CS224v

Agentic AI

Lecture 1: Introduction

Monica Lam

Core User-Facing Capabilities of LLMs

Conversational Q&A

Direct chat with multi-turn context retention across a session

RAG & Retrieval

Grounding answers in external docs and live web search, with citations

Function Calling

The model decides when to call an API, database, or external tool

Agentic Workflows

Multi-step planning, tool use, and self-correction with less supervision

Computer Use

Operating a GUI directly — clicking, typing, navigating on a user's behalf

Code Execution

Running and iterating on code in a live sandbox, not just generating text

Extended Reasoning

Step-by-step chain-of-thought for math, logic, and complex problem solving

Multimodal I/O

Understanding and generating images, audio, and video alongside text

Today's AI Agents Are Not Deployed in High-Stakes Services
(Medicine, Science, Finance, Law)

Fundamental Techniques to Improve:

Accuracy, Accountability, Long Horizon, Long Context

Impact

Affordable High-Stakes Services

Knowledge Discovery Acceleration

LLM Hallucinations

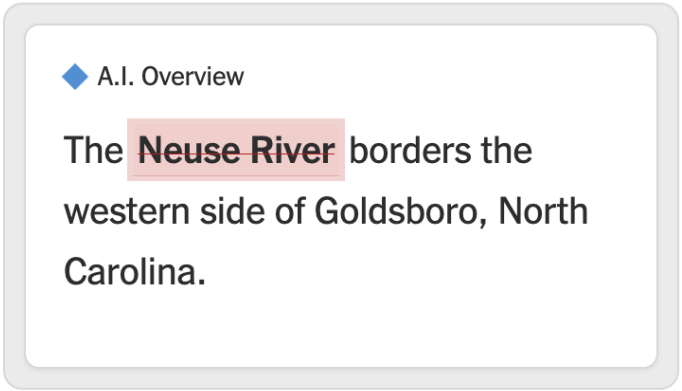
Date	Incident	Why it matters	Citation
May 2023	Mata v. Avianca	ChatGPT-assisted legal filing included **fake cases and citations**; court imposed sanctions.	NYT, 5/27/2023
Dec 2023	Michael Cohen filing	AI-generated **fabricated legal citations** appeared in a high-profile legal matter.	Reuters, 12/29/2023
Dec 2023	Chief Justice Roberts report	Chief Justice Roberts urged **caution when using AI** in legal contexts.	U.S. Supreme Court,12/31/2023
Jan 2024	Stanford legal hallucination study	Found **69%–88% hallucination rates** on legal tasks across **GPT-3.5, PaLM 2, and Llama 2**.	Dan Ho, Stanford arXiv, 1/11/2024
Feb 2024	Moffatt v. Air Canada.	Company held responsible after chatbot gave **incorrect fare/refund guidance**.	The Guardian 2/16/2024
Feb 2024	Colorado sanctions case	Lawyers sanctioned for AI-generated filings with **fake court citations**.	Reuters, 2/22/2024

Hallucination is Still Rampant with RAG Today

SCALED CONSUMER PRODUCTS

NYT, April 7, 2026

The New York Times



Estimated Hallucination Rate: 10%

How Accurate Are Google's A.I. Overviews?

JOURNALISM

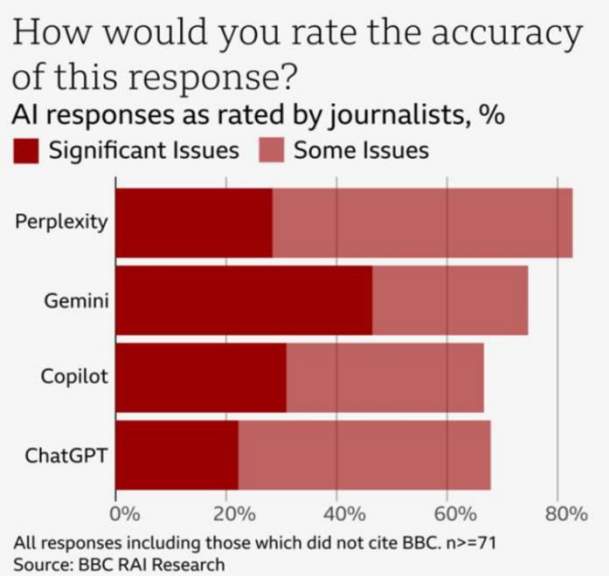
Perplexity, Gemini, Copilot, ChatGPT Hallucinate With RAG

BBC, Feb 11, 2025

51% of AI answers to news problems have significant issues.

19% of AI answers which cited BBC content introduced factual errors.

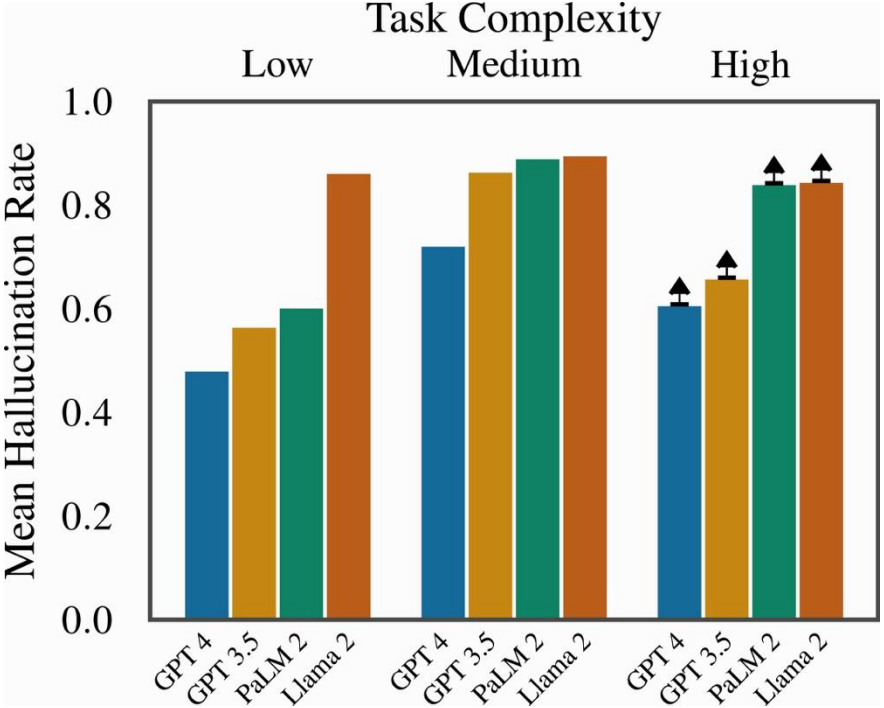
13% of the quotes sourced from BBC articles were either altered or didn't actually exist in that article.



https://www.bbc.com/mediacentre/2025/bbc-research-shows-issues-with-answers-from-artificial-intelligence-assistants

LEGAL

Stanford, June 26, 2024



Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models

Matthew Dahl ✉, Varun Magesh, Mirac Suzgun, Daniel E Ho Author Notes

Journal of Legal Analysis, Volume 16, Issue 1, 2024, Pages 64–93, https://doi.org/10.1093/jla/laae003

Published: 26 June 2024

SCIENCE

NEWS FEATURE | 01 April 2026

Hallucinated citations are polluting the scientific literature. What can be done?

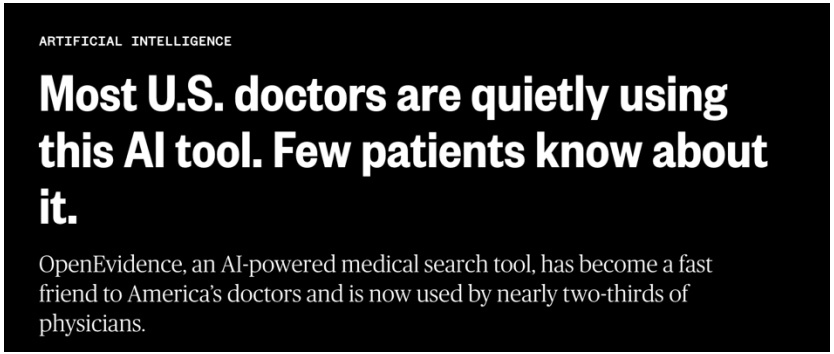
Tens of thousands of publications from 2025 might include invalid references generated by AI, a *Nature* analysis suggests.

By Miryam Naddaf & Elizabeth Quill

https://www.nature.com/articles/d41586-026-00969-z

MEDICINE

NBC News
May 13, 2026


ARTIFICIAL INTELLIGENCE

Most U.S. doctors are quietly using this AI tool. Few patients know about it.

OpenEvidence, an AI-powered medical search tool, has become a fast friend to America's doctors and is now used by nearly two-thirds of physicians.

https://www.nbcnews.com/tech/tech-news/openevidence-ai-doctor-medical-physician-login-app-what-npi-uptodate-rcna341064

Sept 2026

MedXPertQA:
Expert-Level Specialty Board Questions

Accuracy: 72.8%

FINANCE: AI Hallucination not Publicized

An enterprise:

"The bot says 'the investment is made' but it was not."

A large consulting firm:

"Evaluation shows many hallucinations in an enterprise chatbot"

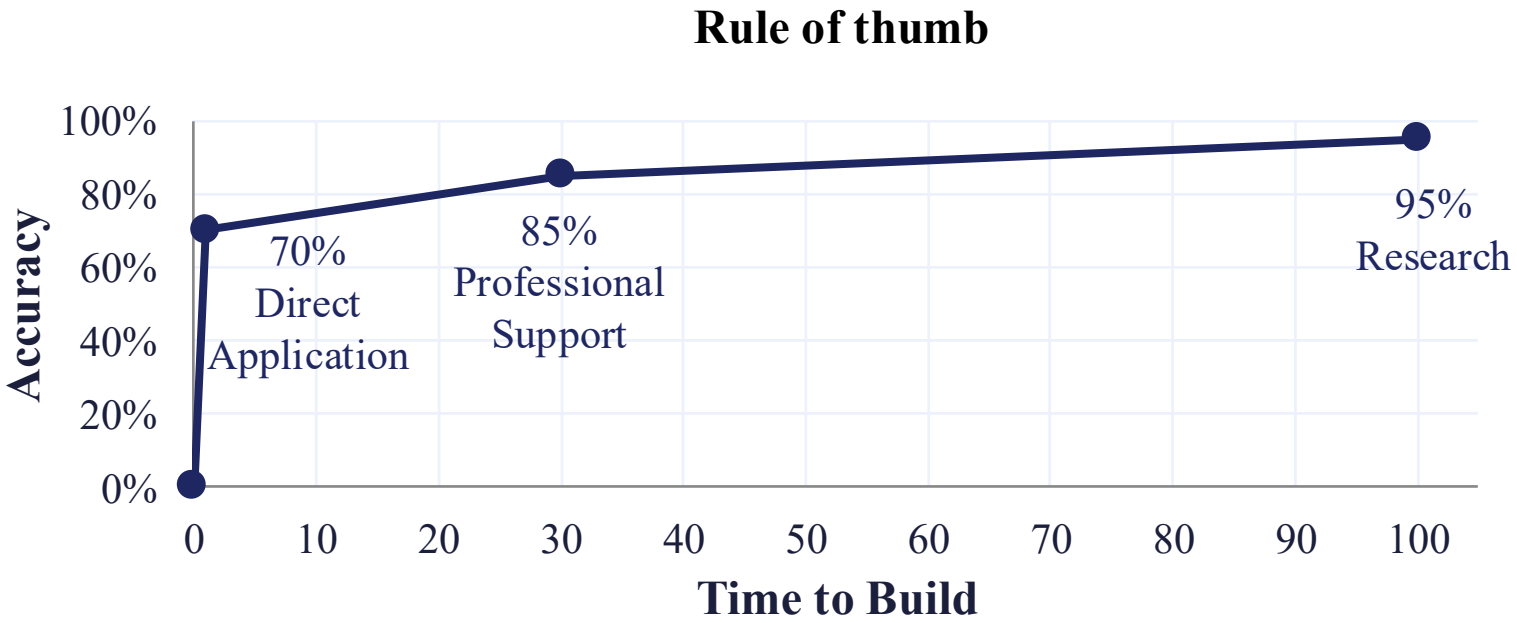
A popular chatbot vendor:

"Does not sell bots for high-stakes applications."

A large enterprise infrastructure vendor:

"No enterprises deployed AI in high-stakes applications."

High-Stakes LLM Agents are Hard to Build!



The Last 10% is 90% of the problem.

LLM Weaknesses

from Next-Word Prediction Training

Out-of-Distribution Tasks

- **New** information
- **Long**-tail information

} RAG (Retrieval Augmented Generation)

- **Long** instructions
- **Long** documents
- **Long** horizon
- **Long** contexts

} Context management
e.g. automatic compression

Focus of CS224V

Today's AI Agents Are Not Deployed in High-Stakes Apps

(Medicine, Science, Finance, Law)

Fundamental Techniques to Improve:

Accuracy, Accountability, Long Horizon, Long Context

Impact

Affordable High-Stakes Services

Knowledge Discovery Acceleration

Outline of This Lecture

1. Problem: LLM Hallucination in High-Stakes Applications
2. Basic concept: Decompose problems into reliable LLM tasks
3. Course information
4. Course syllabus preview
   Stanford A-OS Agentic Operating System
   - Long-Horizon, Long-Context Custom-Facing Agents
   - Long-Horizon, Long-Context Researcher for Science
   - Accountable Decision Making

Why is There Hallucination with RAG

1. Retrieval Does not Return the Answer
2. LLMs Supply the Answer from Memory

WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia

Sina J. Semnani **Violet Z. Yao*** **Heidi C. Zhang*** **Monica S. Lam**
Computer Science Department
Stanford University
Stanford, CA
{sinaj, vyao, chenyuz, lam}@cs.stanford.edu

- Conversational Q&A in open domain with 97% accuracy
- Winner of the Wikimedia's Research of the Year Award, 2024
- In Proceedings of Association for Computational Linguistics: EMNLP, Singapore, December 2023.

WikiChat (7 Prompts)

Traditional (Factuality)

1. Formulate query from input
   - Retrieve documents
2. Filter *each* retrieved doc

LLM (Conversationality + Factuality)

3. Ask GPT to generate answer
4. Extract claims
5. Fact-check/remove *each* claim
   - Retrieve documents

6. Draft
7. Refine

LLM calls: 5 + *n* + *c*
n: # documents retrieved based on user queries
c: # claims generated

The problem is Harder than Expected

3 students, 4 months

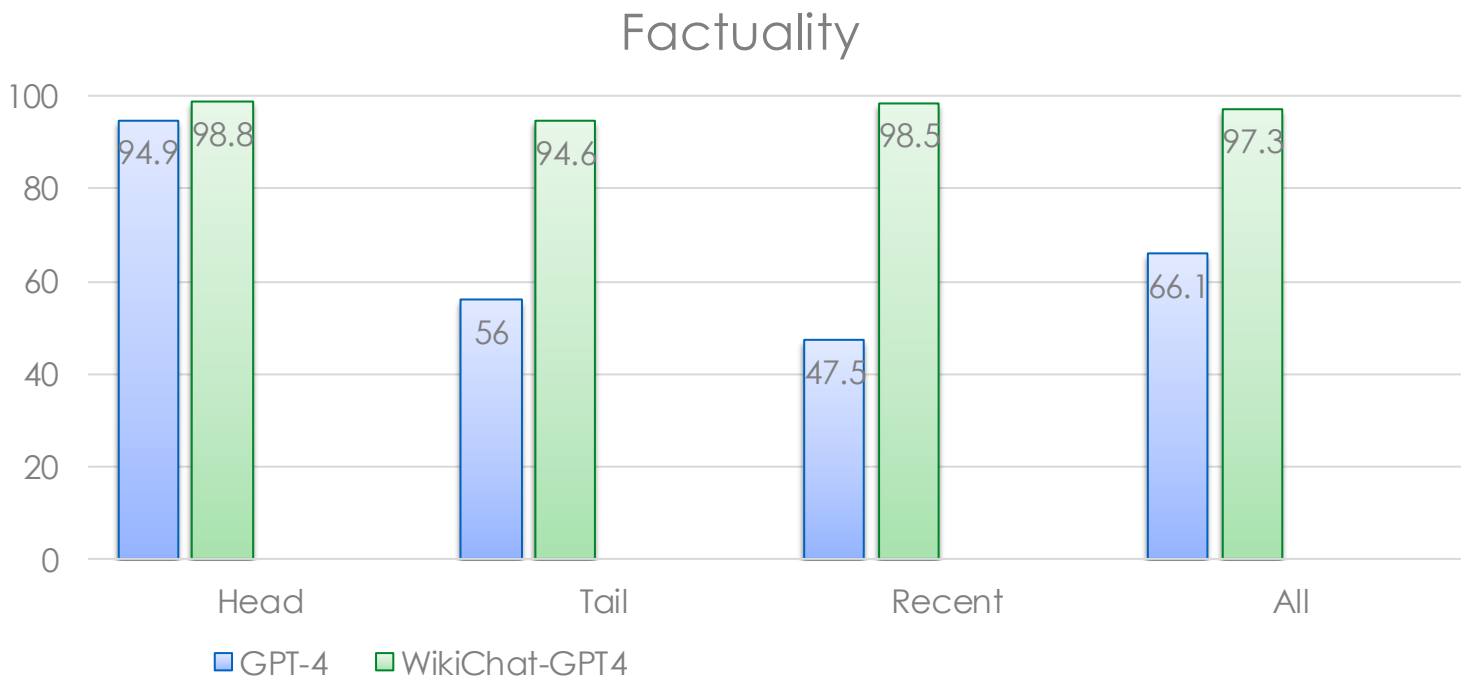
The Devil is In the Details

Decompose Problems into Reliable LLM Task

Speed can be Regained via Distillation

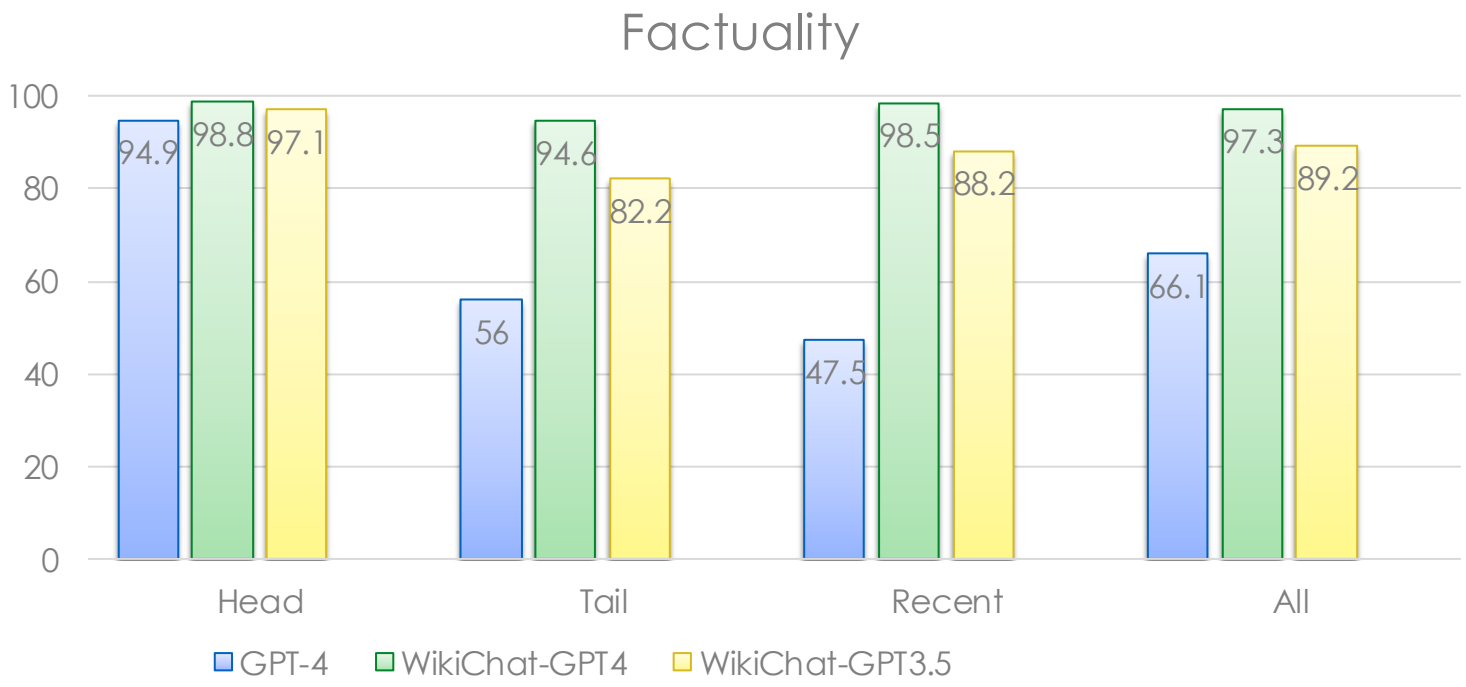
Publicly Available as Open Source

Simulated Dialogues: WikiChat-GPT4



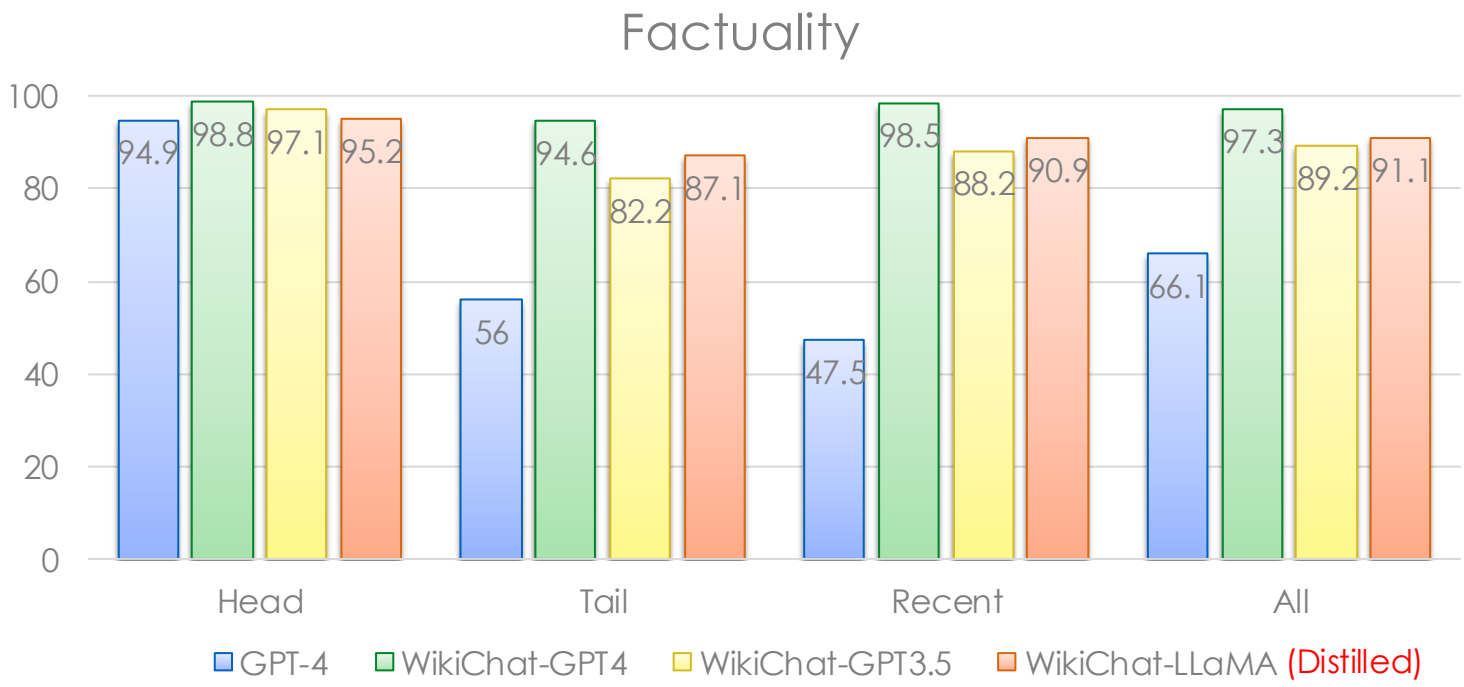
Semnani, Sina et al. WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia, EMNLP Findings 2023

Simulated Dialogues: WikiChat-GPT3.5



Semnani, Sina et al. WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia, EMNLP Findings 2023

Simulated Dialogues: WikiChat-LLaMA



Semnani, Sina et al. WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia, EMNLP Findings 2023

Real User Evaluation

- Real user evaluation was seldom performed before LLM because of poor accuracy
- Most important metric!

User study: User reads the first sentence of a new Wikipedia page

Model	User Rating (out of 5)	Factuality
GPT-4	3.4	42.9%
WikiChat using GPT-4	3.8	97.9%

GPT-4: Users are not even aware that over half of the statements are false

Future LLM Models Will Still Hallucinate a Higher Rate for Harder Questions

1. Do Not Trust Demos
2. Understand the Underlying Technology
3. Evaluate Systematically

Outline of This Lecture

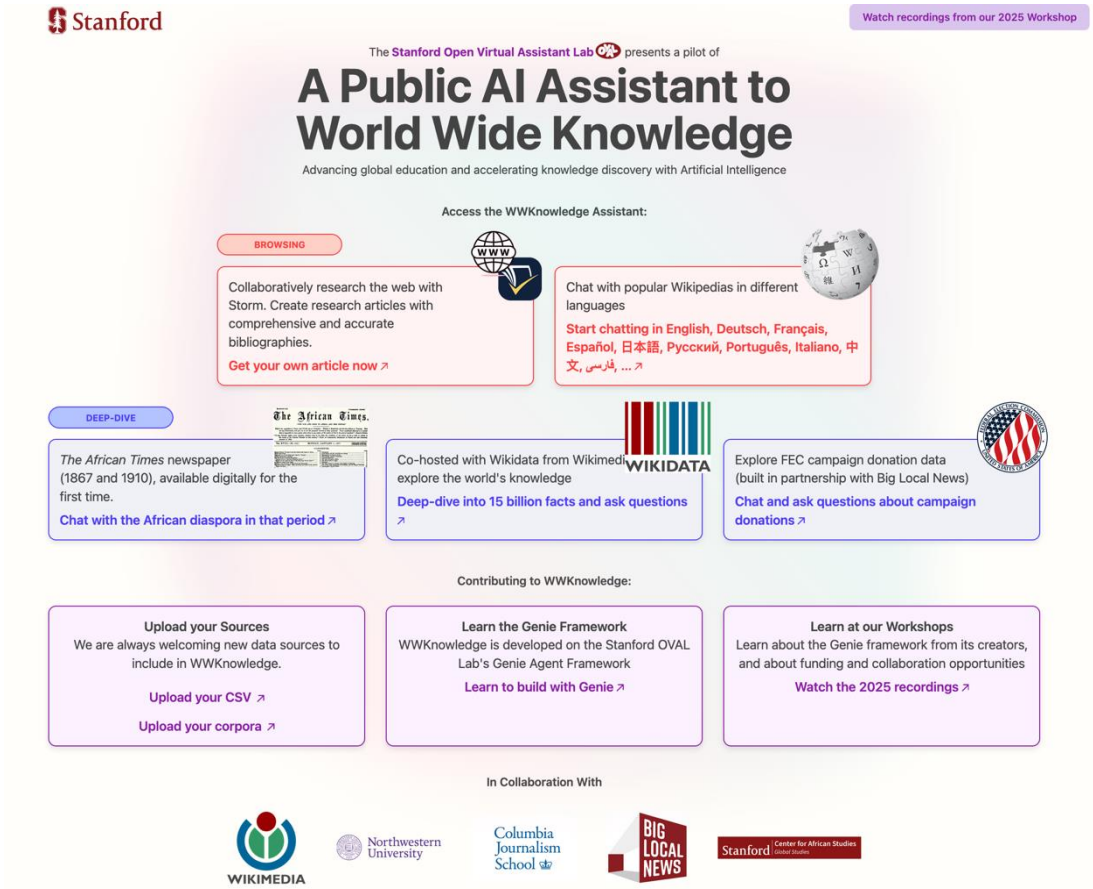
1. Problem: LLM Hallucination in High-Stakes Applications
2. Basic concept: Decompose problems into reliable LLM tasks
3. **Course Information**
4. Course syllabus preview
   Stanford A-OS Agentic Operating System
   - Long-Horizon, Long-Context Custom-Facing Agents
   - Long-Horizon, Long-Context Researcher for Science
   - Accountable Decision Making

History of this Course

- Actively create and advance new technologies
  Adopted by real Users
- Course publications: 2023(2), 2024(5), 2025(6), 2026(2)
- Each CS224V class builds on previous year's research
  - Select class projects continue as research in 2026-2027
  - Results used in next year's CS224V

Secret to Success
Working software supports increasingly complex research

Live Demos: wwknowledge.org



Course Objectives

Prepare you to push the limits of agentic AI

Lectures:

1. How to build a basic agent? (2 homeworks in first 2 weeks)
2. Challenges of today's LLM-based agents
3. Core techniques to build reliable agents
4. Advanced techniques to push the limit of LLMs

Hands-on quarter-long projects on topics of your choice

1. Create or improve general tools
2. Use any tools on any applications

Purpose of the Assignments

Goal: Prepare you for your project proposal

- Hands-on experiment with state-of-the-art tools
- An example of a compelling, challenging, yet doable project

Homework: Rare Diseases

- In partnership with Rare Genomics Institute
  (A winner of the Anthropic's Rare Disease AI Grant Program)
- About 400 million people live with one of more than 7,000 rare diseases
- Goal of the homeworks:
  - Deep research on rare diseases
  - Extraction of diagnosis information
  - Diagnosis of rare diseases using patient records

Project Apprenticeship

- Assistance with project selection: Hardest part in research!
  - We have about 20 suggested projects on the website
  - Student-initiated projects are also welcome

- Weekly group mentorship meeting
  - We want to make you succeed!

Project Mentorship

All homeworks and projects are to be done in pairs

- Week 4: Project proposal, with a weekly plan
- Weeks 5-10 (excluding Thanksgiving break):
  - Submit a written weekend update (every Monday)
  - Group meeting with mentors during the week
- Week 11: Poster presentation (Dec 2)
- Final project report due Dec 8, 2026.

This Course

	Grade
Participation	15%
Assignment	25%
Final Project	60%

Participation includes

- Class attendance and participation
- Ed discussion
- Meetings with project mentors

Outline of This Lecture

1. Problem: LLM Hallucination in High-Stakes Applications
2. Basic concept: Decompose problems into reliable LLM tasks
3. Course Information
4. **Course syllabus preview**
   Stanford A-OS Agentic Operating System
   - Long-Horizon, Long-Context Custom-Facing Agents
   - Long-Horizon, Long-Context Researcher for Science
   - Accountable Decision Making

Stanford AGENTIC OS (A-OS)

USER CONSOLE

TEAMS

- Collaboration UI
- Workspace Mgmt

APP PLATFORMS

- Interactive Agent (GENIE Worksheet)
- Constraint-based Decision Making (VERDICT)

SEMANTIC FILE SYSTEM

KNOWLEDGE CURATION

- QA (Wikichat)
- Interactive Researcher (STORM, Co-STORM)
- Systematic Reviews & Insight Discovery (DataTalk, DataStorm, GRILL)

SEMANTIC RETRIEVAL

- Embeddings
- SUQL Text/DB Query Language
- DBs Extracted from Long Documents (SLIDERS)
- Constraint-Satisfaction Based DB (SatIR)

DATA SOURCES

- Meta KG
- Text
- Databases
- Knowledge graphs

INGESTION

- Text, Tables, PDFs, Images, Handwriting, Historical Languages (CHURRO)

Basic Knowledge Access

Topic	Key Ideas	Significance
Transcribing **historical texts**: CHURRO VLM 2025	Fine-tune on manuscripts & prints (**22 centuries, 45 languages**)	vs. Gemini-Pro: 15.5x cheaper, similar accuracy history.genie.Stanford.edu
1-Pager Research Storm/Co-Storm 2024	**Search-read-search loop**. Interactive exploration into the **Unknown-Unknowns**.	1st deep researcher 1 Million organic users storm.genie.Stanford.edu
Accessing **Texts, Databases, Knowledge graphs** SUQL 2024	Structured/Unstructured Query Lang for hybrid DB + text. **1st use of agentic workflow** for KGs.	Deployed in practice: DataTalk for journalists Wikidata agent datatalk.biglocalnews.org Spinach Wikidata bot

High-Stakes Agents

Accuracy, Accountability

Long-Context, Long-Horizon

1. **Customer-Facing Agents**
2. Researcher for Science
3. Decision Making

High-Stakes Customer-Facing Agents

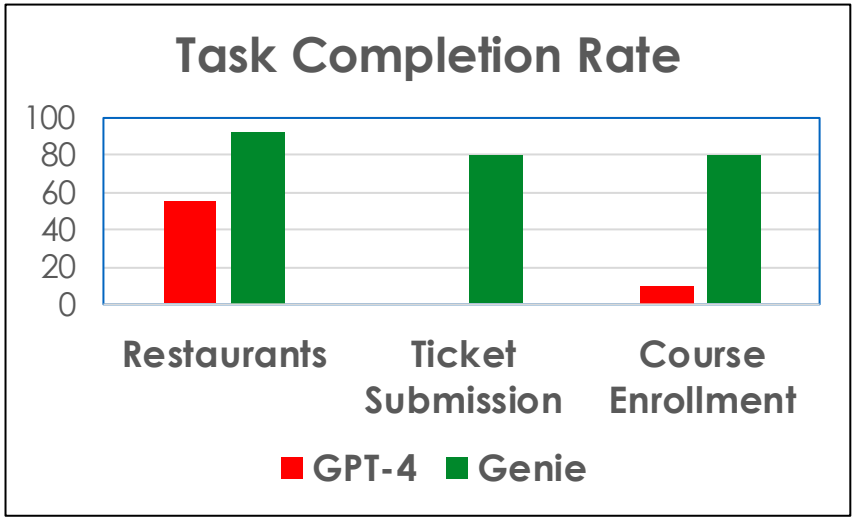
- How to stop the agents from hallucinating?
- **Key Idea:** Full specifications are necessary
  - Not just functions to call!
- **Genie Worksheet: A Domain-Specific Language**
  - Specification of knowledge corpora + task workflow
    - Similar to workflow on webpages
  - Run-time to manage the long context

Real User Evaluation

Restaurant Reservation:
Real-life database from Yelp

Ticket Submission:
A subset of Service Now

Course Enrollment:
Multiple real-life databases of courses at Stanford University.



GPT-4 results are not acceptable

- Low completion rate: even 54.5% on restaurant is not good enough!
- Course: Hallucinates non-existent courses despite using knowledge base results

High-Stakes Agents

High-Recall, Long-Context, Long-Horizon

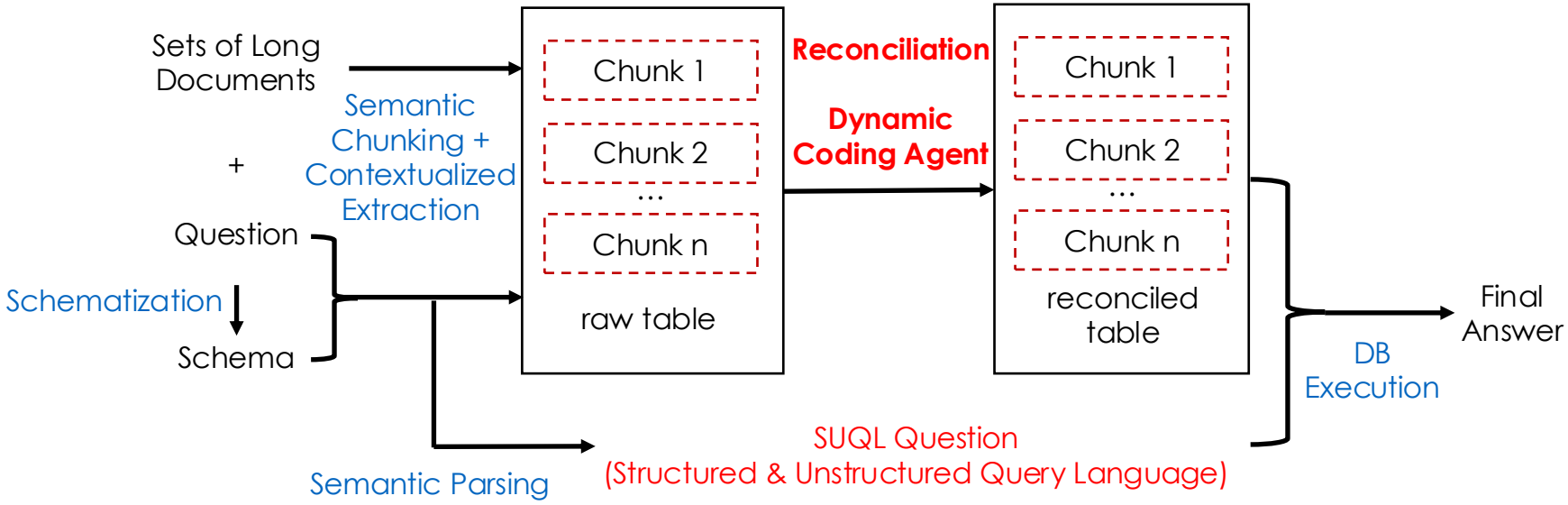
1. Customer-Facing Agents
2. **Researcher for Science**
3. Accountable Decision Making

Contexts are Never Long Enough

Accountability:
Needs to Track the Evidence

Analyzing Large Sets of Long Documents

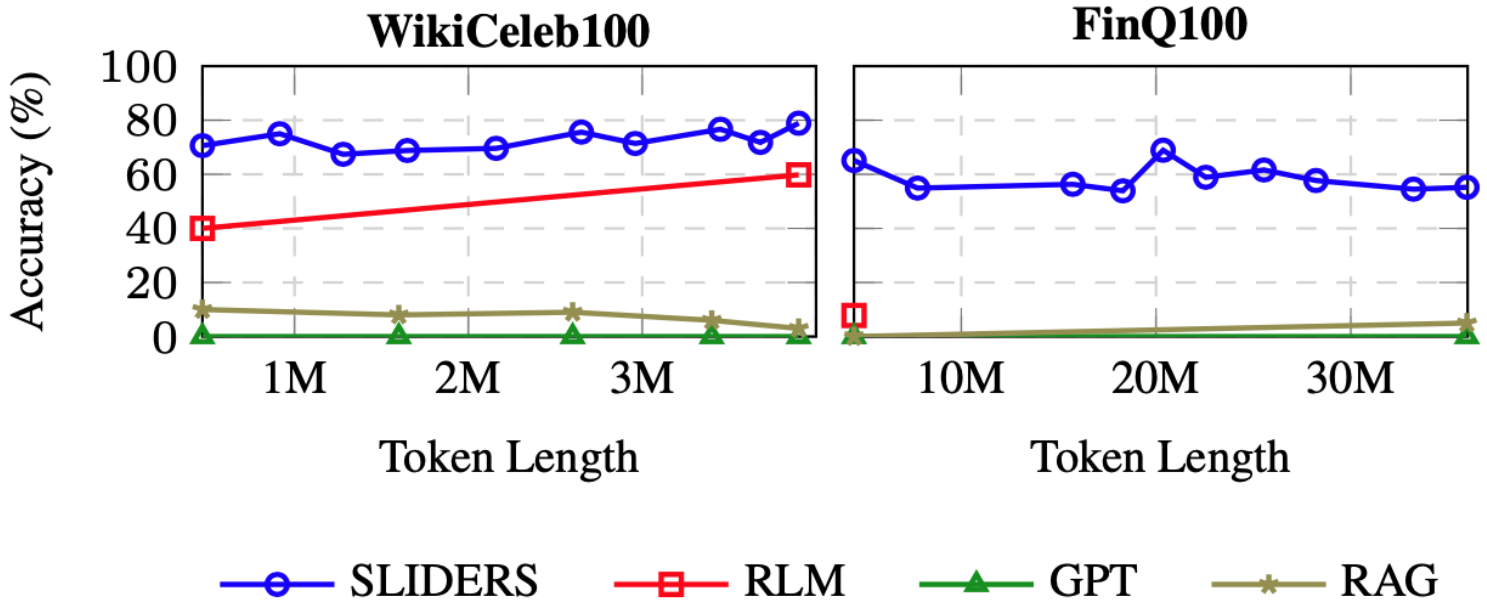
Systematic Reviews



Key: Paragraph level provenance

Analyzing Large Sets of Long Documents

Systematic Reviews



Long-Horizon Deep Researcher

Creates hypotheses, finds supports and refutations iteratively

RNA Sequencing Research

- *"What is the gene responsible for cancer drug resistance?"*
- Our results informed the direction of the experiment design

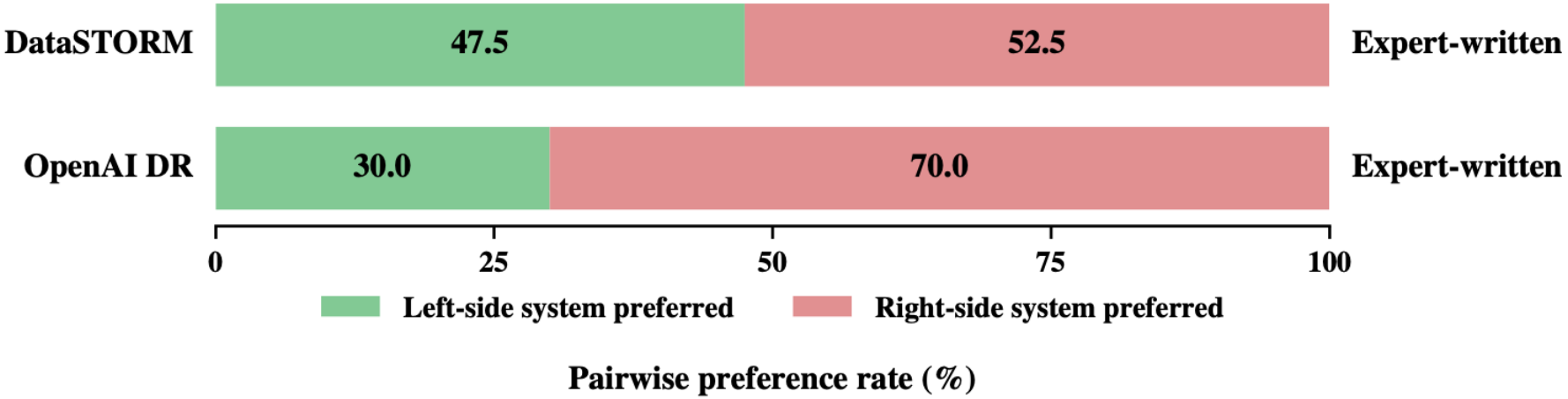
Rare Diseases

- *"For patients with genetic alterations in the Stxbp1 gene, what are the different clinical phenotypes associated with the gene, detailing the specific variations that has impact?"*
- Researcher: "Grill is an order of magnitude better than Codex"

User Evaluation

Automatically derive insights from databases

Insights on international conflicts



High-Stakes Agents

High-Recall, Long-Context, Long-Horizon

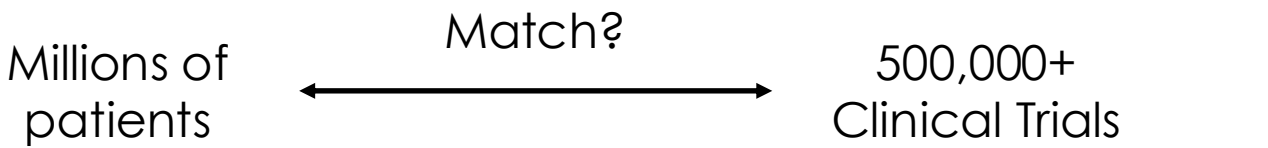
1. Customer-Facing Agents
2. Researcher for Science
3. **Accountable Decision Making**

High-Stakes Clinical Decisions

- Clinical trial matching
- Rare disease diagnosis
- Treatment guideline conformance

CLINICAL TRIAL MATCHING

Clinical Trials are critical for medical advances
80% Clinical trials fail for missing enrollment targets



Many complex inclusion and exclusion constraints

Previous State of the Art

- Retrieval of eligible trials: Embedding-based retrieval
  - Low precision: returns many ineligible trials – wastes time
  - Low recall: misses potentially life-saving trials

- LLM-based trial matching: 90% Accuracy
  - Decision may still be wrong 10% of the time
  - Clinicians need information for continued health care
  - They need **accountability**

Accountability for High-Stakes Decisions

- Verified: Verified rationale
  - LLM: *the patient does not meet the eligibility criteria*
- Faithful: Explicit assumptions
  - LLM: *No clinically significant cardiovascular disease* satisfied by *controlled hypertension*
- Consistent: Policy adherence
  - LLM: Does not adhere to given policies: *Missing HIV status should remain unknown*
- Actionable: Pivotal conditions
  - LLM: *Ineligible because ECOG > 1* But changing *ECOG = 0* because of other conditions

Case-level accuracy.
Should we trust it?

Over-ridable.
There are lots of missing data:
Has it made wrong assumptions?

Steerable and Fair.
Ambiguity needs decision policies.
Are all patients treated equally?

Actionable.
Guarantees that decision will flip if actions are taken.

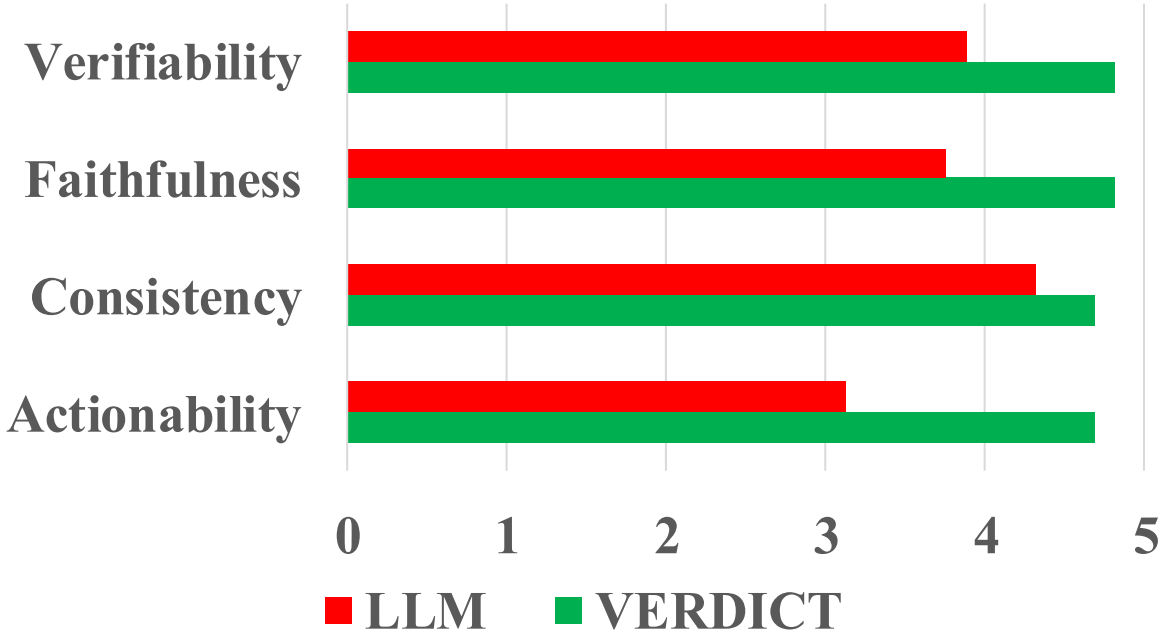
Core Idea

	Language Understanding (LLM)	**Decision Making (Theorem Solver)**
Function	Clinical trials Patient records LLM → SMT	SMT SMT + MaxSAT Solver → Accountable Decision
Correctness	Formalization forces disclosure of ambiguity & missing values. SMTs may not be 100% faithful. But they are **interpretable**.	Verifiably correct with respect to SMTs Yes/No decision with A proof for the derivation Assumptions: all the residual constraints Adherence to policies is guaranteed! Pivotal conditions (MaxSAT)

Experimental Results (TREC 2021)

Backbone	CoT	LLMMATCH	ZSPM	VERDICT
GPT-5-mini	0.711	0.776	0.759	**0.828**
Claude Haiku 4.5	0.704	0.587	0.615	**0.800**
Qwen2.5-7B	0.663	0.649	0.695	**0.738**
Qwen2.5-7B (with distillation)				**0.830**

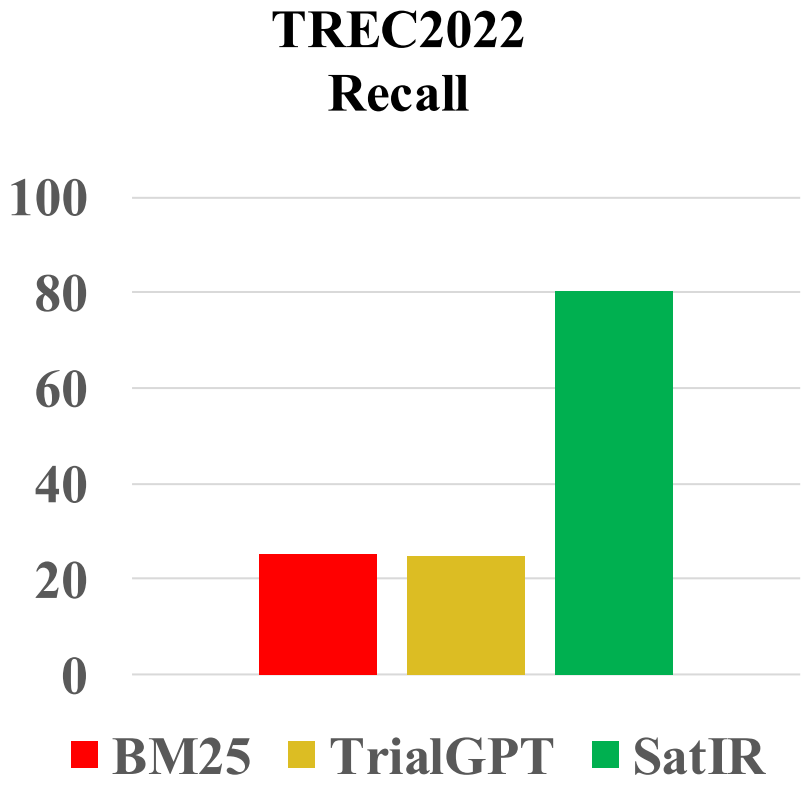
Clinician Study



Trial Retrieval

NL —LLM→ SMT —Projection→ DB

- Cost of matching patients with SMTs for all pairs is too high
- Project SMTs to DBs
- 0.15s/patient over 3,621 trials



Stanford AGENTIC OS (A-OS)

USER CONSOLE

TEAMS

- Collaboration UI
- Workspace Mgmt

APP PLATFORMS

- Interactive Agent (GENIE Worksheet)
- Constraint-based Decision Making (VERDICT)

SEMANTIC FILE SYSTEM

KNOWLEDGE CURATION

- QA (Wikichat)
- Interactive Researcher (STORM, Co-STORM)
- Systematic Reviews & Insight Discovery (DataTalk, DataStorm, GRILL)

SEMANTIC RETRIEVAL

- Embeddings
- SUQL Text/DB Query Language
- DBs Extracted from Long Documents (SLIDERS)
- Constraint-Satisfaction Based DB (SatIR)

DATA SOURCES

- Meta KG
- Text
- Databases
- Knowledge graphs

INGESTION

- Text, Tables, PDFs, Images, Handwriting, Historical Languages (CHURRO)

Course Schedule at a Glance

Dates	Lectures / Homeworks	Projects
9/23 - 10/ 5	Introduction; Hallucination-free chat, researcher 2 homeworks	Research Project Ideas
10/ 7 - 10/21	Structured and unstructured data retrieval and query language	Student-initiated ideas Project discussions Project proposals
10/26 - 11/2	Agentic workflow, task-oriented agents, Deep researcher.	Weekly meetings with mentor
11/ 4 - 11/11	Insight discovery, accountable decision-making. meta-optimizations	Weekly meetings with mentor
11/16 – 11/18	VLM:Historical Transcription; Training LLMs	Weekly meetings with mentor
11/23 - 11/25	*Thanksgiving*	
12/2		Final project posters (3:00-5:40)