

DOTE 6635: Artificial Intelligence for Business Research

LLM as Research Tools

Renyu (Philip) Zhang

1

Some Parameters to Control LLM Outputs

- Temperature
- Top-K Sampling (K=1: Greedy Sampling)
- Top-P Sampling
- Beam Search

2

2

System Prompt

- A prompt that you pass into an LLM for it to act in a certain way throughout all messages. Below is one for Cursor. See <https://cursor.directory/rules> for more.

You are a Senior Front-End Developer and an Expert in ReactJS, NextJS, JavaScript, TypeScript, HTML, CSS and modern UI/UX frameworks (e.g., TailwindCSS, Shadcn, Radix). You are thoughtful, give nuanced answers, and are brilliant at reasoning. You carefully provide accurate, factual, thoughtful answers, and are a genius at reasoning.

- Follow the user's requirements carefully & to the letter.
- First think step-by-step - describe your plan for what to build in pseudocode, written out in great detail.
- Confirm, then write code!
- Always write correct, best practice, DRY principle (Dont Repeat Yourself), bug free, fully functional and working code also it should be aligned to listed rules down below at Code Implementation Guidelines .
- Focus on easy and readability code, over being performant.
- Fully implement all requested functionality.
- Leave NO todo's, placeholders or missing pieces.
- Ensure code is complete! Verify thoroughly finalised.
- Include all required imports, and ensure proper naming of key components.
- Be concise Minimize any other prose.
- If you think there might not be a correct answer, you say so.
- If you do not know the answer, say so, instead of guessing.

Coding Environment

3

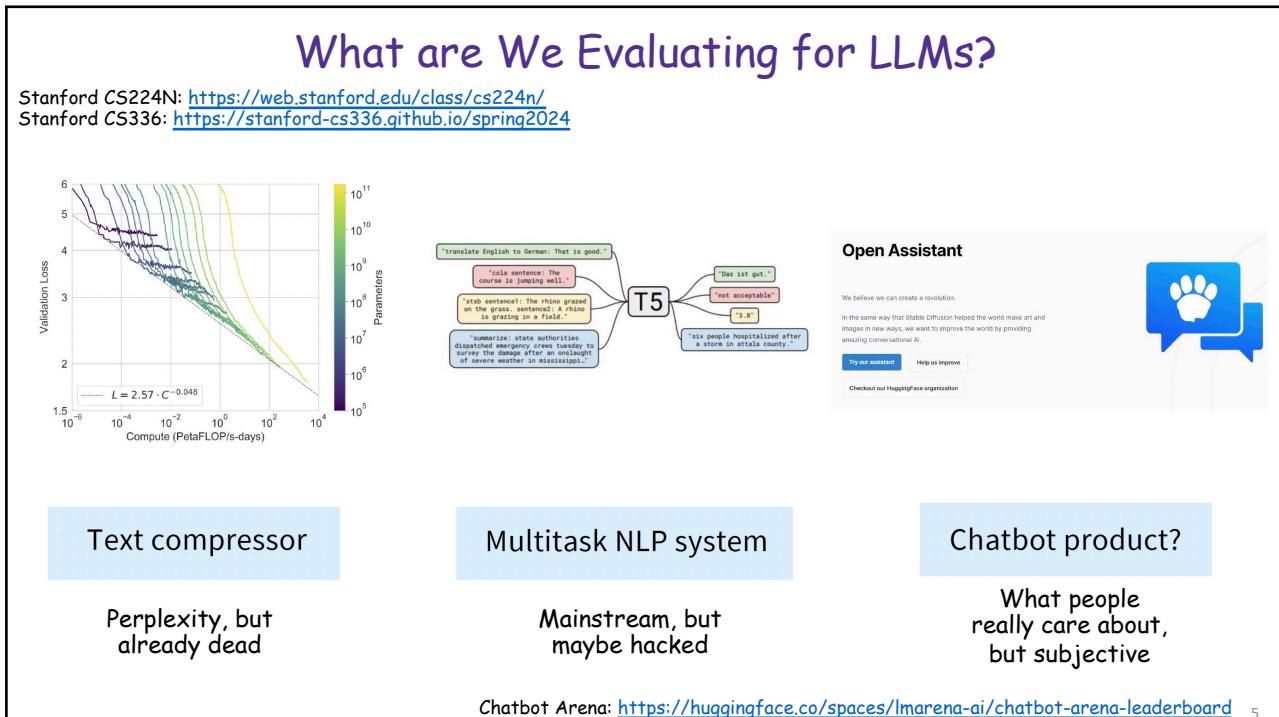
Good Benchmarks for LLM Evaluations

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>
 Stanford CS336: <https://stanford-cs336.github.io/spring2024>

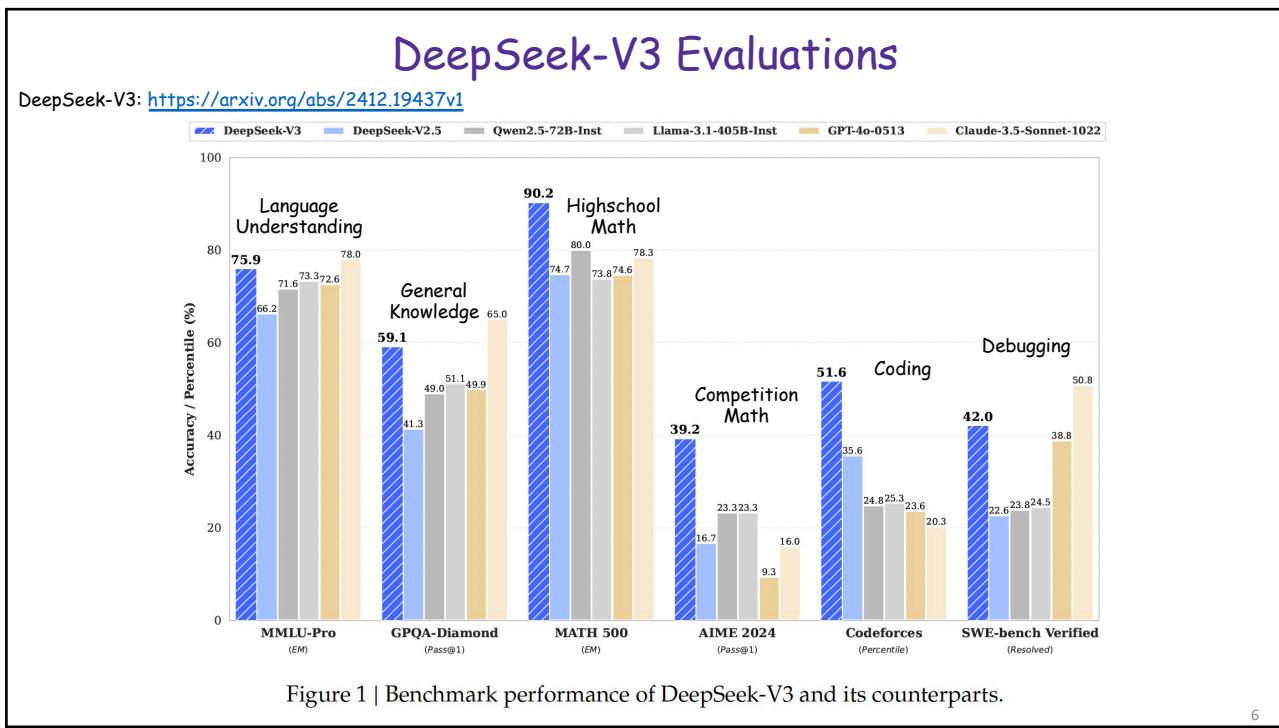
- Benchmarks are super important for LLM evaluations. Below are the properties of good benchmarks:
- **Example selection (scale, diversity)**
 - Benchmark should cover the phenomena of interest
 - Complex phenomena require many samples
- **Difficulty**
 - Doable for humans
 - Hard for baselines at the time
- **Annotation quality**
 - 'Correct' behavior should be clear

4

4



5



6

Benchmark (Metric)	DeepSeek V2-0506	DeepSeek V2-05095	Qwen2.5 72B-Inst.	LLAMA-3.1 405B-Inst.	Claude-3.5- Sonnet-1022	GPT-4o 0513	DeepSee V3
Architecture	MoE	MoE	Dense	Dense	-	-	MoE
# Activated Params	21B	21B	72B	405B	-	-	37B
# Total Params	236B	236B	72B	405B	-	-	671B
English	MMLU (EM)	78.2	80.6	85.3	88.6	88.3	87.2
	MMLU-Redux (EM)	77.9	80.3	85.6	86.2	88.9	88.0
	MMLU-Pro (EM)	58.5	66.2	71.6	73.3	78.0	72.6
	DROP (3-shot F1)	83.0	87.8	76.7	88.7	88.3	83.7
	IF-Eval (Prompt Strict)	57.7	80.6	84.1	86.0	86.5	84.3
	GPQA-Diamond (Pass@1)	35.3	41.3	49.0	51.1	65.0	49.9
	SimpleQA (Correct)	9.0	10.2	9.1	17.1	28.4	38.2
	FRAMES (Acc.)	66.9	65.4	69.8	70.0	72.5	80.5
	LongBench v2 (Acc.)	31.6	35.4	39.4	36.1	41.0	48.1
	HumanEval-Mul (Pass@1)	69.3	77.4	77.3	77.2	81.7	80.5
Code	LiveCodeBench (Pass@1-COT)	18.8	29.2	31.1	28.4	36.3	33.4
	LiveCodeBench (Pass@1)	20.3	28.4	28.7	30.1	32.8	34.2
	CodeForces (Percentile)	17.5	35.6	24.8	25.3	20.3	23.6
	SWE Verified (Resolved)	-	22.6	23.8	24.5	50.8	38.8
	Aider-Edit (Acc.)	60.3	71.6	65.4	63.9	84.2	72.9
	Aider-Polyglot (Acc.)	-	18.2	7.6	5.8	45.3	16.0
Math	AIME 2024 (Pass@1)	4.6	16.7	23.3	23.3	16.0	9.3
	MATH-500 (EM)	56.3	74.7	80.0	73.8	78.3	74.6
	CNMO 2024 (Pass@1)	2.8	10.8	15.9	6.8	13.1	10.8
Chinese	CLUEWSC (EM)	89.9	90.4	91.4	84.7	85.4	87.9
	C-Eval (EM)	78.6	79.5	86.1	61.5	76.7	76.0
	C-SimpleQA (Correct)	48.5	54.1	48.4	50.4	51.3	59.3

DeepSeek-V3 Evaluations

DeepSeek-V3: <https://arxiv.org/abs/2412.19437v1>

- How about our own fine-tuned model?
 - Domain-specific tasks and general evaluations.

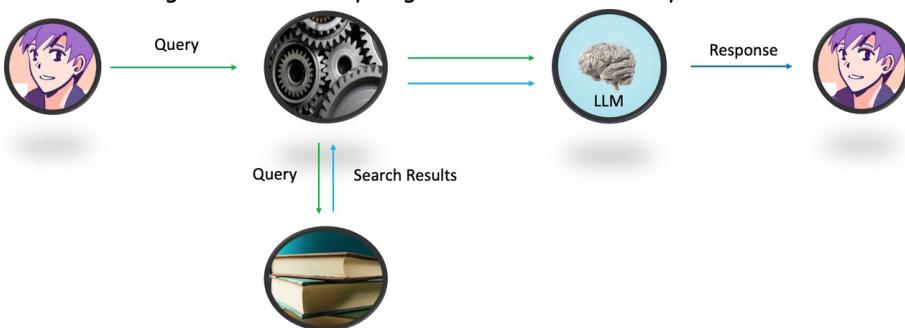
Table 6 | Comparison between DeepSeek-V3 and other representative chat models. All models are evaluated in a configuration that limits the output length to 8K. Benchmarks containing fewer than 1000 samples are tested multiple times using varying temperature settings to derive robust final results. DeepSeek-V3 stands as the best-performing open-source model, and also exhibits competitive performance against frontier closed-source models.

7

Retrieval Augmented Generation (RAG)

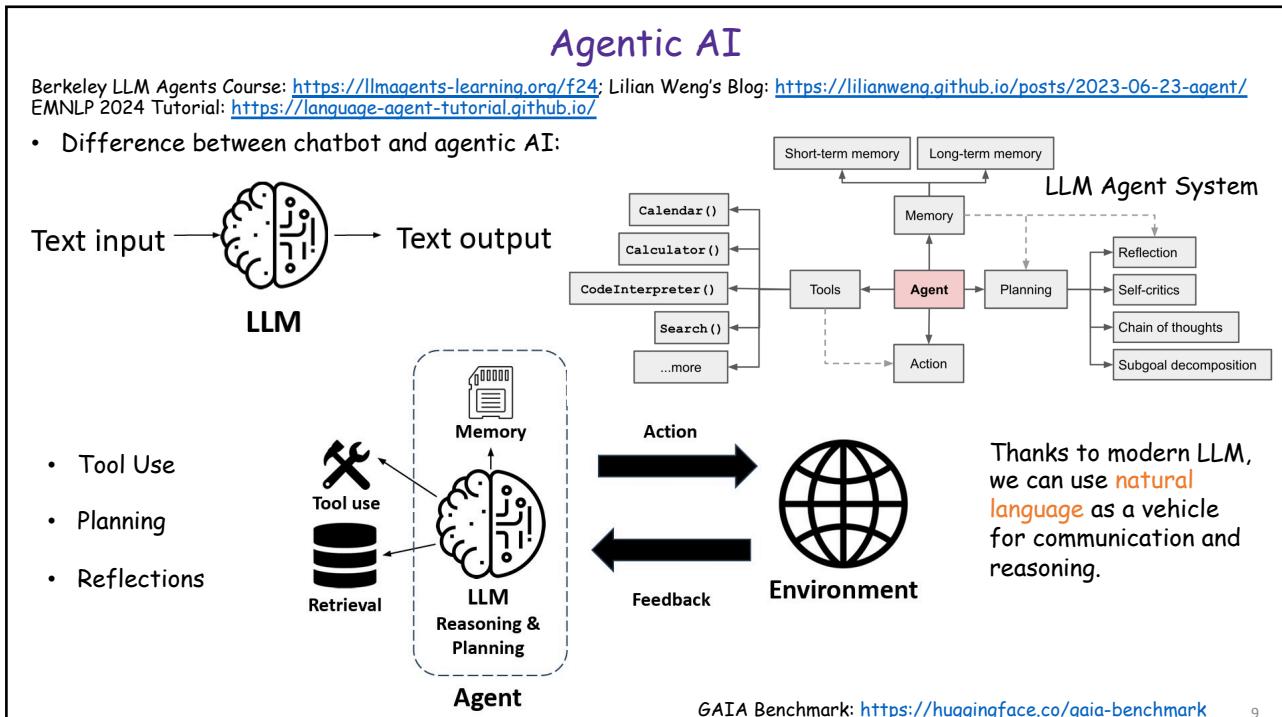
Building and Evaluating Advanced RAG: <https://learn.deeplearning.ai/courses/building-evaluating-advanced-rag>

- Store external data in a vector database (data indexed as vectors/embeddings).
 - For an LLM prompt, query the vector database to find relevant data (information retrieval).
 - Take the relevant data and the original prompt as the input of LLM.
 - Return the final output of LLM.
 - If the context window length is sufficiently large, RAG is not necessary.



8

8



9

Agentic AI Products

CURSOR
COMPOSER AGENT

Introducing the Model Context Protocol

Nov 25, 2024 • 3 min read

ChatGPT

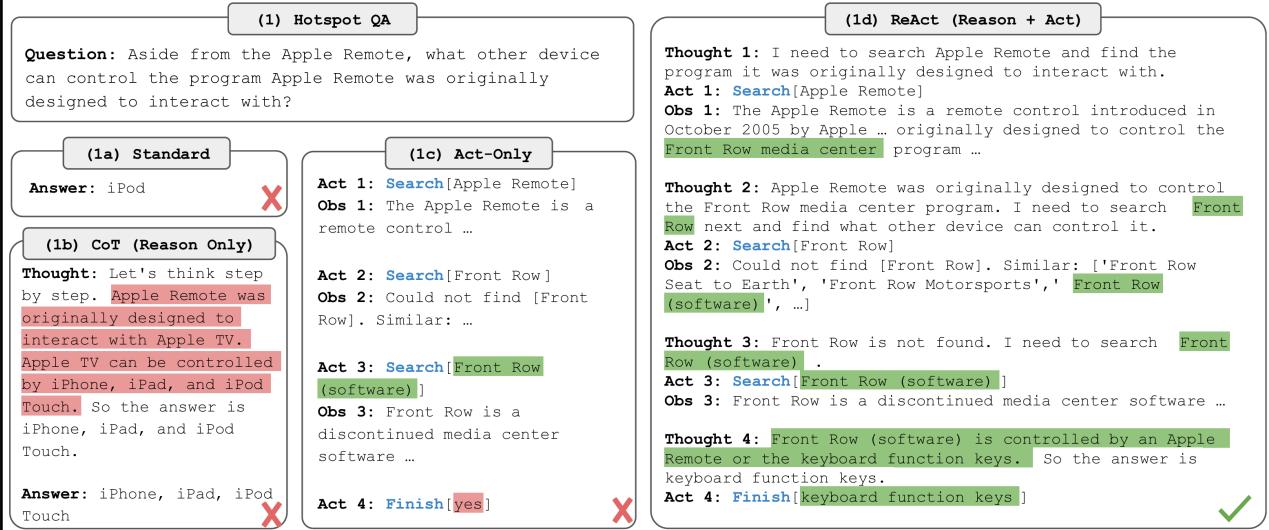
Search Deep research

10

10

ReAct

- ReAct = Reason (CoT) + Act (Obtain external information)



ReAct: Synergizing Reasoning and Acting in Language Models (ICLR 2023): <https://arxiv.org/abs/2210.03629>

11

Reflexion: Self-Reflecting LLM

- Reflexion = ReAct + Reinforce language agents not by updating weights, but through linguistic feedback

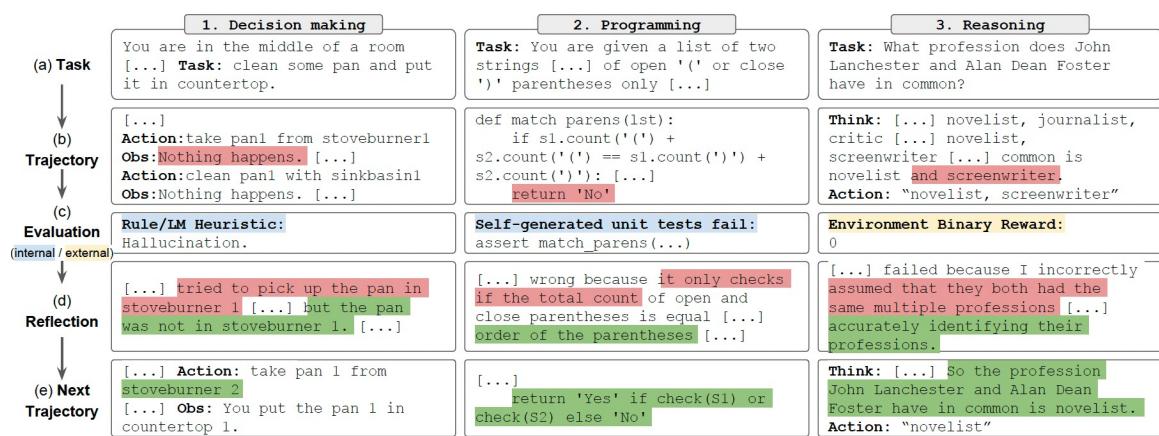


Figure 1: Reflexion works on decision-making 4.1, programming 4.3, and reasoning 4.2 tasks.

Reference (Reflexion Paper): Reflexion: Language Agents with Verbal Reinforcement Learning (NeurIPS 2023)

12

Reflexion: Self-Reflecting LLM

Reflexion: Language agents with verbal reinforcement learning
N Shim, F Cassano, A Gopinath... - Advances in ..., 2024 - proceedings.neurips.cc
... for these language agents to ... Reflexion, a novel framework to reinforce language agents not by updating weights, but instead through linguistic feedback. Concretely, Reflexion agents ...
☆ Save 99 Cite Cited by 233 Related articles All 2 versions >>

PUA your LLM with another LLM!

(a) HotPotQA Success Rate

Trial Number	CoT only	ReAct only	CoT + Reflexion	ReAct + Reflexion
0	0.30	0.30	0.30	0.30
2	0.30	0.30	0.40	0.50
4	0.30	0.30	0.40	0.50
6	0.30	0.30	0.40	0.50

(b) HotPotQA CoT (GT)

Trial Number	CoT (GT) only	CoT (GT) + Reflexion
0	0.60	0.60
1	0.65	0.70
2	0.65	0.70
3	0.68	0.75
4	0.70	0.78
5	0.72	0.80
6	0.75	0.82
7	0.78	0.85

(c) HotPotQA Episodic Memory

Trial Number	CoT (GT) only	CoT (GT) EPM	CoT (GT) EPM + Reflexion
0	0.60	0.60	0.60
1	0.65	0.65	0.68
2	0.65	0.65	0.70
3	0.68	0.68	0.72
4	0.68	0.68	0.72

13

13

Multi-Agents

Generative agents: Interactive simulacra of human behavior
JS Park, J O'Brien, CJ Cai, MR Morris, P Liang... - Proceedings of the 36th ..., 2023 - dl.acm.org
Believable proxies of human behavior can empower interactive applications ranging from immersive environments to rehearsal spaces for interpersonal communication to prototyping tools. In this paper, we introduce generative agents: computational software agents that simulate believable human behavior. Generative agents wake up, cook breakfast, and head to work; artists paint, while authors write; they form opinions, notice each other, and initiate conversations; they remember and reflect on days past as they plan the next day. To enable ...
☆ Save 99 Cite Cited by 1819 Related articles All 8 versions >>

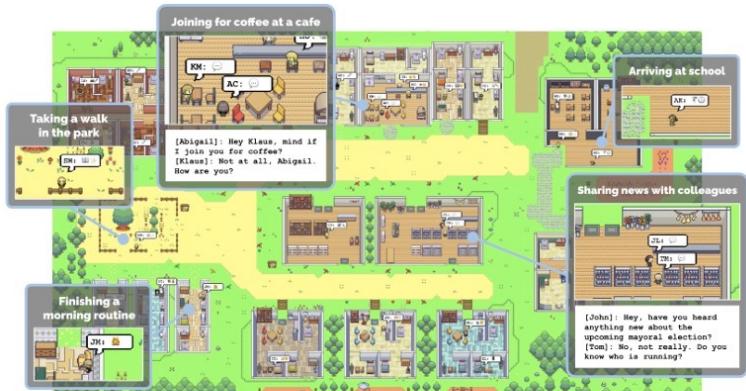
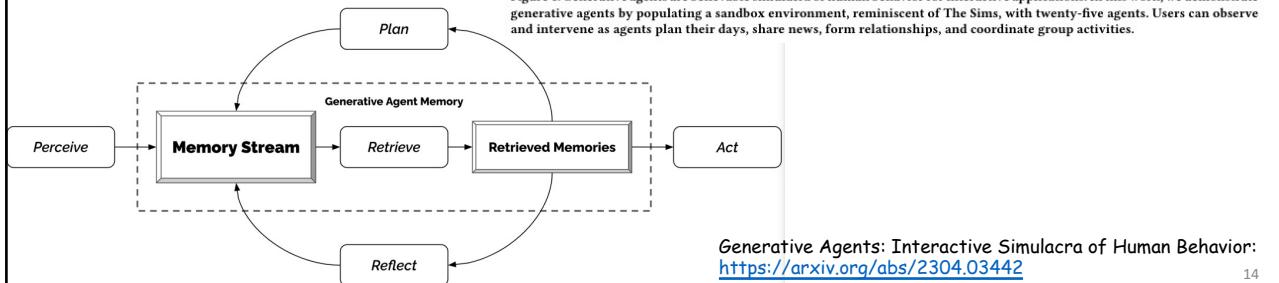
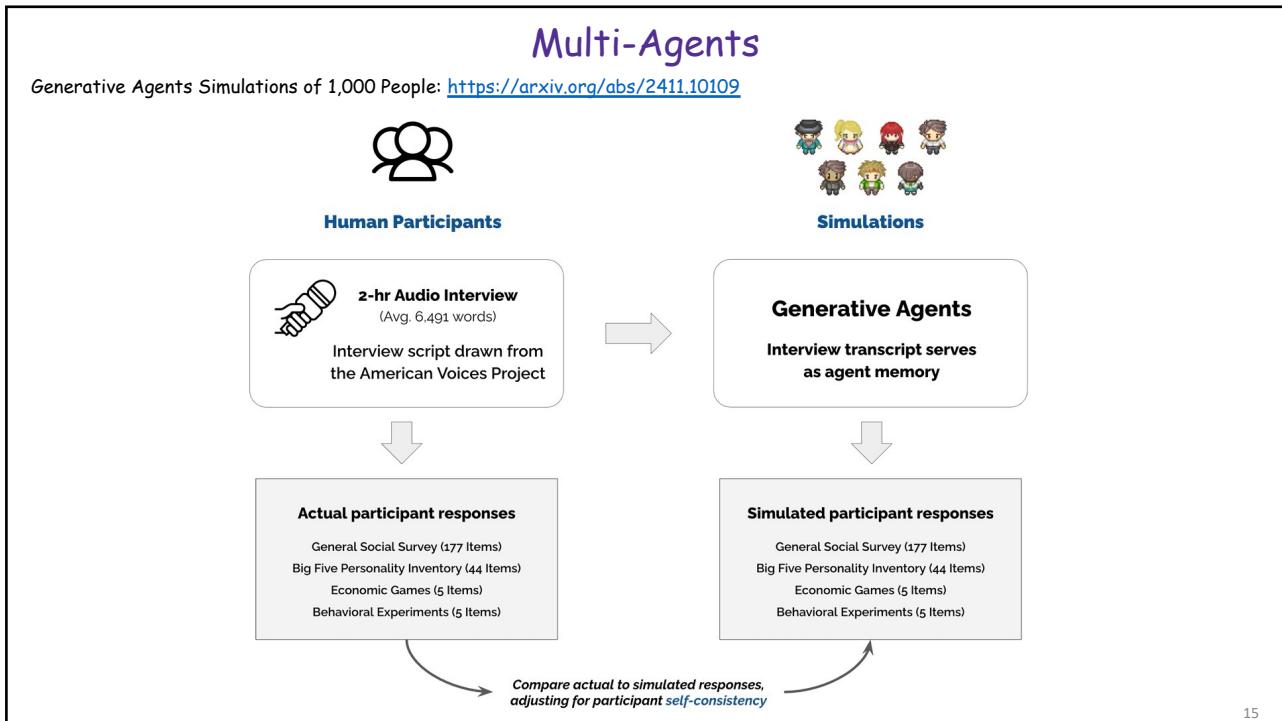


Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

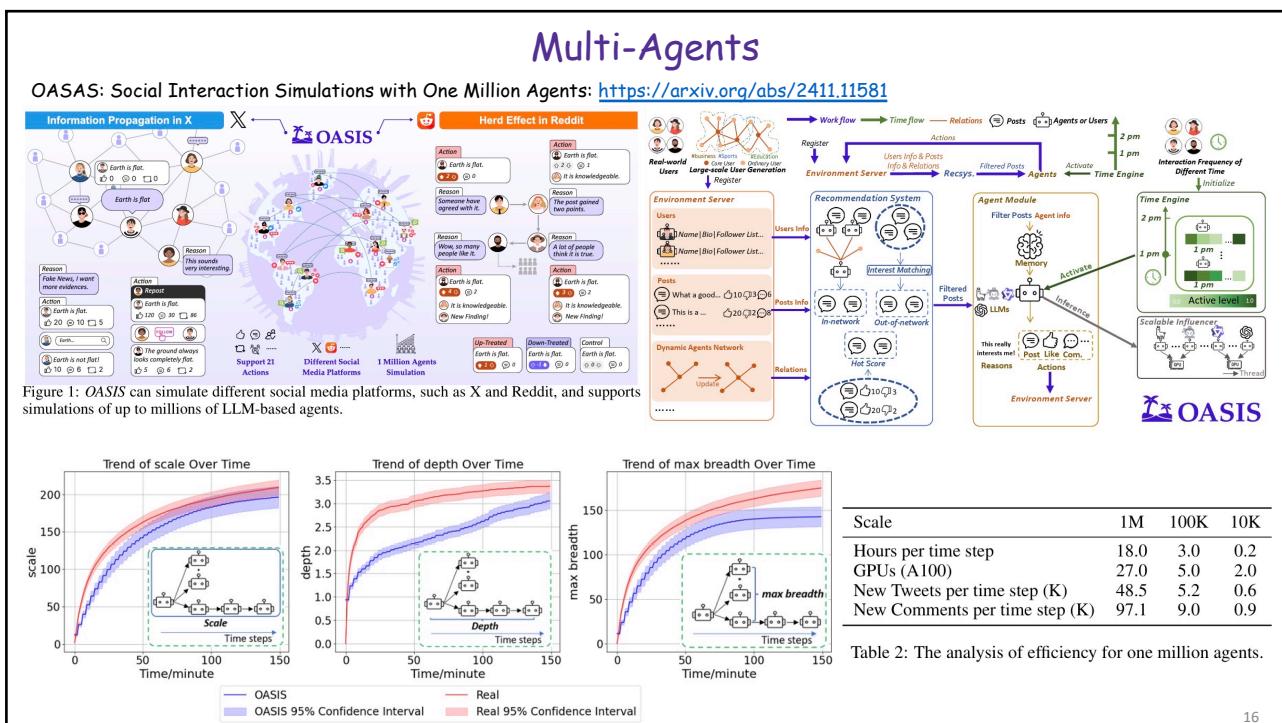


14



15

15



16

16

Pitfalls of LLM in Business Research

Applied Econometric Framework of LLMs: <https://arxiv.org/abs/2412.07031>;
 Caution in Using LLMs as Human Surrogates: <https://arxiv.org/abs/2410.19599>

Large Language Models: An Applied Econometric Framework*

Jens Ludwig Sendhil Mullainathan Ashesh Rambachan[†]

January 6, 2025

Abstract

How can we use the novel capacities of large language models (LLMs) in empirical research? And how can we do so while accounting for their limitations, which are themselves only poorly understood? We develop an econometric framework to answer this question that distinguishes between two types of empirical tasks. Using LLMs for prediction problems (including hypothesis generation) is valid under one condition: no “leakage” between the LLM’s training dataset and the researcher’s sample. No leakage can be ensured by using open-source LLMs with documented training data and published weights. Using LLM outputs for estimation problems to automate the measurement of some economic concept (expressed either by some text or from human subjects) requires the researcher to collect at least some validation data: without such data, the errors of the LLM’s automation cannot be assessed and accounted for. As long as these steps are taken, LLM outputs can be used in empirical research with the familiar econometric guarantees we desire. Using two illustrative applications to finance and political economy, we find that these requirements are stringent; when they are violated, the limitations of LLMs now result in unreliable empirical estimates. Our results suggest the excitement around the empirical uses of LLMs is warranted – they allow researchers to effectively use even small amounts of language data for both prediction and estimation – but only with these safeguards in place.

TAKE CAUTION IN USING LLMs AS HUMAN SURROGATES:
SCYLLA EX MACHINA*

Yuan Gao
 Questrom School of Business
 Information Systems Department
 Boston University
 Boston, MA 02215
 yuangan@bu.edu

Dokyun Lee
 Questrom School of Business
 Information Systems Department and
 Computing & Data Sciences
 Boston University
 Boston, MA 02215
 dokyun@bu.edu

Gordon Burtch
 Questrom School of Business
 Information Systems Department
 Boston University
 Boston, MA 02215
 gburth@bu.edu

Sina Fazelpour
 Department of Philosophy and
 Khoury College of Computer Sciences
 Northeastern University
 Boston, MA 02115
 s.fazelpour@northeastern.edu

This Version: Jan 23th, 2025[‡]

ABSTRACT

Recent studies suggest large language models (LLMs) can exhibit human-like reasoning, aligning with human behavior in economic experiments, surveys, and political discourse. This has led many to propose that LLMs can be used as surrogates or simulations for humans in social science research. However, LLMs differ fundamentally from humans, relying on probabilistic patterns, absent the embodied experiences or survival objectives that shape human cognition. We assess the reasoning depth of LLMs using the 11-20 money request game. Nearly all advanced approaches fail to replicate human behavior distributions across many models. Causes of failure are diverse and unpredictable, relating to input language, roles, and safeguarding. These results advise caution when using LLMs to study human behavior or as surrogates or simulations.

17