

Course Overview

Anupam Datta (Snowflake)

John Mitchell (Stanford)

Instructors



Anupam Datta

Email: danupam@stanford.edu

Office hours: TBD



John Mitchell

Email: jcm@stanford.edu

Office hours: TBD

thanks to



Ankur Taly

(Google)

With



Jason Goldberg

Email:

Office hours: TBD



Ishita Gupta

Email:

Office hours: TBD



Josh Reini

(Snowflake)

History of CS329T

- Spring 2021, 2022
 - Online course developed during the pandemic, pre-ChatGPT
 - Trust dimensions in model training: explainability, fairness, privacy, robustness
- Fall 2023
 - Project-based course to draw on student ideas for using LLMs, post-ChatGPT
 - Emphasis on trust: relevance, groundedness, confidence, calibration, uncertainty, explainability, privacy, fairness, toxicity, adversarial attacks
- Fall 2024
 - Increasing sophistication about use of LLMs: from prompting to RAG, agents
 - Emphasis on RAG triad - grounding, context relevance, query relevance - and other trust-related measures

Course goal

Learn how to build and evaluate agentic AI systems

- Homework assignments to build and evaluate a sample agentic system
- **Quarter-long course project** to try out methods and tools in small teams

Area / Eval	Grounding	Planning	Consistency	Tool choice	Alignment
Science					
Education					
Healthcare					
Coding					
...					

Pick at least one row and two columns; use fine-tuning/RAG/agentic workflows to build apps

Outline for today

- **Course structure and requirements**
 - Topics and projects
 - Schedule and deadlines
- **How can we build AI applications?**
 - What applications are now possible?
 - What architectures and approaches are needed to build them?
- **Sample projects from past years**
 - A few examples to help you think ahead
 - Quick online poll on project directions that interest you

Course Structure

Pace of the course

- **Two-week introduction to building and evaluating agentic systems**
 - Two-week overview, with one hands-on homework
 - LLMs, Retrieval-Augmented Generation, agents; evaluation criteria
- **Week three: choose project directions and teams**
 - Present your project proposal in class; feedback from class
- **Additional topics**
 - Concepts and methods that will help you with your projects
 - Emerging directions and perspectives
 - Multi-modality
- **Project presentations and final project fair**
 - Mid-term presentations focus on functionality, evaluation approach
 - Final run-through covers evaluation results; project fair to show others

Requirements

- One two-part homework assignments and quarter-long course project
 - Assignments due in early week of the course
 - Project proposal (week 3), midterm (w 7), review (w 9), fair (week 10)
- Prerequisites:
 - Introductory Python-based ML class (CS229)
 - Knowledge of deep learning (e.g. CS230, CS231N, etc.)
 - Familiarity with ML frameworks in Python (scikit-learn, Keras)
- Grade: 70% Project, 15% Homework, 15% Class participation

Course Projects

- Work in small team: two or possibly three
- Choose an application area
 - Scientific research, healthcare, security, education, or other
- Characterize two or more needs for trustworthiness
 - Identify relevant properties, based on class presentations and concepts
- Develop approach building on methods presented in class
 - Find ways to measure, increase or ensure trustworthiness of AI apps

Additional details

- Class meetings: Tue/Thur 4:30 - 5:50pm, Lathrop Room 180
- Web page: <http://web.stanford.edu/class/cs329t/>
- Gradescope: assignment submissions
- Canvas: grades
- Ed: Announcements, all other communication and class discussion
- [Stanford Honor Code](#)

Building AI Applications

TruEra was founded in 2014 by Anupam Datta, Somesh Jha, and Tomer Geva. Some key details about the founders:

- Anupam Datta - Professor of Computer Science and Electrical and Computer Engineering at Carnegie Mellon University. He served as TruEra's President and Chief Scientist.
- Somesh Jha - Professor of Computer Science at the University of Wisconsin-Madison. He was the Vice President of Research at TruEra.
- Tomer Geva - Previously worked at Microsoft Research and VMware. He was TruEra's CEO.

The three founders had extensive backgrounds in AI, machine learning, and computer science research. Their goal with TruEra was to develop AI solutions that could help enterprises manage risks and ensure compliance with regulations. The company was headquartered in Redwood City, California.

In 2019, TruEra was acquired by SAS, the analytics software company, for an undisclosed sum. This allowed SAS to incorporate TruEra's AI auditing and risk intelligence capabilities into its software offerings for financial services, healthcare, government, and other regulated industries. While the TruEra brand was absorbed into SAS, the founders

A Who are the founders of TruEra?

TruEra founders

10 results

The founders of TruEra are:

Anupam Datta - Co-founder, President, and Chief Scientist. He was previously a professor at Carnegie Mellon University for 15 years, most recently as Professor and Director of the Accountable Systems Lab. [Snowflake to Acquire Truer...](#)

Shayak Sen - Co-founder and Chief Technology Officer. [Snowflake to Acquire Truer...](#)

Will Uppington - Co-founder and Chief Executive Officer. He previously held the role of Head of Product at Bloomreach. [Snowflake to Acquire Truer...](#)

TruEra was acquired by Snowflake in 2024, and all three co-founders joined Snowflake as part of the acquisition to continue developing AI observability capabilities for Snowflake's AI Data Cloud. [Snowflake](#) [TruEra](#) The company was founded around 2019 and specialized in AI quality management solutions that test, analyze, and monitor machine learning models.



   Retry

Claude can make mistakes. Please double-check cited sources.

Reply to Claude...



 Research

Claude Sonnet 4



From Prompting LLMs to Agentic AI

Insights, Design Patterns & Evaluation

TruEra was founded in 2014 by Anupam Datta, Somesh Jha, and Tomer Geva. Some key details about the founders:

- Anupam Datta - Professor of Computer Science and Electrical and Computer Engineering at Carnegie Mellon University. He served as TruEra's President and Chief Scientist.
- Somesh Jha - Professor of Computer Science at the University of Wisconsin-Madison. He was the Vice President of Research at TruEra.
- Tomer Geva - Previously worked at Microsoft Research and VMware. He was TruEra's CEO.

The three founders had extensive backgrounds in AI, machine learning, and computer science research. Their goal with TruEra was to develop AI solutions that could help enterprises manage risks and ensure compliance with regulations. The company was headquartered in Redwood City, California.

In 2019, TruEra was acquired by SAS, the analytics software company, for an undisclosed sum. This allowed SAS to incorporate TruEra's AI auditing and risk intelligence capabilities into its software offerings for financial services, healthcare, government, and other regulated industries. While the TruEra brand was absorbed into SAS, the founders



A Who are the founders of TruEra?

TruEra founders

10 results

The founders of TruEra are:

Anupam Datta - Co-founder, President, and Chief Scientist. He was previously a professor at Carnegie Mellon University for 15 years, most recently as Professor and Director of the Accountable Systems Lab. [Snowflake to Acquire Truer...](#)

Shayak Sen - Co-founder and Chief Technology Officer. [Snowflake to Acquire Truer...](#)

Will Uppington - Co-founder and Chief Executive Officer. He previously held the role of Head of Product at Bloomreach. [Snowflake to Acquire Truer...](#)

TruEra was acquired by Snowflake in 2024, and all three co-founders joined Snowflake as part of the acquisition to continue developing AI observability capabilities for Snowflake's AI Data Cloud. [Snowflake](#) [TruEra](#) The company was founded around 2019 and specialized in AI quality management solutions that test, analyze, and monitor machine learning models.



📄 🗨️ 🔄 Retry

Claude can make mistakes. Please double-check cited sources.

2023

2025

An Intermediate Step...

This overlap is really murky

AI research
optimizes LLMs for
Generalization

And actively
penalizes
Memorization

Focus LLMs on 'General' Tasks

Generalization

Memorization

- ✓ Summarization
- ✓ Text Embedding
- ✓ Logical Inference
- ✓ Planning

Leave memorization
to something else

LLMs Need a Knowledge Source

 Pinecone

 chroma

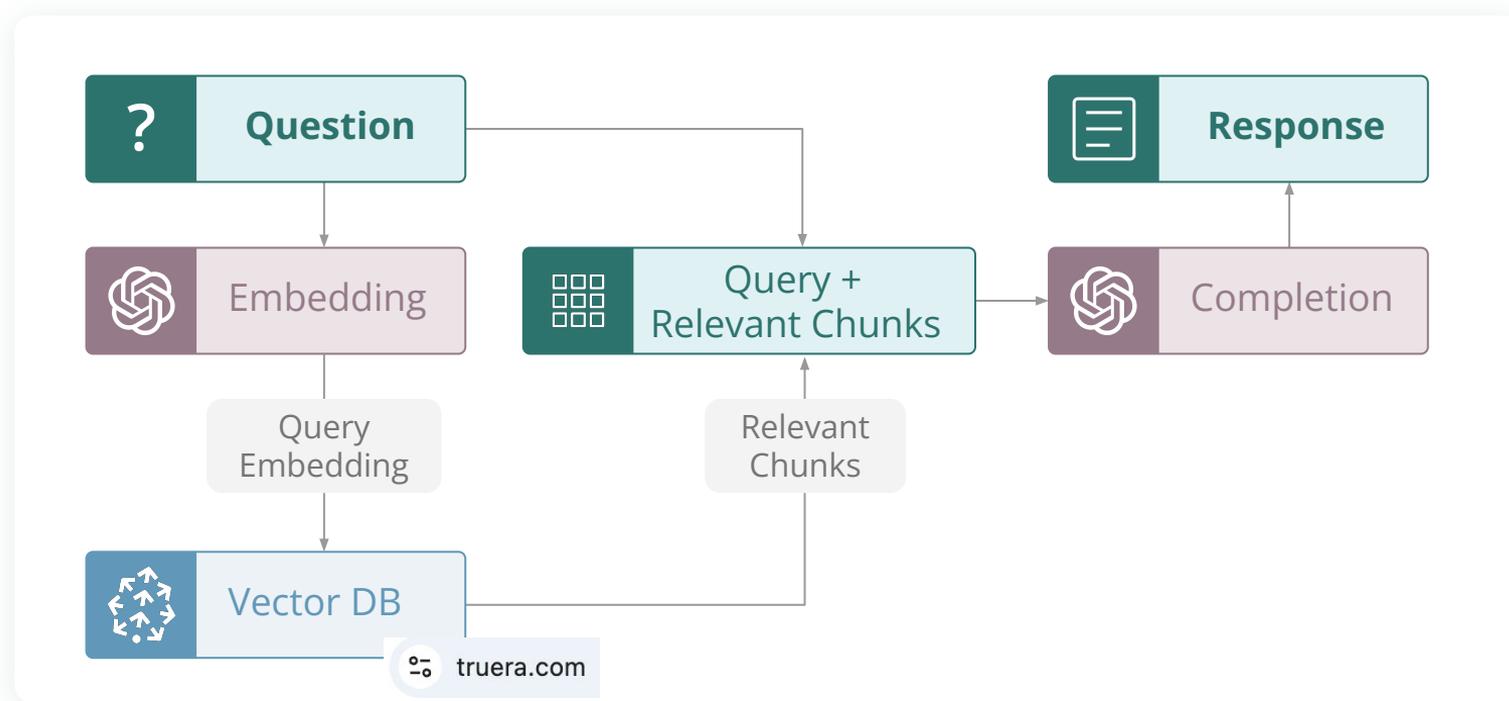
 Weaviate


Milvus

Vector Databases

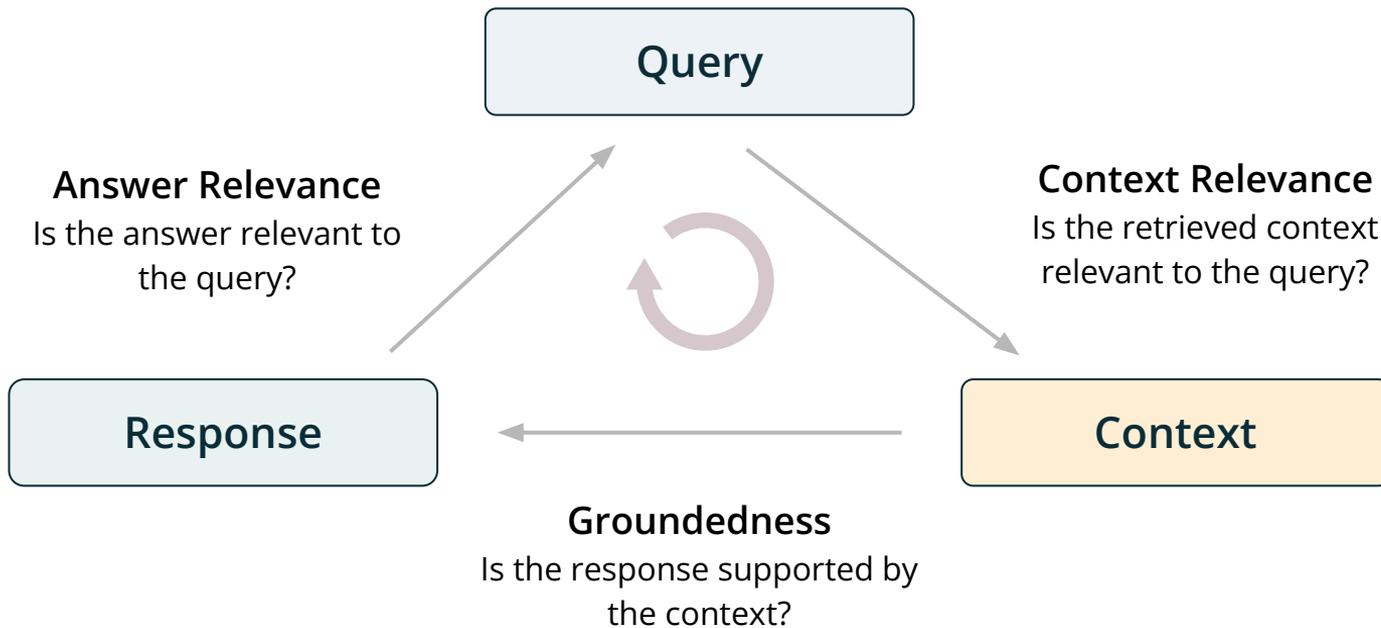
Enter Retrieval Augmented Generation (RAGs)

Example: Question Answering ChatBot



Evaluating RAGs

The RAG Triad



From RAGs to Agents

A Who are the founders of TruEra?

TruEra founders

10 results

The founders of TruEra are:

Anupam Datta - Co-founder, President, and Chief Scientist. He was previously a professor at Carnegie Mellon University for 15 years, most recently as Professor and Director of the Accountable Systems Lab. [Snowflake to Acquire Truer...](#)

Shayak Sen - Co-founder and Chief Technology Officer. [Snowflake to Acquire Truer...](#)

Will Uppington - Co-founder and Chief Executive Officer. He previously held the role of Head of Product at Bloomreach. [Snowflake to Acquire Truer...](#)

TruEra was acquired by Snowflake in 2024, and all three co-founders joined Snowflake as part of the acquisition to continue developing AI observability capabilities for Snowflake's AI Data Cloud. [Snowflake](#) [TruEra](#) The company was founded around 2019 and specialized in AI quality management solutions that test, analyze, and monitor machine learning models.



Copy Like Comment Retry

Claude can make mistakes. Please double-check cited sources.

Reply to Claude...

+ ⌵ 🔍 Research

Claude Sonnet 4



Groundedness

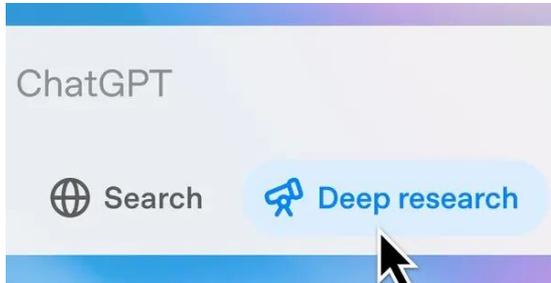
Answer Relevance

Context Relevance

Agent Building Blocks

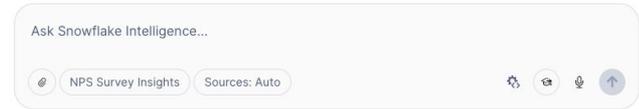
- Planning
- Tool Use (incl. web search)
- Reflection
- Memory
- Multi agent collaboration

Data Agents Gaining Widespread Adoption



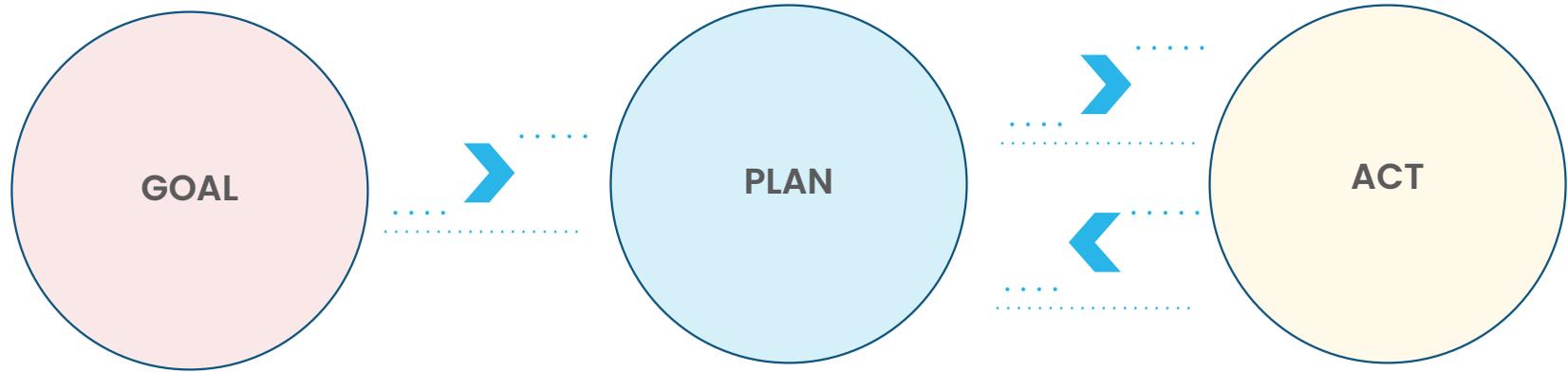
Deep Research

Good evening, Josh
What insights can I help with?



Snowflake Intelligence

How Do Agents Work?

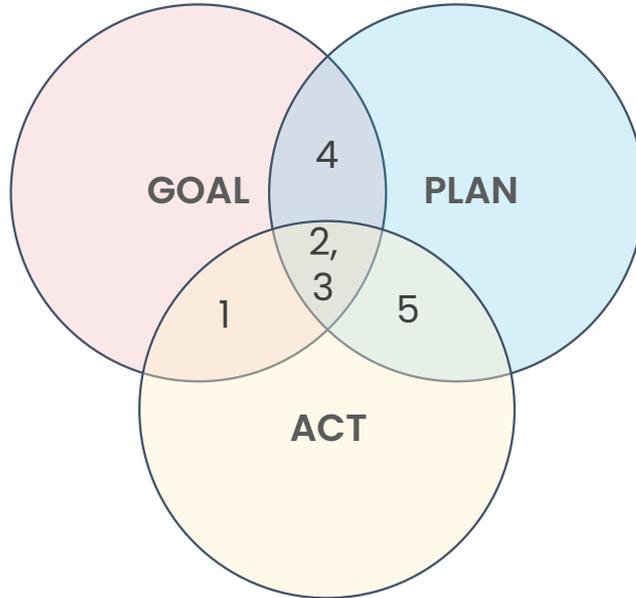


How Do Agents Work Well?



Trustworthy Agents Execute with **G**oals, **P**lans and **A**ctions Aligned.

What is Your Agent's GPA?



1. Goal Fulfillment

1A. Answer Relevance

2. Logical Consistency

3. Execution Efficiency

4. Plan Quality

4A. Tool Selection

5. Plan Adherence

5A. Tool Calling

Results

- Agent GPA framework implemented in the TruLens open source framework

- Validated on the TRAIL/GAIA dataset and an internal dataset for Snowflake Intelligence

Key Result 1

- **The Agent GPA framework provides a systematic way to cover a broad range of agent failures**
- In particular, it captures all the 281 agent internal errors on our test set split of the TRAIL/GAIA dataset

Key Result 2

- **LLM Judges for measuring Agent GPA show strong agreement with human judgments**
- TRAIL/GAIA dataset:
 - 94% (264/281) agreement
 - Baseline: TRAIL LLM Judge shows 55% (154/281) agreement
- Snowflake Intelligence:
 - 82% agreement when grading on a 3-point scale
 - 97% agreement when grading on a 2-point scale

Key Result 3

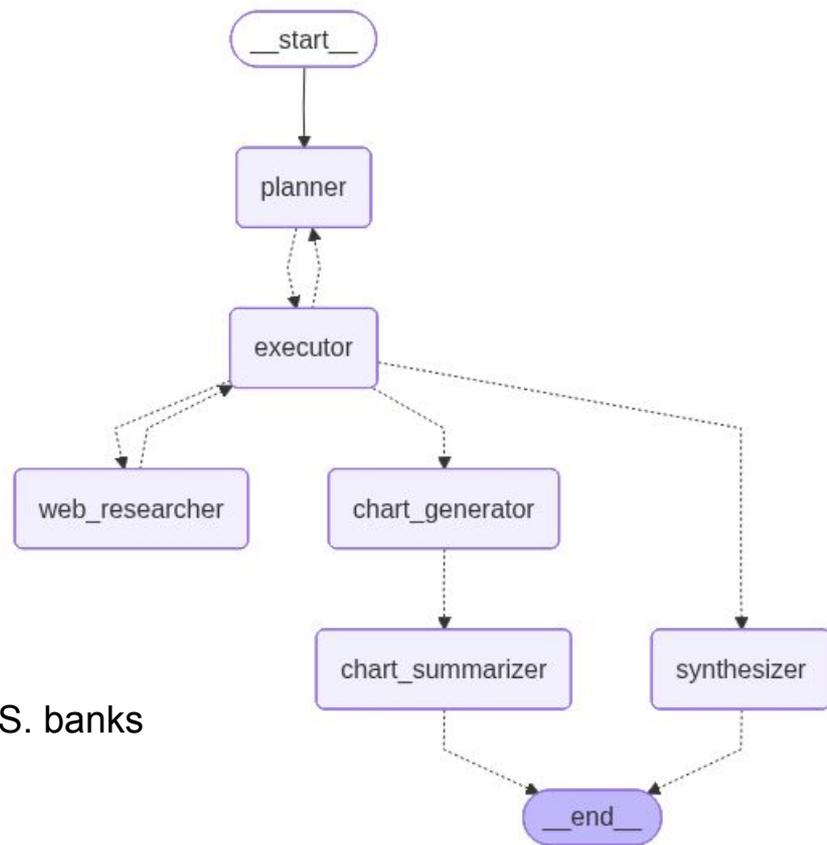
- **LLM Judges for measuring Agent GPA localize errors with strong agreement with human annotations**
- TRAIL/GAIA dataset:
 - 85% (240/281) agreement
 - Baseline: TRAIL LLM Judge shows 49% (138/281) agreement

Hands-on Activity: Build & Evaluate a Data Agent

- Build with [LangGraph](#)
- Evaluate with [TruLens](#)
- [Notebook](#)

Running Example: Web Research Agent

- Memory
 - Augmented LangGraph State
- Reflection
 - Performed by Executor/Planner
- Tool Use
 - Web Researcher uses Tavily Tool
 - Charting uses Python Tool
- Planning
 - Performed by Planner
- Multi agent collaboration
 - Coordinated through Planner and Executor



Query: Chart the market cap (in USD) of the top five U.S. banks

Refresh Data

Info

Web Search Data Agent / Base / Record: fb4d0bd9-0b27-4902-a30d-22e477cebdde

Record ID: fb4d0bd9-0b27-4902-a30d-22e477cebdde

Record Input

Chart the current market capitalization of the top 5 banks in the US?

Record Output

Here is the chart visualizing the market capitalizations of the top 5 U.S. banks. You can find the chart at the provided path:

[top\5_us_banks_market_cap.png](#). Each bank's position on the chart highlights JPMorgan Chase & Co.'s substantial lead in market capitalization, with more than double the market cap of its closest competitor, Bank of America Corporation.

Total tokens (#) ?

223

Total cost (USD) ?

\$0.00

↓ -0.0122 USD

Latency (s) ?

68.757...

↓ -87.6s

Feedback Results

Feedback Functions (click to learn more)

Answer Relevance 0.00

Context Relevance 1.00

Execution Efficiency 1.00

Groundedness 0.33

Logical Consistency 1.00

Plan Adherence 0.33

Plan Quality 1.00

Feedback Results

Feedback Functions (click to learn more)

Answer Relevance 0.00 Context Relevance 1.00 Execution Efficiency 1.00 Groundedness 0.33

Logical Consistency 1.00 Plan Adherence 0.33 **Plan Quality 1.00**

trace	score	explanation
Trace(events= record event_id \ 0 {		

Trace Details

Tree Timeline

graph
⌚ 69 s Record root

planner_node
⌚ 21 s Graph node

executor_node
⌚ 6 s Graph node

Criteria: Score the quality of the plan. 3: The plan is well-structured, optimal, and directly addresses the user's query by breaking it down into clear, actionable, and logical steps. Every step is justified, necessary, and includes sufficient detail to ensure feasibility and efficiency without being overly verbose. If replanning occurs, the revised plan and rationale are provided with precise references to observed gaps or triggers in prior results or context. Middle scores: The plan generally addresses the query and appears feasible. Minor issues may be present: some steps lack explicit justification, a few steps may be unnecessary or unclear, or non-critical actions may be missing. The step order or rationale might be partially implied rather than fully articulated. If replanning occurs, it is mentioned but may lack thorough explanation or explicit ties to prior context. 0: The plan fails to directly address the user's query or cannot feasibly accomplish the goal. Critical steps are missing, irrelevant, unsupported, or based on fabricated reasoning. Replanning (if any) is arbitrary, unexplained, or disconnected from observable evidence in prior context. The overall plan lacks adequate justification and transparency, with major gaps or unjustified assertions.

Supporting Evidence: The plan consists of three steps: (1) Use a web researcher to search for and extract the current market capitalizations of the top 5 US banks, producing a structured table; (2) Use a chart generator to create a horizontal bar chart visualizing the market capitalizations using the data from step 1; (3) Use a chart summarizer to provide a concise narrative explaining the differences in market capitalization among the top 5 US banks. Each step is necessary, clearly justified, and directly addresses the user's query to chart the current market capitalization of the top 5 US banks. The plan is logically ordered, with explicit pre- and post-conditions for each step, ensuring data flows correctly between agents. There is no unnecessary or missing step, and no evidence of arbitrary or unexplained replanning. The plan is neither verbose nor lacking in detail, and all actions are feasible and efficient for the stated goal.

Feedback Results

Feedback Functions (click to learn more)

Answer Relevance 0.00

Context Relevance 1.00

Execution Efficiency 1.00

Groundedness 0.33

Logical Consistency 1.00

Plan Adherence 0.33

Plan Quality 1.00



trace	score	explanation
Trace(events= record event_id \ 0 {		

Trace Details

Tree Timeline

graph
69 s Record root

planner_node
21 s Graph node

executor_node
6 s Graph node

Criteria: Score the adherence of the execution to the plan. 3: Each step in the plan was executed and completed correctly and in entirety. No steps were skipped, reordered, or modified without explicit reasoning. Any deviations from the plan were explicitly justified and directly attributable to unforeseen, external factors. If replanning was necessary, the revised plan was followed exactly. Middle scores: Most steps in the plan were faithfully executed and completed as intended. Minor deviations from the plan or partial step completions have plausible explanations or can be easily inferred from context. If replanning was necessary, the revised plan was generally followed. 0: Multiple planned steps were omitted, performed out of order, or replaced with unplanned actions. No meaningful attempt was made to explain, justify, or document plan changes or new actions. The plan was largely ignored or disregarded in execution, or steps were not completed as intended. If replanning was necessary, the revised plan was not followed.

Supporting Evidence: The plan consists of three steps: (1) web research to obtain a structured list of the top 5 US banks by market capitalization, (2) generate a horizontal bar chart using that data, and (3) summarize the key takeaways from the chart. Step 1 was executed by the web_researcher, who provided the required structured data. Step 2 was executed by the chart_generator, who created and saved the horizontal bar chart as specified. However, step 3, which requires the chart_summarizer to provide a concise narrative summarizing the differences in market capitalization, was not completed. The trace shows the chart_summarizer node was called, but no summary was produced or included in the final output. There is no explicit justification for omitting this step, and the final output only includes the chart and a brief note, not a full summary as required by the plan.

Answer Relevance 0.00

Context Relevance 1.00

Execution Efficiency 1.00

Groundedness 0.50

Logical Consistency 1.00

Plan Adherence 0.33

Plan Quality 1.00



trace score explanation

Trace(events= record event_id \ 0 {'name':

Trace(events= record event_id \ 0 {'name':

Criteria: Score the logical consistency of the trace, including both the plan and execution. (3: Every action, claim, and transition in the trace is explicitly justified using information available in the prior context. Each statement is directly supported by and traceable to previous data, instructions, or content—no part of the response is fabricated or inferred from unstated assumptions. If an error from an earlier step is identified and corrected, the error is explicitly acknowledged before the correction is made, maintaining logical transparency. Each system instruction is followed. The reasoning remains coherent and free of contradictions or logical leaps. Middle scores: Occasional lapses in logic, minor unsupported assertions, or isolated explanatory gaps. Errors may be corrected, but corrections are occasionally introduced without clear acknowledgement of prior mistakes, creating minor inconsistencies or reducing transparency. Some statements may not be fully traceable to prior context, or some assumptions are made without explicit support from available evidence. Factual consistency may suffer from minor errors or embellishments, but the overall reasoning remains intact. Most previously assigned tasks and instructions remain intact. 0: Frequent or severe breakdown in the logical flow; many statements are either unsupported by, or cannot be grounded in, the prior context. Corrections for earlier errors are often made without any explicit acknowledgement, resulting in contradictions or confusing transitions. Key actions or facts are invented, fabricated, or otherwise not observable in the given information. Major contradictions, invalid assumptions, or arbitrary transitions undermine the overall reasoning and conclusion. Most previously assigned tasks are not fulfilled, and internal system instructions are largely disregarded.)

Supporting Evidence: The trace begins with a clear plan: (1) research the top 5 US banks by market cap, (2) generate a bar chart, (3) summarize the chart. Each agent's system instructions are followed: the web researcher only uses the search

Trace Details

Tree Timeline

graph

59 s

Record root

planner_node

18 s

Graph node

executor_node

Answer Relevance 0.00

Context Relevance 1.00

Execution Efficiency 1.00

Groundedness 0.50

Logical Consistency 1.00

Plan Adherence 0.33

Plan Quality 1.00



trace	score	explanation
Trace(events= record event_id \ 0 {'name':		
Trace(events= record event_id \ 0 {'name':		

Trace Details

Tree

Timeline

graph

59 s

Record root

planner_node

18 s

Graph node

executor_node

Criteria: Score the efficiency of the execution. 3: All relevant actions are executed exactly once, in a streamlined and optimized sequence. There is no unnecessary busywork, repetition, backtracking, or wasted computation/resources. Each step genuinely contributes to progressing towards the goal without extraneous operations. Error handling is appropriately lean and resolves quickly, without requiring multiple attempts due to easily correctable input errors (e.g., incorrect tool arguments). Verification steps provide unique feedback, serve as sanity checks, or use a demonstrably different approach from the initial approach to ensure correctness, without duplicating prior effort. Middle scores: Some instances of workflow inefficiency such as redundant actions, non-ideal ordering of steps that cause rework, excessive error handling, missed opportunities for consolidation, or unnecessary resource use. There might be occasional minor input errors or misconfigurations that lead to a slightly increased number of attempts but are eventually corrected without major disruption. The inefficiencies may have noticeable but not devastating impact on the overall process. 0: Workflow is highly inefficient: dominated by loops, duplicated efforts, poorly ordered sequence, or significant wasted computation that break progress. Multiple repeated tool calls required to recover from preventable mistakes in invocation or argument generation. Verification steps are highly redundant and do not provide any value. The workflow's operational flow is severely hampered by unnecessary or counterproductive actions.

Supporting Evidence: The execution proceeds in a linear, stepwise fashion: (1) planning, (2) web research for current market cap data, (3) chart generation, and (4) chart summarization. Each agent is invoked once per its role, and there is no evidence of redundant tool calls, unnecessary retries, or backtracking. The web search is performed once, the chart is generated with a single Python code execution. and the summary is produced directly from the chart output. There are

From Prompting LLMs to Agentic AI

Insights, Design Patterns & Evaluation

TruEra was founded in 2014 by Anupam Datta, Somesh Jha, and Tomer Geva. Some key details about the founders:

- Anupam Datta - Professor of Computer Science and Electrical and Computer Engineering at Carnegie Mellon University. He served as TruEra's President and Chief Scientist.
- Somesh Jha - Professor of Computer Science at the University of Wisconsin-Madison. He was the Vice President of Research at TruEra.
- Tomer Geva - Previously worked at Microsoft Research and VMware. He was TruEra's CEO.

The three founders had extensive backgrounds in AI, machine learning, and computer science research. Their goal with TruEra was to develop AI solutions that could help enterprises manage risks and ensure compliance with regulations. The company was headquartered in Redwood City, California.

In 2019, TruEra was acquired by SAS, the analytics software company, for an undisclosed sum. This allowed SAS to incorporate TruEra's AI auditing and risk intelligence capabilities into its software offerings for financial services, healthcare, government, and other regulated industries. While the TruEra brand was absorbed into SAS, the founders



2023

2025

A Who are the founders of TruEra?

TruEra founders

10 results

The founders of TruEra are:

Anupam Datta - Co-founder, President, and Chief Scientist. He was previously a professor at Carnegie Mellon University for 15 years, most recently as Professor and Director of the Accountable Systems Lab. [Snowflake to Acquire Truer...](#)

Shayak Sen - Co-founder and Chief Technology Officer. [Snowflake to Acquire Truer...](#)

Will Uppington - Co-founder and Chief Executive Officer. He previously held the role of Head of Product at Bloomreach. [Snowflake to Acquire Truer...](#)

TruEra was acquired by Snowflake in 2024, and all three co-founders joined Snowflake as part of the acquisition to continue developing AI observability capabilities for Snowflake's AI Data Cloud. [Snowflake](#) [TruEra](#) The company was founded around 2019 and specialized in AI quality management solutions that test, analyze, and monitor machine learning models.



📄 🗨️ 🔄 Retry

Claude can make mistakes. Please double-check cited sources.

Snowflake is Hiring Interns & Full-time Engineers



Sample Projects

Chain of Cases (Jason Goldberg, Pooja Sethi, 2023)

- Goal of the application
 - Answer queries about US Supreme Court decisions
 - Provide citations to specific cases so lawyers have quotable precedents
- Trustworthiness requirements
 - **Grounding** - accurately reflect Supreme Court decisions
 - **Interpretability** - identify specific cases that provide basis for the answer
- Background
 - In June 2023, two New York lawyers were sanctioned when they submitted a legal brief that included six fake cases (Merken, 2023)
 - The lawyers had used ChatGPT to help prepare their brief
 - The cases were “hallucinated” by the model and not checked by attorneys

Basic approach

- Construct index once based on shortened overviews
 - Use LLM to condense Supreme Court rulings to shorter overview
 - Build RAG vector database using condensed overviews
- Produce each response to user query in three steps
 - Query index for top k nodes
 - Summarize cases corresponding to retrieved nodes, with query context.
 - Order summaries and use to iteratively refine LLM response.

Report: https://drive.google.com/file/d/11fFvIBmEPoP3r3Fide9M_d5WmxbRcmlp/view?usp=drive_link

Slides: https://docs.google.com/presentation/d/1D9GiGpCPu6Qjoj9_XWubnycG1ORQ9f7kn3DQR9mOS50/edit?usp=sharing

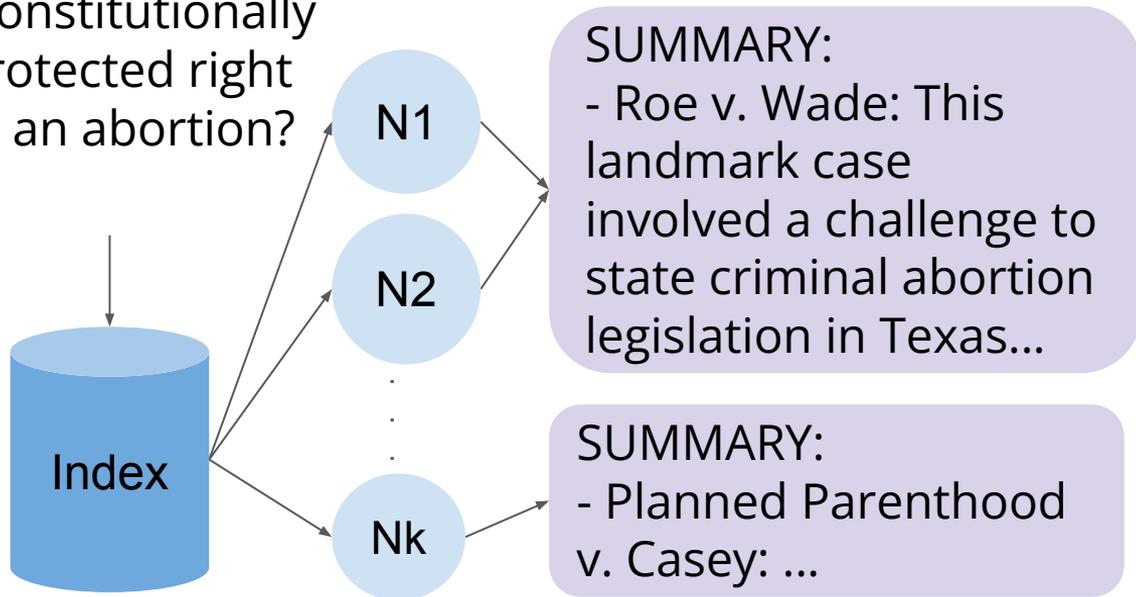
Querying Strategy: Chain-of-Cases Prompting

1. **Query** index for top k nodes.

2. **Summarize** cases corresponding to retrieved nodes, *with* query context.

3. **Response synthesis**: order summaries and use to iteratively refine LLM response.

Do women have a constitutionally protected right to an abortion?



Roe v. Wade, 1972-048 ...

Maher v. Roe, 1976-155 ...

Planned Parenthood v. Casey, 1992-027 ...

ANSWER: It is established law that women...

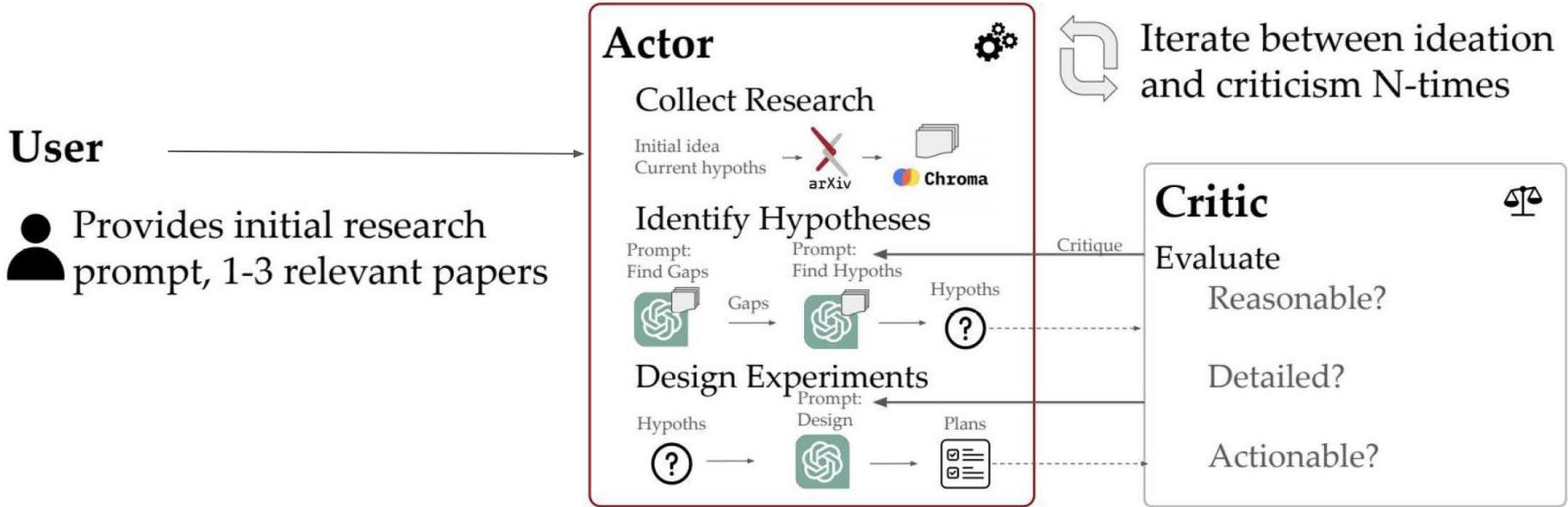
Results: Chain-of-Cases gives relevant, grounded responses

	Toy Corpus + Vanilla Prompt	SCOTUS + Vanilla Prompt	SCOTUS + Chain-of-Cases Prompting
	...whether women have a constitutionally protected right to an abortion is a matter that is not definitively addressed by the Constitution.	...the constitutionality of specific abortion regulations depends on their purpose and effect in relation to a legitimate state interest.	...states are permitted to regulate abortion as long as the regulations do not create an undue burden... as clarified in Planned Parenthood v. Casey (1992-027).
AR	0.75	0.83	0.97
G	0.76	0.79	1
CR	0.57	0.73	0.66

Agentic Research Planner (Shounak Ray, Harrison Delecki 2024)

- Motivation and Goal
 - Planning new research projects can require significant time and expertise
 - Aim to enable more efficient exploration of research directions
- Methods
 - User provides initial ideas to agentic workflow composed of actor and critic
 - Actor and critic iteratively collect research, propose hypothesis, refine plan
 - Evaluation: RAG trio, alignment to creativity, specificity, completeness
- Results
 - Agentic workflow shows improvement across LLM evaluation metrics

Methodology



Output Creative, specific, research plans involving:
Hypothesis, methods, datasets, experiments, etc

Evaluation

- Generate research plans in familiar areas
- RAG evaluation
 - Relevance of retrieved paper context
 - Answer relevance to original query
 - Response groundedness in retrieved paper context
- Qualitative evaluation
 - LLM judges for creativity, specificity, completeness

Results

- Baselines for comparison
 - Prompt ChatGPT to produce research plans directly
 - One iteration of Actor without Critic feedback
- Results after 10 iterations

Method	Answer Relevance	Context Relevance	Groundedness
Actor Only	0.70 ± 0.40	0.92 ± 0.07	0.43 ± 0.31
Actor-Critic (10)	0.73 ± 0.41	0.98 ± 0.04	0.40 ± 0.36

Method	Creativity	Specificity	Completeness
ChatGPT	0.80 ± 0.37	0.34 ± 0.26	0.57 ± 0.54
Actor Only	0.6	0.65 ± 0.09	0.68 ± 0.10
Actor-Critic (10)	0.6	0.76 ± 0.14	0.76 ± 0.14

Low groundedness may be a proxy for creativity.

Relevance scores alongwith LLM evaluations marginally improved in agentic workflow, indicating the potential benefit of an iterative approach.



Agentic Research Project Planner

Shounak Ray and Harrison Delecki

Overview

Motivation and Goal

- Planning new research projects can require significant time and expertise.
- We aim to enable more efficient exploration of research directions.

Methods

- User provides initial ideas to agentic workflow composed of actor and critic.
- Actor and critic iteratively collect research, propose hypothesis, and refine plan.
- Trustworthiness: RAG trio, alignment to creativity, specificity, completeness

Results

- Built end-to-end, adaptable “actor-critic” update abstraction
- Agentic workflow shows potential improvement across LLM evaluations and trustworthiness metrics

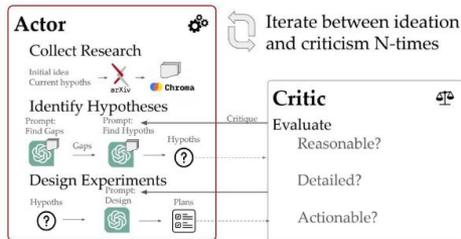
Background

- Preliminary studies suggest that LLM tools may be beneficial for rapid ideation [1, 2].
- Existing research co-pilots focus on search, summarization, or writing [3].
- Research project plans typically involve:
 - Literature on current approaches
 - Gaps or limitations of current work
 - Hypothesis about how to address the gap
 - Experiments to test the hypothesis
 - Evaluation on experimental outcomes

Methodology

Key Idea: Inspired by real-life research ideation, we propose an iterative, agentic approach.

User Provides initial research prompt, 1-3 relevant papers



Output Creative, specific, research plans involving: Hypothesis, methods, datasets, experiments, etc

Experiments and Metrics

- We generate research plans in familiar areas
- RAG evaluation
 - Relevance of retrieved paper context
 - Answer relevance to original query
 - Response groundedness in retrieved paper context
- Qualitative evaluation (Should we call this alignment)
 - LLM judges to creativity, specificity, completeness

Results and Discussion

Baselines

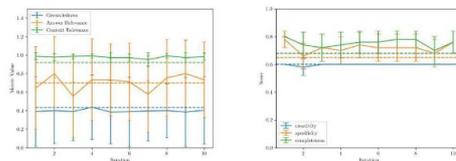
- Prompt ChatGPT to produce research plans directly
- One iteration of Actor without Critic feedback

Results after 10 iterations:

Method	Answer Relevance	Context Relevance	Groundedness
Actor Only	0.70 ± 0.40	0.92 ± 0.07	0.43 ± 0.31
Actor-Critic (10)	0.73 ± 0.41	0.98 ± 0.04	0.40 ± 0.36

Method	Creativity	Specificity	Completeness
ChatGPT	0.80 ± 0.37	0.34 ± 0.26	0.57 ± 0.54
Actor Only	0.6	0.65 ± 0.09	0.68 ± 0.10
Actor-Critic (10)	0.6	0.76 ± 0.14	0.76 ± 0.14

Low groundedness may be a proxy for creativity. Relevance scores along with LLM evaluations marginally improved in agentic workflow, indicating the potential benefit of an iterative approach.



All LLM evaluation and trustworthiness metrics are roughly constant across agentic iterations, suggesting ineffectiveness of critic-update mechanism.

References

- [1] Gonzalez, G. E., Moran, D. A. S., Houde, S., He, J., Ross, S. L., Muller, M. J., ... & Weisz, J. D. (2024, March). Collaborative Canvas: A Tool for Exploring LLM Use in Group Ideation Tasks. In *IJL Workshops*.
- [2] He, J., Houde, S., Gonzalez, G. E., Silva Moran, D. A., Ross, S. L., Muller, M., & Weisz, J. D. (2024, June). AI and the Future of Collaborative Work: Group Ideation with an LLM in a Virtual Canvas. In Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (pp. 1-14).
- [3] Wang, B., Zhang, X., Li, S., & Wang, Y. (2024, April). The Practice of Enhancing Learning and Scientific Innovative Abilities Using LLM-Based AI Tools. In 2024 6th International Conference on Computer Science and Technologies in Education (CSTE) (pp. 166-170). IEEE.

Your initial thoughts on projects

- Open the Google doc using this QR code 
- Add one or two project ideas
- Include your name, or add your name to a project someone else already added



Welcome to CS329T
for 2025 !!