taxonomy Generation and Log Analysis



Objective

User intent

Potential Solutions

Human annotations

ML models

LLM

Problems

Expensive,
Not scalable

Need large training data,
Need retraining

Black boxes,
Feedback loop

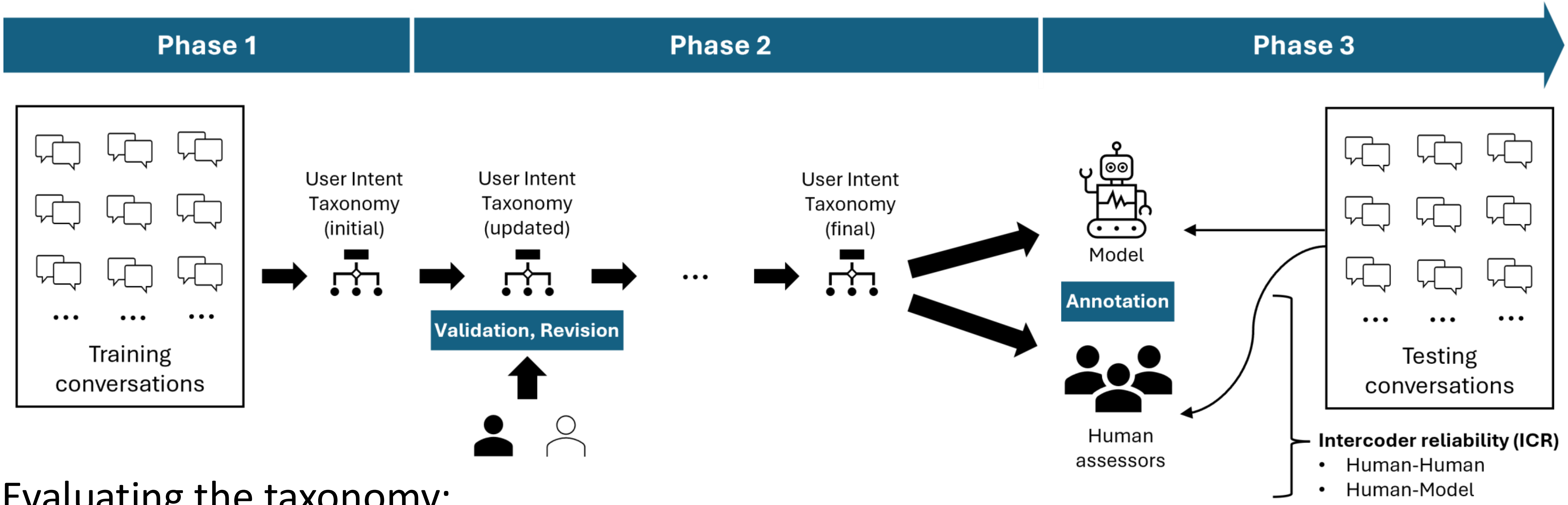
New Method

Generation ⟷ Curation

Annotations ⟷ Validation

Cost-effective, scalable,
verifiable user intent
taxonomies

Taxonomy Generation and Log Analysis

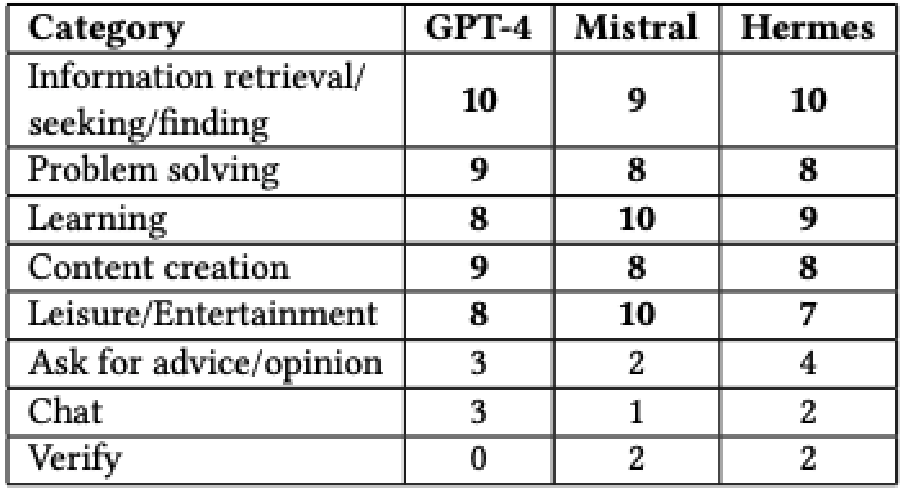


Evaluating the taxonomy:

- Comprehensiveness
- Consistency
- Clarity
- Accuracy
- Conciseness

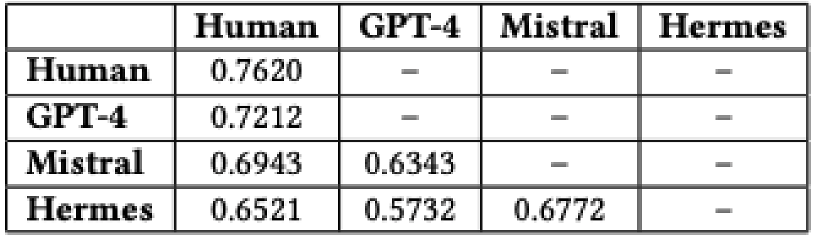
Generating and Validating LLM-based Taxonomies

Table 5: Bootstrapping experiments showing frequency of different intent categories over 30 total runs, 10 runs for each of the three LLMs. Top 5 categories in each are bolded.



Category	GPT-4	Mistral	Hermes
Information retrieval/ seeking/finding	**10**	**9**	**10**
Problem solving	**9**	**8**	**8**
Learning	**8**	**10**	**9**
Content creation	**9**	**8**	**8**
Leisure/Entertainment	**8**	**10**	**7**
Ask for advice/opinion	3	2	4
Chat	3	1	2
Verify	0	2	2

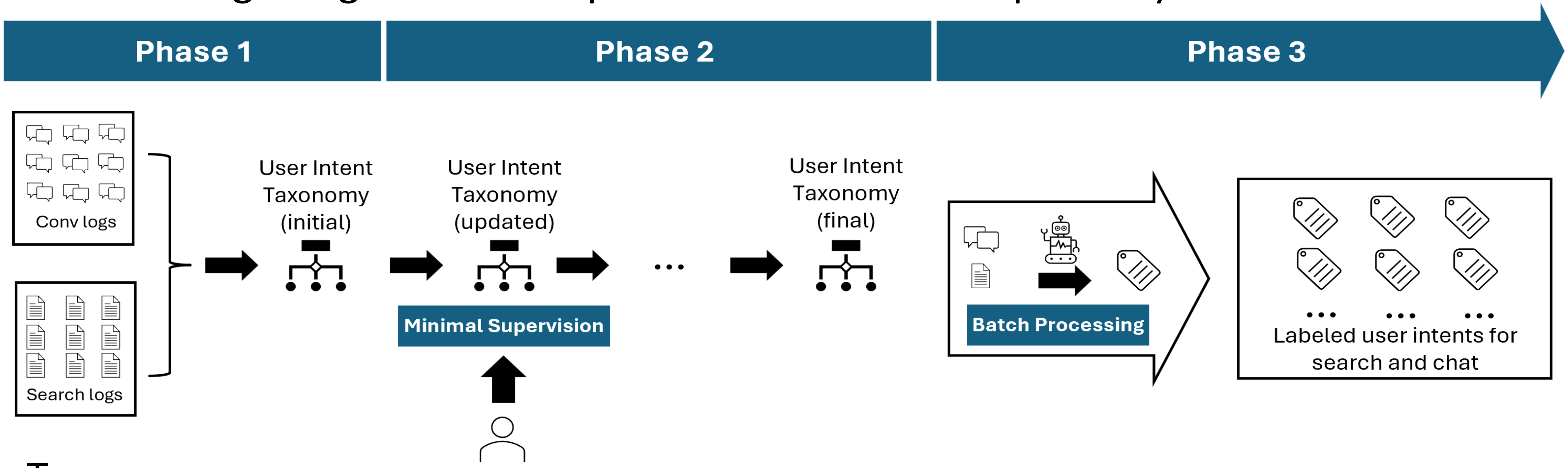
Table 6: ICR using Cohen's Kappa.



	Human	**GPT-4**	**Mistral**	**Hermes**
Human	0.7620	–	–	–
GPT-4	0.7212	–	–	–
Mistral	0.6943	0.6343	–	–
Hermes	0.6521	0.5732	0.6772	–

Taxonomy Generation and Log Analysis

Training using 500 search queries and 500 chat requests by the same users

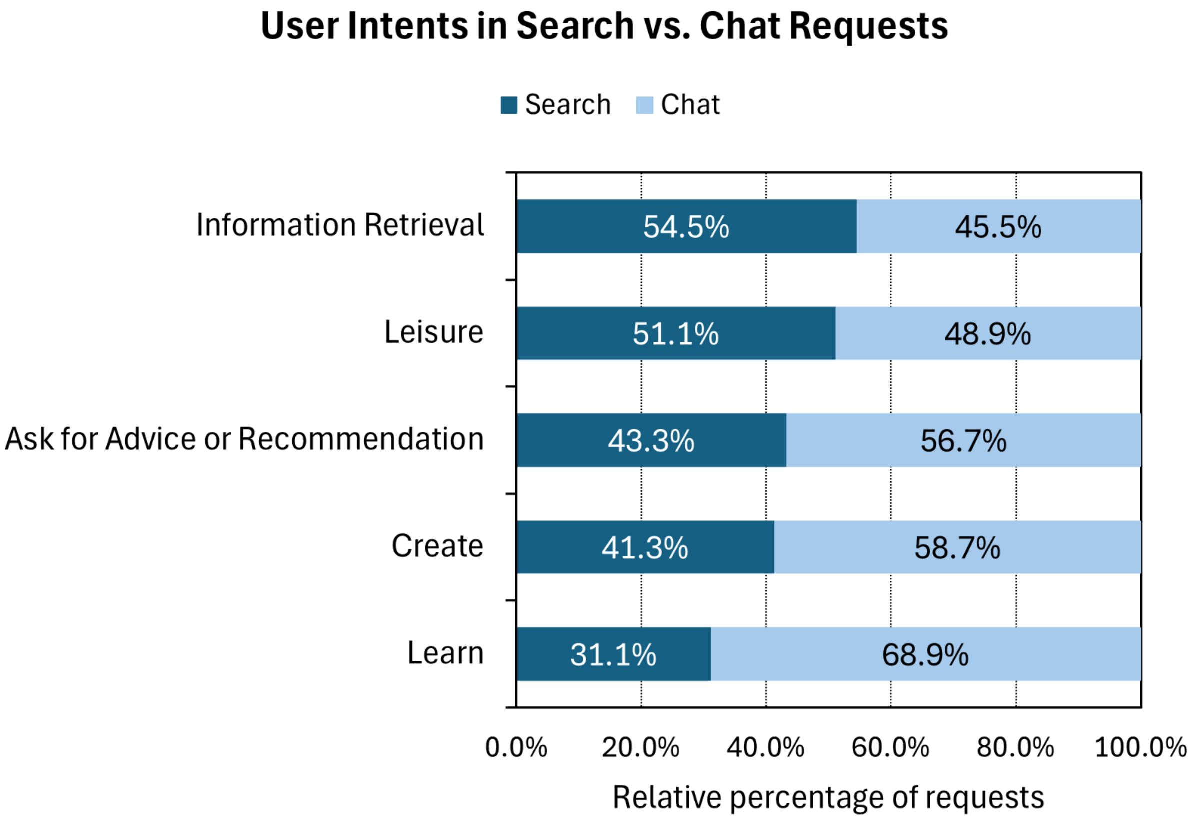

Phase 1
Phase 2
Phase 3
Conv logs
Search logs
User Intent Taxonomy (initial)
User Intent Taxonomy (updated)
Minimal Supervision
…
User Intent Taxonomy (final)
Batch Processing
Labeled user intents for search and chat

Taxonomy:

- Ask for advice/recommendation
- Create
- Information retrieval
- Learn
- Leisure

Taxonomy Generation and Log Analysis

Annotations on 1,956 search queries and 15,031 chat requests by the same users

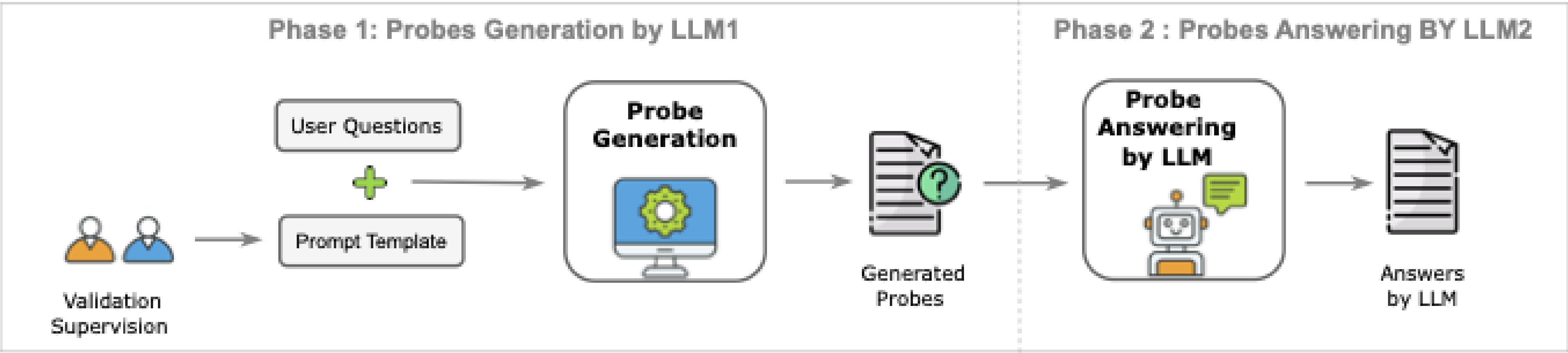


Case Study 2

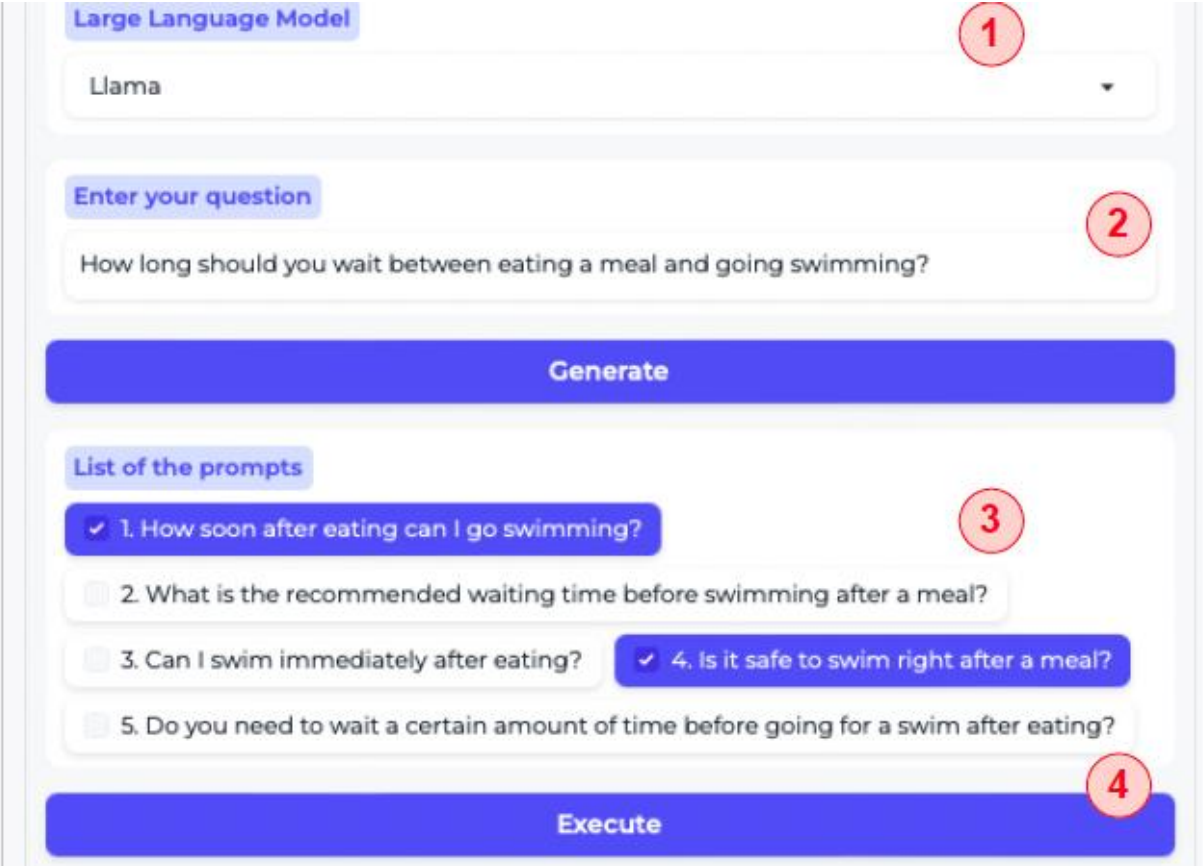
AUDITING AN LLM

Amiritziani et al. (2024). AuditLLM: A Tool for Auditing Large Language Models Using Multiprobe Approach. *Proceedings of ACM CIKM 2024*. Boise, ID. October 2024.

Multiprobe Auditing



Multiprobe Auditing



Large Language Model

Llama

Enter your question

How long should you wait between eating a meal and going swimming?

List of the prompts

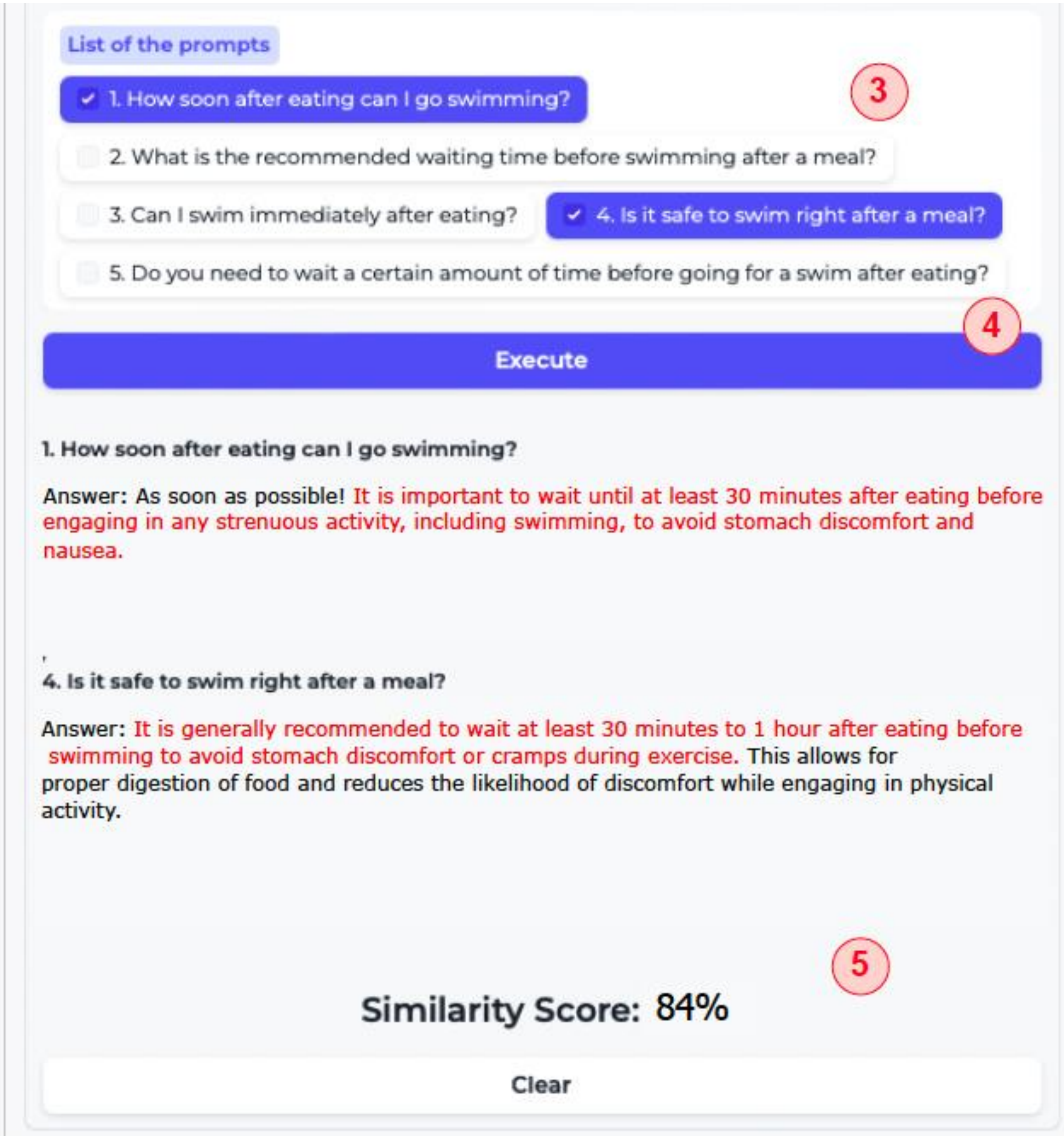
☑ 1. How soon after eating can I go swimming?

□ 2. What is the recommended waiting time before swimming after a meal?

□ 3. Can I swim immediately after eating?

☑ 4. Is it safe to swim right after a meal?

□ 5. Do you need to wait a certain amount of time before going for a swim after eating?



List of the prompts

☑ 1. How soon after eating can I go swimming?

□ 2. What is the recommended waiting time before swimming after a meal?

□ 3. Can I swim immediately after eating?

☑ 4. Is it safe to swim right after a meal?

□ 5. Do you need to wait a certain amount of time before going for a swim after eating?

1. How soon after eating can I go swimming?

Answer: As soon as possible! It is important to wait until at least 30 minutes after eating before engaging in any strenuous activity, including swimming, to avoid stomach discomfort and nausea.

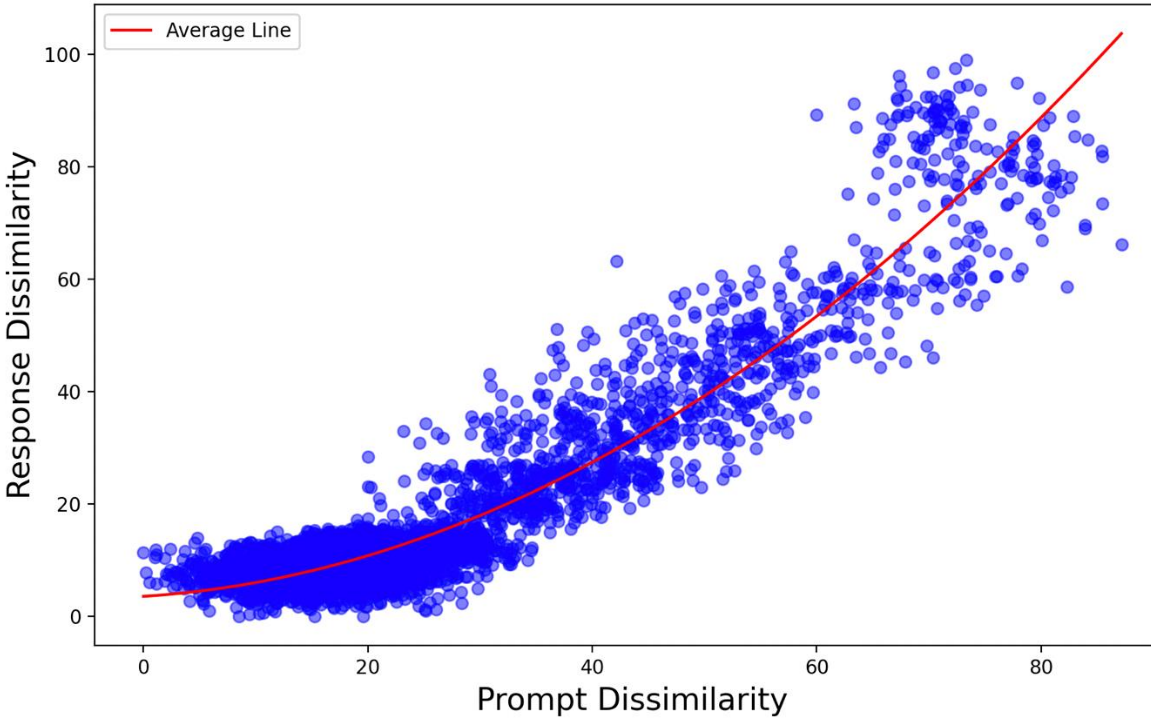
4. Is it safe to swim right after a meal?

Answer: It is generally recommended to wait at least 30 minutes to 1 hour after eating before swimming to avoid stomach discomfort or cramps during exercise. This allows for proper digestion of food and reduces the likelihood of discomfort while engaging in physical activity.

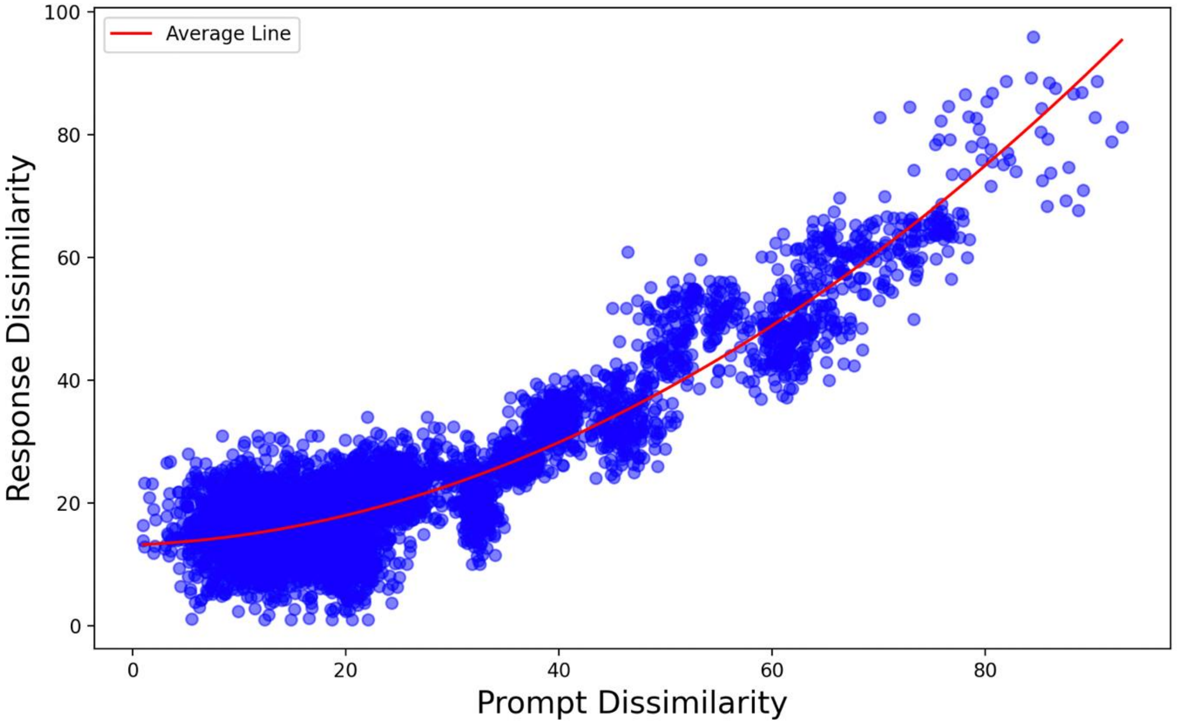
Similarity Score: 84%

Multiprobe Auditing

- Data: TruthfulQA with 817 questions spanning 38 diverse categories, including health, law, finance, and politics



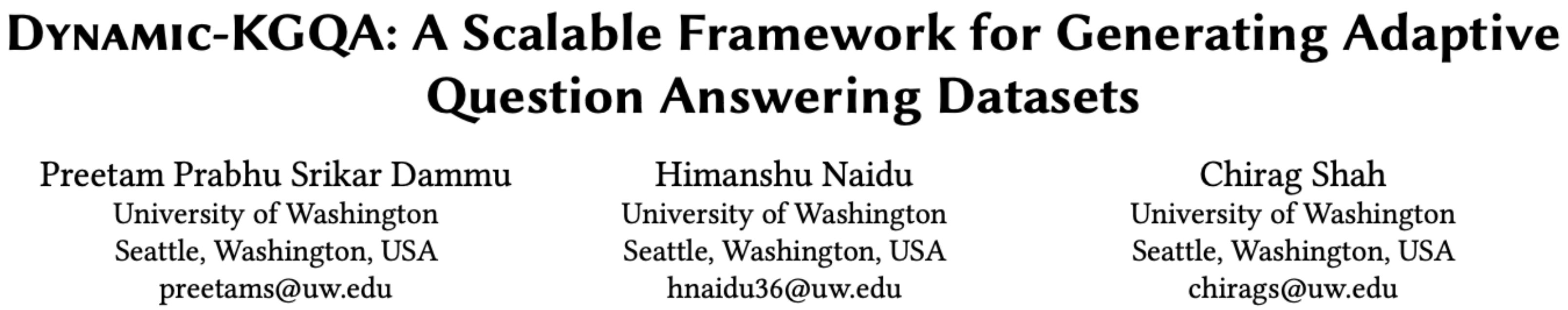
Falcon



Llama 2

Are You Using The Right Benchmarks?

- Question your benchmarks. Data contaminations happen.

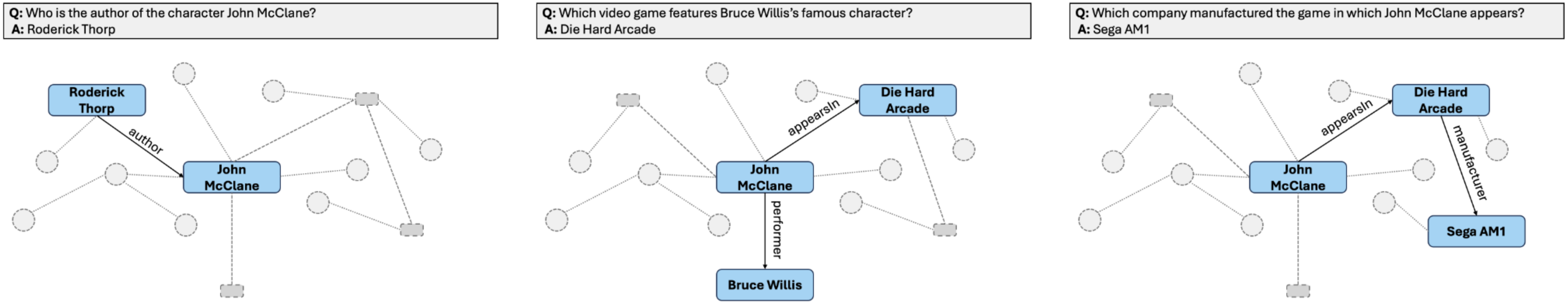


DYNAMIC-KGQA: A Scalable Framework for Generating Adaptive Question Answering Datasets

Preetam Prabhu Srikar Dammu
University of Washington
Seattle, Washington, USA
preetams@uw.edu

Himanshu Naidu
University of Washington
Seattle, Washington, USA
hnaidu36@uw.edu

Chirag Shah
University of Washington
Seattle, Washington, USA
chirags@uw.edu



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

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

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

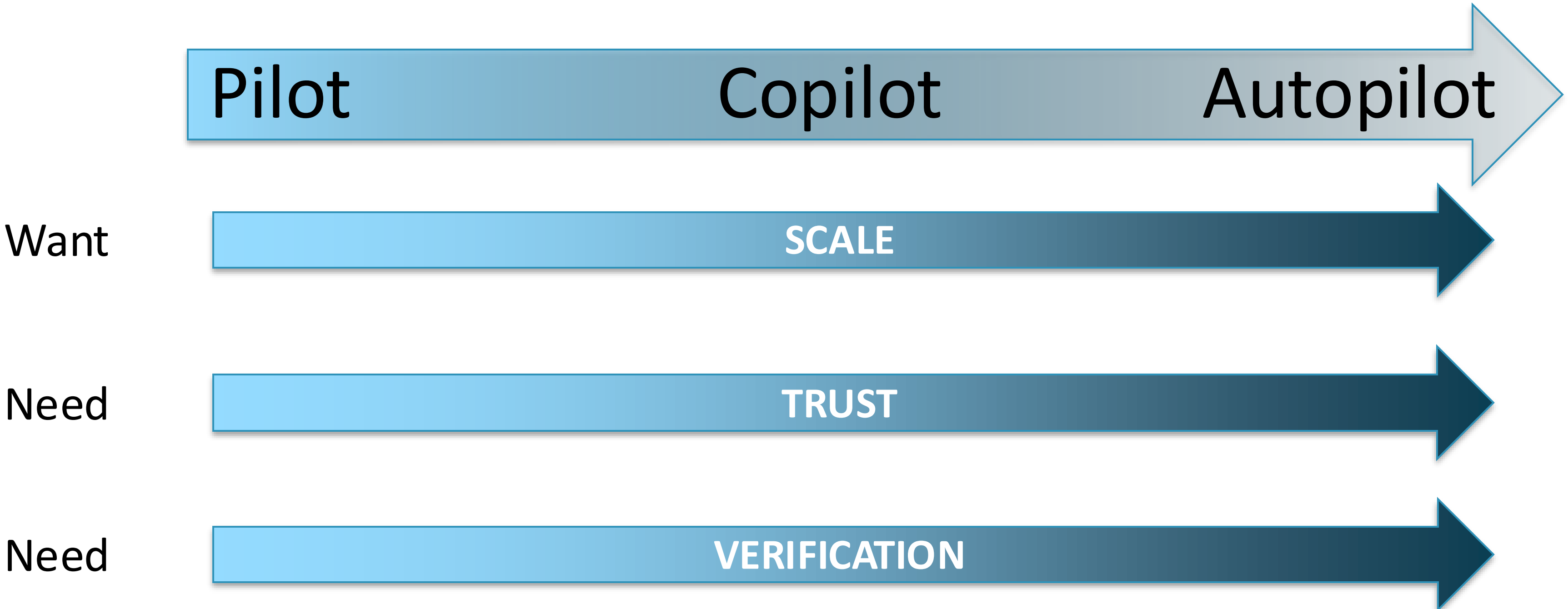
(a) Answer Path 1

(b) Answer Path 2

(c) Answer Path 3

Lessons

- Using LLMs as an evaluator must be done with a LOT of caution and verification.
- Judge the LLM (and benchmarks) before LLM can judge for us.
- Think about humans in the loop, accountability, and trustworthiness.
- “Trust, but verify” – Ronald Regan



UNIVERSITY of WASHINGTON

Thank you!

RAISE
RESPONSIBILITY IN AI SYSTEMS & EXPERIENCES

INFO SEEKING