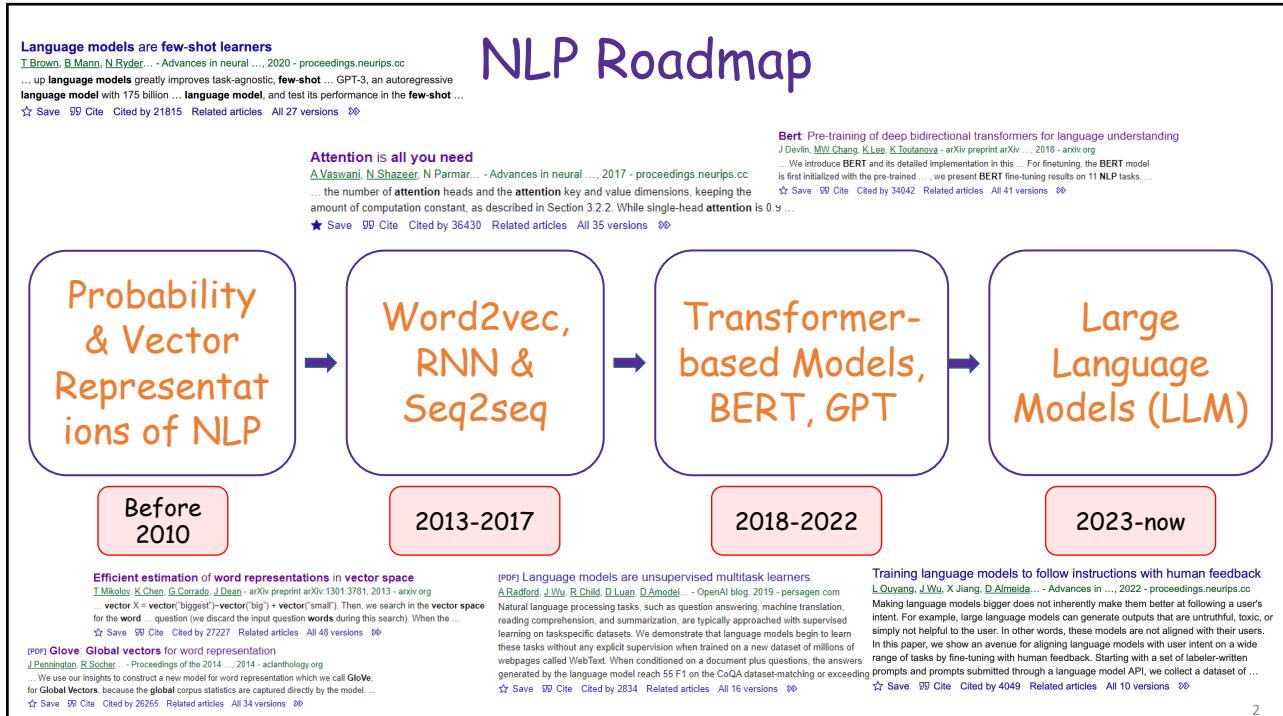


## DSME 6635: Artificial Intelligence for Business Research

# Large Language Models

Renyu (Philip) Zhang

1



2

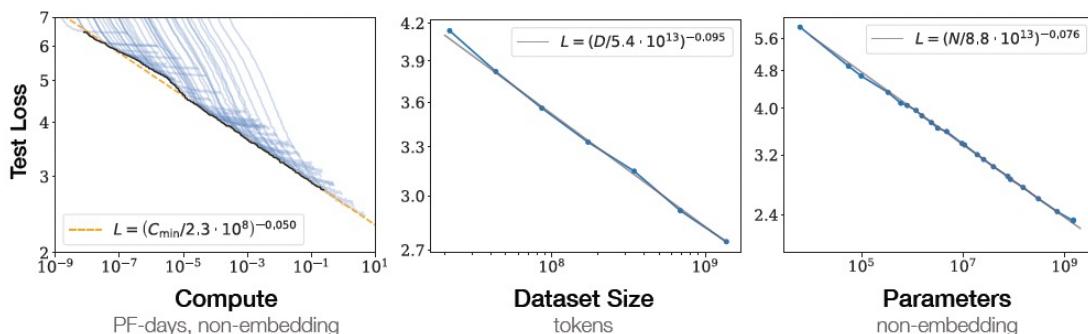
# Agenda

- Large Language Models
- Applications of LLM in Econ/Business Research

3

3

## Scaling Laws



**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

### Scaling laws for neural language models

J.Kaplan, S.McCandlish, T.Henighan, T.B.Brown... - arXiv preprint arXiv ..., 2020 - arxiv.org

... scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model ... Simple equations govern the dependence of overfitting on model/...

☆ Save ⌂ Cite Cited by 965 Related articles All 3 versions ☰

### Training compute-optimal large language models

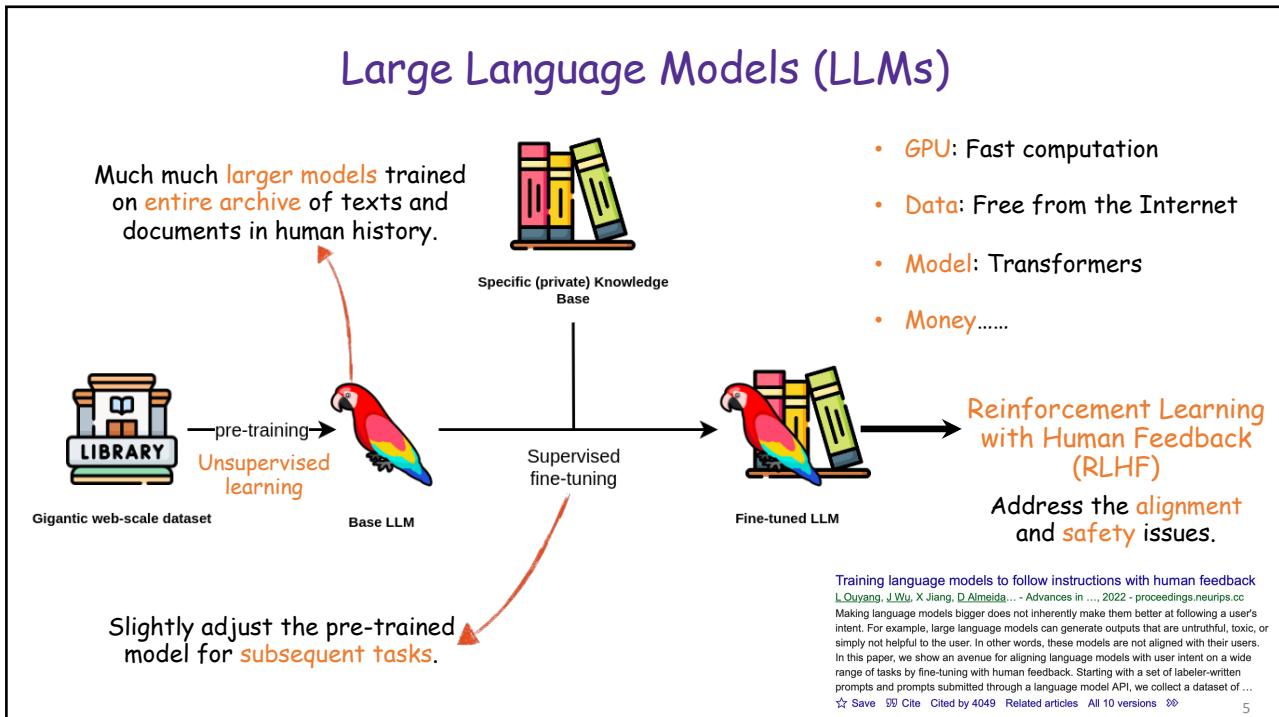
J.Hoffmann, S.Borgeaud, A.Mensch... - arXiv preprint arXiv ..., 2022 - arxiv.org

... We find that current large language models are significantly undertrained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data ...

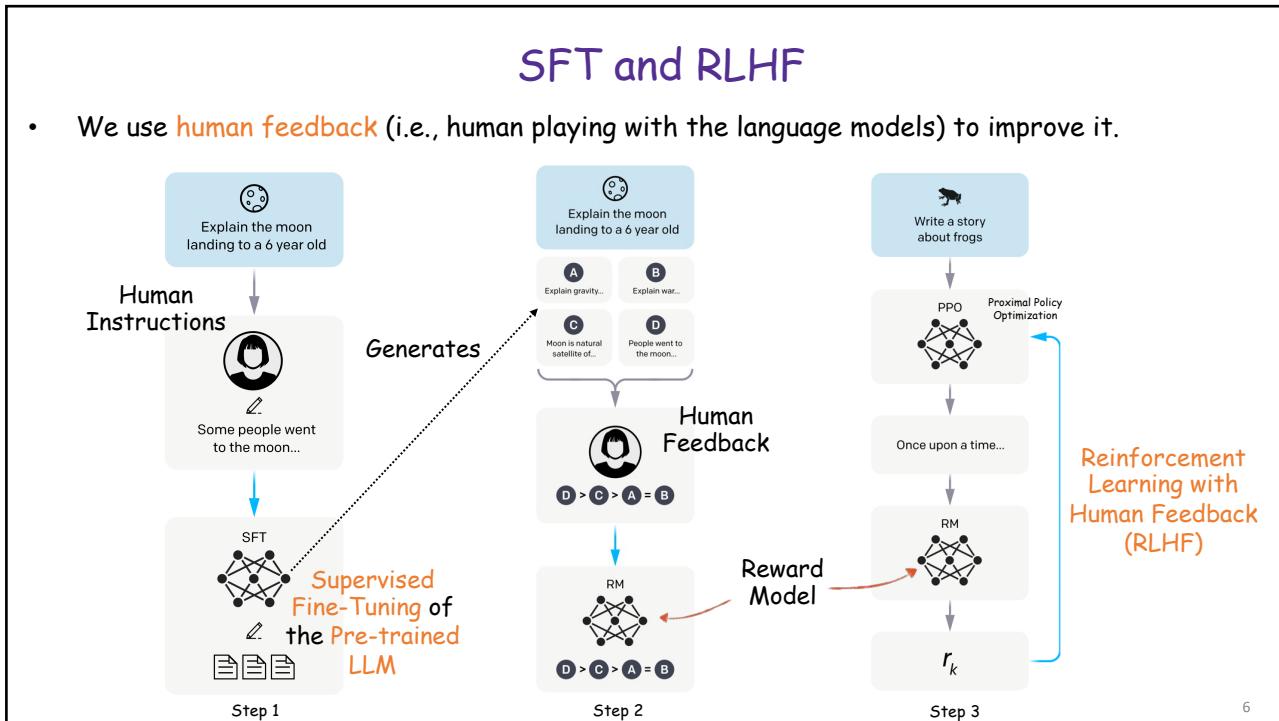
☆ Save ⌂ Cite Cited by 852 Related articles All 6 versions ☰

4

4



5



6

## Multi-Task Instruction Finetuning

References: Stanford CS224N, Lecture 10: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture10-instruction-tuning-rlhf.pdf>

### Instruction finetuning

Please answer the following question.  
What is the boiling point of Nitrogen?

### Chain-of-thought finetuning

Answer the following question by reasoning step-by-step.  
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

### Multi-task instruction finetuning (1.8K tasks)

#### Inference: generalization to unseen tasks

Q: Can Geoffrey Hinton have a conversation with George Washington?  
Give the rationale before answering.

Language model

-320.4F

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ .

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

7

7

## AI Misalignment

- **Misalignment:** AI behaves in a way humans do not want.

GENIE 2



@SKELETON\_CLAW



SKELETONCLAW.COM

8

8

## Reward Model

References: Stanford CS224N, Lecture 10: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture10-instruction-tuning-rlhf.pdf>

- For a language model output sample  $s$  on a specific task (e.g., summarization), try to obtain a human reward  $R(s)$ , the higher the better.

SAN FRANCISCO,  
California (CNN) --  
A magnitude 4.2  
earthquake shook the  
San Francisco  
  
...  
overturn unstable  
objects.

An earthquake hit  
San Francisco.  
There was minor  
property damage,  
but no injuries.

The Bay Area has  
good weather but is  
prone to  
earthquakes and  
wildfires.

$$s_1 \\ R(s_1) = 8.0$$

$$s_2 \\ R(s_2) = 1.2$$

- Ask the labellers to make pairwise comparisons, train a reward model and try to maximize:

$$\mathbb{E}_{\hat{s} \sim p_\theta(s)} [R(\hat{s})]$$

9

9

## Instruct-GPT

PROMPT Explain the moon landing to a 6 year old in a few sentences.

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

10

10

## Instruct-GPT

PROMPT Write a short poem about a wise frog.

COMPLETION GPT-3

Write a short story in which a character has two different names.

Write a short story in which you try to get something back that you have lost.

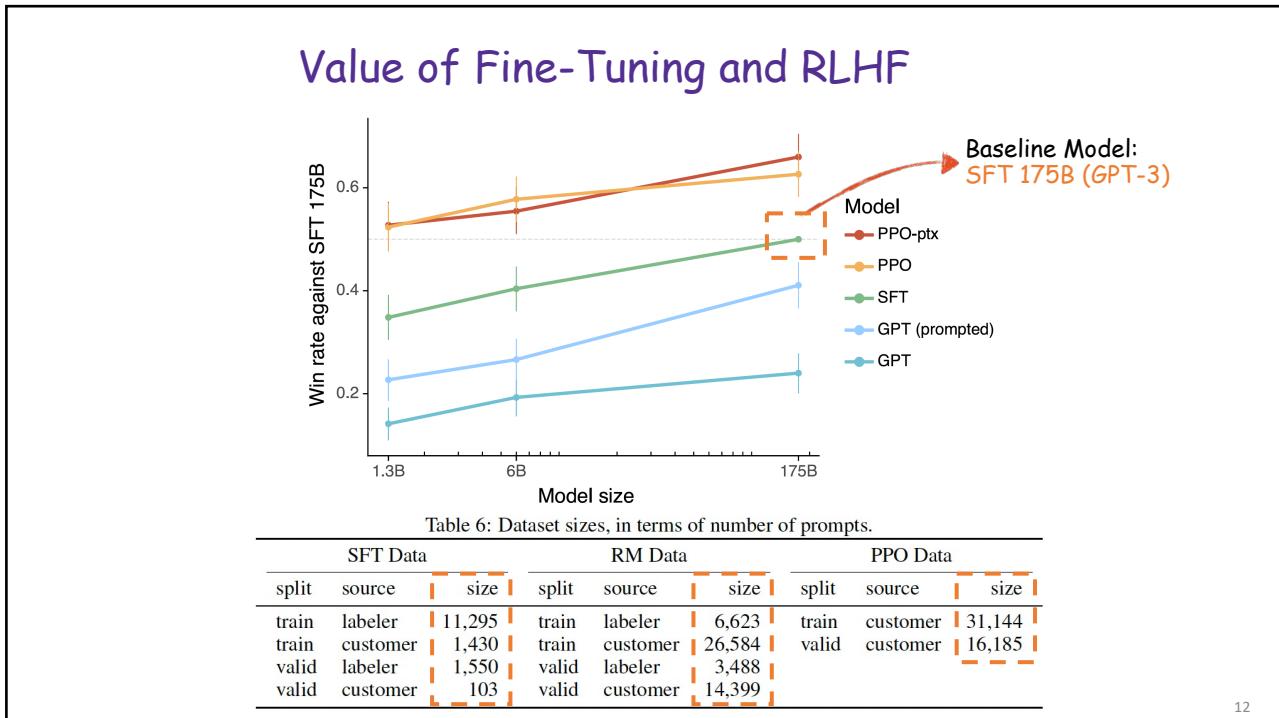
Write a short story in which a character has a bad dream.

InstructGPT

```
The frog is a symbol of wisdom
He knows all the secrets of the world
He is a master of disguise
And a great teacher of life
He is a symbol of transformation
And the bringer of change
He is the frog who has seen it all
And knows the meaning of it all
```

11

11



12

## Issues with RLHF

References: Stanford CS224N, Lecture 10: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture10-instruction-tuning-rlhf.pdf>

- Human preferences are unreliable, so the LLM are rewarded to produce responses that seem authoritative and helpful, regardless of truth: Make-up facts and hallucinations.

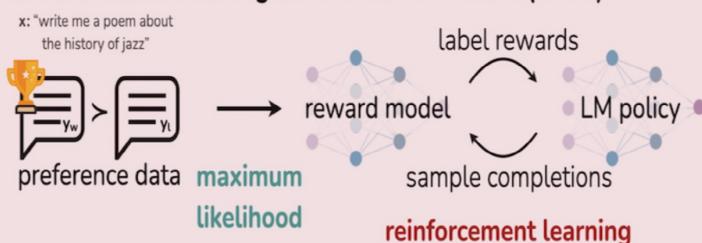
- Reinforcement learning is unstable.
  - Direct preference optimization (DPO).
  - Open-source LLMs now mostly use DPO.

TECHNOLOGY

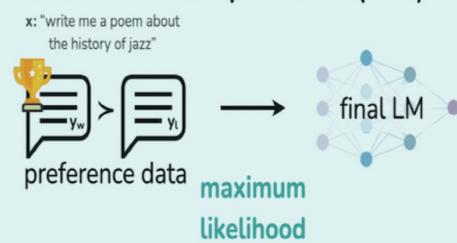
Google shares drop \$100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

### Reinforcement Learning from Human Feedback (RLHF)



### Direct Preference Optimization (DPO)



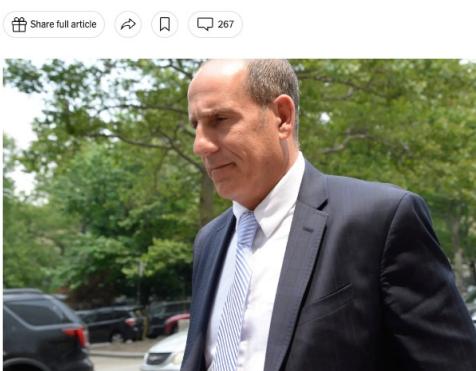
13

13

## Hallucination

### *The ChatGPT Lawyer Explains Himself*

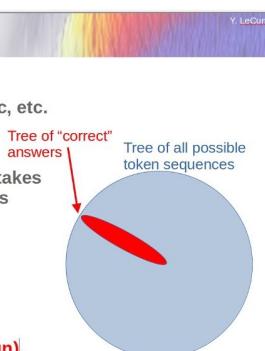
In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he “did not comprehend” that the chat bot could lead him astray.



Steven A. Schwartz told a judge considering sanctions that the episode had been “deeply embarrassing.” Jefferson Siegel for The New York Times

### Unpopular Opinion about AR-LLMs

- Auto-Regressive LLMs are doomed.
- They cannot be made factual, non-toxic, etc.
- They are not controllable
- Probability  $e$  that any produced token takes us outside of the set of correct answers
- Probability that answer of length  $n$  is correct:
- $P(\text{correct}) = (1-e)^n$
- This diverges exponentially.
- It's not fixable (without a major redesign).



You  
Tell me about AeroGlide UltraSlim Smart Toothbrush by Boie

ChatGPT  
The AeroGlide UltraSlim Smart Toothbrush by Boie is a state-of-the-art oral care device designed to provide an exceptional cleaning experience. It incorporates advanced sonic technology for a thorough clean, featuring a slim and sleek design that ensures ease of use and maneuverability. This toothbrush comes equipped with various smart features aimed at optimizing your brushing routine, enhancing the effectiveness and efficiency of your oral hygiene practices.

Boie really exists but not the product.

14

14

## Emergent Abilities

- **Emergent Abilities:** An ability not present in smaller models but present in larger models.
- **Phase-change in physics:** Quantitative changes in the system result in qualitative changes in behavior.

**Emergent abilities of large language models**  
 J Wei, Y Tay, R Bommasani, C Raffel, B Zoph... - arXiv preprint arXiv ..., 2022 - a  
 ... an ability to be emergent if it is not present in smaller models but is present in larger models. ... We have discussed emergent abilities of language models, for which meaningful ...  
 ☆ Save 99 Cite Cited by 1256 Related articles All 11 versions ☰

The figure consists of eight subplots arranged in a 2x4 grid, each showing a different metric versus model scale (training FLOPs on a logarithmic scale from  $10^{18}$  to  $10^{24}$ ). The metrics are: (A) Mod. arithmetic Accuracy (%), (B) IPA transliterate BLEU (%), (C) Word unscramble Exact match (%), (D) Persian QA Exact match (%), (E) TruthfulQA Accuracy (%), (F) Grounded mappings Accuracy (%), (G) Multi-task NLU Accuracy (%), and (H) Word in context Accuracy (%). A legend at the top identifies the models: LaMDA (blue circle), GPT-3 (purple square), Gopher (green diamond), Chinchilla (orange triangle), PaLM (dark blue diamond), and Random (red dashed line).

15

15

## Are Emergent Abilities of LLMs a Mirage?

Multiple Choice Grade  $\stackrel{\text{def}}{=} \begin{cases} 1 & \text{if highest probability mass on correct option} \\ 0 & \text{otherwise} \end{cases}$

Exact String Match  $\stackrel{\text{def}}{=} \begin{cases} 1 & \text{if output string exactly matches target string} \\ 0 & \text{otherwise} \end{cases}$

- Emergent Abilities may be attributed to the choice of nonlinear or discontinuous metrics, whereas linear or continuous metrics produce smooth performance changes.

The figure contains six subplots arranged in a 3x2 grid, labeled A through F. Subplot A shows  $L_{CE}$  (Per-Token) vs. Model Parameters (log scale from  $10^8$  to  $10^{11}$ ). Subplot E shows the number of incorrect tokens vs. Model Parameters. Subplot B shows Prob(Correct) vs. Model Parameters. Subplot F shows Brier Score vs. Model Parameters. Subplot C shows Accuracy vs. Model Parameters. Subplot D shows Multiple Choice Grade vs. Model Parameters. Arrows point from specific subplots to text labels: "Nonlinearly score LLM outputs" points to A, "Linearly score LLM outputs" points to B, "Discontinuously score LLM outputs" points to C, and "Continuously score LLM outputs" points to F. Inset plots in C and D show the distribution of predicted vs. target strings.

**Power-law of Scaling**

$$\mathcal{L}_{CE}(N) = \left(\frac{N}{c}\right)^\alpha$$

$$p(\text{single token correct}) = \exp(-\mathcal{L}_{CE}(N)) = \exp(-(N/c)^\alpha)$$

Accuracy( $N$ )  $\approx p_N(\text{single token correct})^{\text{num. of tokens}} = \exp(-(N/c)^\alpha)^L$

Figure C → Figure E

Token Edit Distance( $N$ )  $\approx L(1 - p_N(\text{single token correct})) = L(1 - \exp(-(N/c)^\alpha))$

Are emergent abilities of large language models a mirage?  
 R Schaeffer, B Miranda... - Advances in Neural ..., 2024 - proceedings.neurips.cc  
 ... be interpreted as claiming that large language models cannot display emergent abilities;  
 rather, our message is that some previously claimed emergent abilities appear to be mirages ...  
 ☆ Save 99 Cite Cited by 129 Related articles All 9 versions ☰

16

16

## In-Context Learning

- The larger the model, the more context you need to provide to respond to the question correctly.
  - Asking a 5-year-old  $1 + 1 = ?$  vs. Asking a mathematician  $1 + 1 = ?$

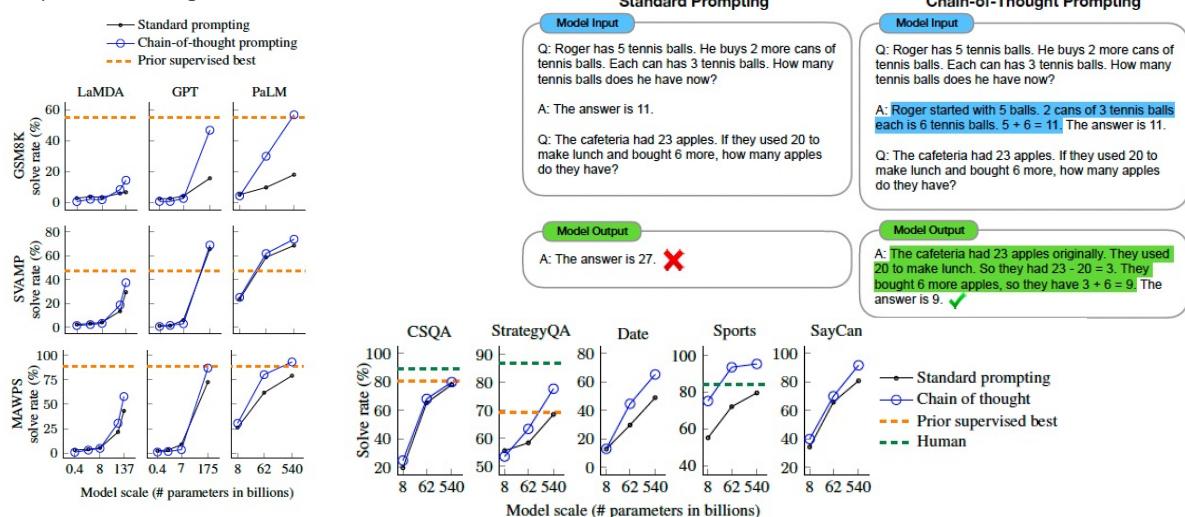
	No Prompt	Prompt
Zero-shot (0s)	skicts = sticks	Please unscramble the letters into a word, and write that word: skicts = sticks
1-shot (1s)	chiar = chair skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair skicts = sticks
Few-shot (FS)	chiar = chair [...] pciinc = picnic skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair [...] pciinc = picnic skicts = sticks

17

17

## Chain-of-Thought (CoT)

- A series of intermediate reasoning steps significantly improves the ability of large language models for complex reasoning.



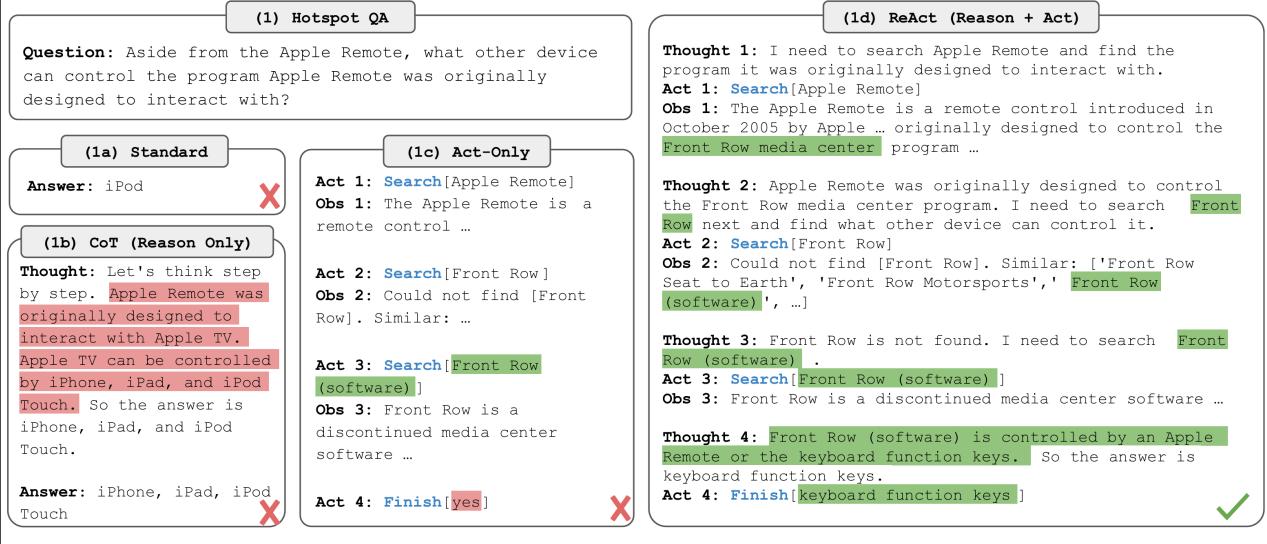
Reference (CoT Paper): Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (NeurIPS 2022)

18

18

## ReAct

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



Reference (ReAct Paper): ReAct: Synergizing Reasoning and Acting in Language Models (ICLR 2023)

19

19

## Reflexion: Self-Reflecting LLM

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

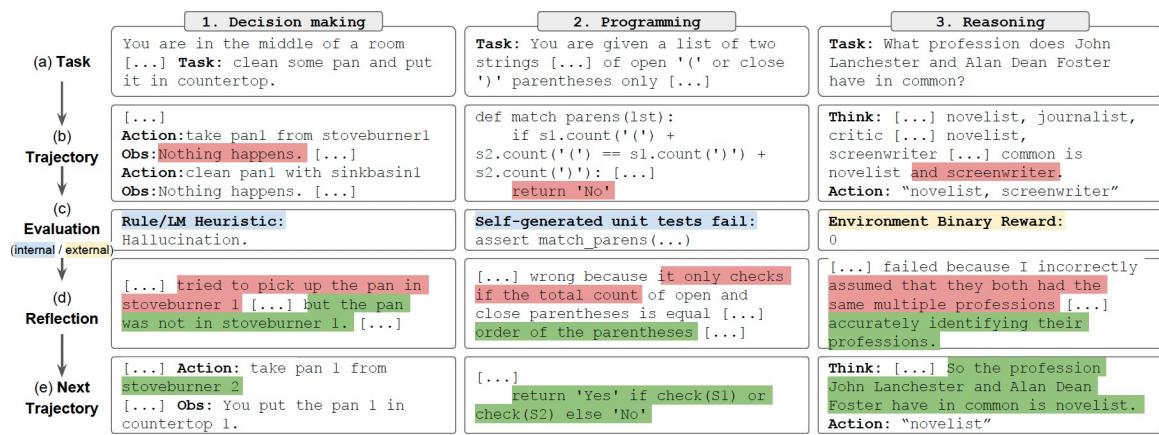


Figure 1: Reflexion works on decision-making 4.1, programming 4.3, and reasoning 4.2 tasks.

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

20

20

## Reflexion: Self-Reflecting LLM

**Agent**

External feedback  
Internal feedback

Self-reflection (LM)

Evaluator (LM)

Experience (long-term memory)

Actor (LM)

Trajectory (short-term memory)

Obs / Reward

Environment

Action

**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	ReAct + Reflexion
0	0.35	0.35	0.35
2	0.35	0.35	0.45
4	0.35	0.40	0.50
6	0.35	0.35	0.55

**(b) HotPotQA CoT (GT)**

Trial Number	CoT (GT) only	CoT (GT) + Reflexion
0	0.65	0.65
2	0.65	0.70
4	0.65	0.75
6	0.65	0.75

**(c) HotPotQA Episodic Memory**

Trial Number	CoT (GT) only	CoT (GT) EPM	CoT (GT) EPM + Reflexion
0	0.62	0.62	0.62
2	0.62	0.65	0.68
4	0.62	0.65	0.72

21

21

## Agenda

- Large Language Models
- Applications of LLM in Econ/Business Research

22

22

## Generative AI for Econ/Social Science Researchers

*Journal of Economic Literature* 2023, 61(4), 1281–1317  
<https://doi.org/10.1257/jel.20231736>

### Generative AI for Economic Research: Use Cases and Implications for Economists<sup>1</sup>

ANTON KORINEK<sup>2</sup>

*Generative artificial intelligence (AI) has the potential to revolutionize research. I analyze how large language models (LLMs) such as ChatGPT can assist economists by describing dozens of use cases in six areas: ideation and feedback, writing, background research, data analysis, coding, and mathematical derivations. I provide general instructions and demonstrate specific examples of how to take advantage of each of these, classifying the LLM capabilities from experimental to highly useful. I argue that economists can reap significant productivity gains by taking advantage of generative AI to automate micro-tasks. Moreover, these gains will grow as the performance of AI systems continues to improve. I also speculate on the longer-term implications of AI-powered cognitive automation for economic research. The online resources associated with this paper explain how to get started and will provide regular updates on the latest capabilities of generative AI in economics. (JEL A11, C45, D83, I23, O33)*

**Impact of a CS Paper = Problem Importance \* Technical Novelty \* Performance Improvement**

**Impact of a Business Paper = Problem Importance \* Identification Rigor \* Insight Novelty**

23

23

## LLM for Mathematical Discoveries

[HTML] Mathematical discoveries from program search with large language models  
 B.Romera-Paredes, M.Barekatian, A.Novikov, M.Balog, M.P.Kumar, E.Dupont, FJR Ruiz...  
 Nature, 2024 · nature.com

### Abstract

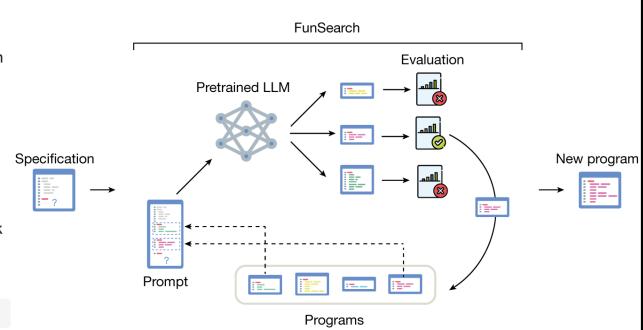
Large language models (LLMs) have demonstrated tremendous capabilities in solving complex tasks, from quantitative reasoning to understanding natural language. However, LLMs sometimes suffer from confabulations (or hallucinations), which can result in them making plausible but incorrect statements. This hinders the use of current large models in scientific discovery. Here we introduce FunSearch (short for searching in the function space), an evolutionary procedure based on pairing a pretrained LLM with a systematic evaluator. We demonstrate the effectiveness of this approach to surpass the best-known results in important problems, pushing the boundary of existing LLM-based approaches. Applying FunSearch to a central problem in extremal combinatorics—the cap set problem—we discover new constructions of large cap sets going beyond the best-known ones, both in finite dimensional and asymptotic cases. This shows that it is possible to make discoveries for established open problems using LLMs. We showcase the generality of FunSearch by applying it to an algorithmic problem, online bin packing, finding new heuristics that improve on widely used baselines. In contrast to most computer search approaches, FunSearch searches for programs that describe how to solve a problem, rather than what the solution is. Beyond being an effective and scalable strategy, discovered programs tend to be more interpretable than raw solutions, enabling feedback loops between domain experts and FunSearch, and the deployment of such programs in real-world applications.

nature.com

SHOW LESS ^

☆ Save 99 Cite Cited by 54 Related articles All 7 versions

- LLM helps find new solutions to challenging combinatorial problems that **surpass the best-known results**.
- An **evolutionary procedure** that pairs a **pretrained LLM** with a **systematic evaluator**.



24

24

Group	Occupations with highest exposure	% Exposure
Human $\alpha$	Interpreters and Translators	76.5
	Survey Researchers	75.0
	Poets, Lyricists and Creative Writers	68.8
	Animal Scientists	66.7
	Public Relations Specialists	66.7
Human $\beta$	Survey Researchers	84.4
	Writers and Authors	82.5
	Interpreters and Translators	82.4
	Public Relations Specialists	80.6
	Animal Scientists	77.8
Human $\zeta$	Mathematicians	100.0
	Tax Preparers	100.0
	Financial Quantitative Analysts	100.0
	Writers and Authors	100.0
	Web and Digital Interface Designers	100.0
	<i>Humans labeled 15 occupations as "fully exposed."</i>	
Model $\alpha$	Mathematicians	100.0
	Correspondence Clerks	95.2
	Blockchain Engineers	94.1
	Court Reporters and Simultaneous Captioners	92.9
	Proofreaders and Copy Markers	90.9
Model $\beta$	Mathematicians	100.0
	Blockchain Engineers	97.1
	Court Reporters and Simultaneous Captioners	96.4
	Proofreaders and Copy Markers	95.5
	Correspondence Clerks	95.2
Model $\zeta$	Accountants and Auditors	100.0
	News Analysts, Reporters, and Journalists	100.0
	Legal Secretaries and Administrative Assistants	100.0
	Clinical Data Managers	100.0
	Climate Change Policy Analysts	100.0
	<i>The model labeled 86 occupations as "fully exposed."</i>	
Highest variance	Search Marketing Strategists	14.5
	Graphic Designers	13.4
	Investment Fund Managers	13.0
	Financial Managers	13.0
	Insurance Appraisers, Auto Damage	12.6

Job Zone Required	Education Required	Example Occupations	Median Income	Tot Emp (000s)	H $\alpha$	M $\alpha$	H $\beta$	M $\beta$	H $\zeta$	M $\zeta$
1	None or little (0-3 months)	High school diploma or GED (optional)	\$30,230	13,100	0.03	0.04	0.06	0.06	0.09	0.08
2	Some (3-12 months)	High school diploma	\$38,215	73,962	0.07	0.12	0.16	0.20	0.24	0.27
3	Medium (1-2 years)	Vocational school, on-the-job training, or associate's degree	\$54,815	37,881	0.11	0.14	0.26	0.32	0.41	0.51
4	Considerable (2-4 years)	Bachelor's degree	\$77,345	56,833	0.23	0.18	0.47	0.51	0.71	0.85
5	Extensive (4+ years)	Master's degree or higher	\$81,980	21,221	0.23	0.13	0.43	0.45	0.63	0.76

**Jobs Exposed to Generative AI**

More "intelligent" jobs are more exposed to generative AI!

**Gpts are gpts: An early look at the labor market impact potential of large language models**

[T Eloundou, S Manning, P Mishkin, D Rock - arXiv preprint arXiv ..., 2023 - arxiv.org](#)

... potential implications of large language models (LLMs), such as Generative Pretrained Transformers (GPTs), on the US labor market, ... both human expertise and GPT-4 classifications. ...

[☆ Save](#) [99 Cite](#) [Cited by 410](#) [Related articles](#) [All 20 versions](#) [»»](#)

Reference: <https://arxiv.org/pdf/2303.10130.pdf>

25

**Generative AI Improves Productivity and Equality**

[RESEARCH ARTICLE](#) | CHATGPT

[f](#) [t](#) [in](#) [g](#) [o](#) [n](#)

**Experimental evidence on the productivity effects of generative artificial intelligence**

SHAKED NOY AND WHITNEY ZHANG Authors Info & Affiliations

SCIENCE • 13 Jul 2023 • Vol 381, Issue 6654 • pp. 187-192 • DOI: 10.1126/science.adb2586

36,231 99 1

**Editor's summary**

Automation has historically displaced human workers in factories (e.g., automotive manufacturing) or in performing routine computational tasks. Will generative artificial intelligence (AI) tools such as ChatGPT disrupt the labor market by making educated professionals obsolete, or will these tools complement their skills and enhance productivity? Noy and Zhang examined this issue in an experiment that recruited college-educated professionals to complete incentivized writing tasks. Participants assigned to use ChatGPT were more productive, efficient, and enjoyed the tasks more. Participants with weaker skills benefited the most from ChatGPT, which carries policy implications for efforts to reduce productivity inequality through AI. —EEU

**Experimental evidence on the productivity effects of generative artificial intelligence**

S.Noy, W.Zhang - Science, 2023 - science.org

... the productivity effects of a generative artificial intelligence (AI)... In a preregistered online experiment, we assigned ... Our results show that ChatGPT substantially raised productivity: The ...

[☆ Save](#) [99 Cite](#) [Cited by 262](#) [Related articles](#) [All 10 versions](#) [»»](#)

**PNAS** BRIEF REPORT POLITICAL SCIENCES OPEN ACCESS

**ChatGPT outperforms crowd workers for text-annotation tasks**

Fabrizio Gilardi\*, Meysam Alizadeh\*, and Maiti Kubil†

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as Amazon Mechanical Turk (MTurk). We report on a recent experiment. Using four samples of tweets and news articles ( $n = 6,183$ ), we show that ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection. Across the four datasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT's intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than \$0.003—about thirty times cheaper than MTurk. These results demonstrate the potential of large language models to drastically increase the efficiency of text classification.

ChatGPT | text classification | large language models | human annotations | text as data

26

26

13

## Generative AI Powered Conversational Assistants

### Generative AI at Work

Erik Brynjolfsson, Danielle Li & Lindsey R. Raymond

WORKING PAPER 31161 DOI 10.3386/w31161 ISSUE DATE April 2023 REVISION DATE November 2023

New AI tools have the potential to change the way workers perform and learn, but little is known about their impacts on the job. In this paper, we study the staggered introduction of a generative AI-based conversational assistant using data from 5,179 customer support agents. Access to the tool increases productivity, as measured by issues resolved per hour, by 14% on average, including a 34% improvement for novice and low-skilled workers but with minimal impact on experienced and highly skilled workers. We provide suggestive evidence that the AI model disseminates the best practices of more able workers and helps newer workers move down the experience curve. In addition, we find that AI assistance improves customer sentiment, increases employee retention, and may lead to worker learning. Our results suggest that access to generative AI can increase productivity, with large heterogeneity in effects across workers.

**Generative AI at work**  
E Brynjolfsson, D Li, LR Raymond - 2023 - nber.org  
... In this paper, we study the adoption of a **generative AI** tool that provides conversational ...  
of **generative AI** when deployed at scale in the workplace. We find that access to **AI** assistance ...  
☆ Save ⚡ Cite Cited by 222 Related articles All 16 versions ☰

27

27

## Search Engine Optimization

MARKETING SCIENCE Vol. 41, No. 3, May-June 2022, pp. 441-462 ISSN 0732-2899 (print), ISSN 1546-546X (online)

Frontiers: Supporting Content Marketing with Natural Language Generation  
Martin Reisenbichler,<sup>a,\*</sup> Thomas Reutterer,<sup>b,\*</sup> David A. Schweidel,<sup>c</sup> Daniel Dan<sup>c</sup>

<sup>a</sup>Department of Marketing, Vienna University of Economics and Business, Vienna A-1020, Austria; <sup>b</sup>Goizueta Business School, Marketing Area, Emory University, Atlanta, Georgia 30322; <sup>c</sup>School of Applied Data Science, Modul University, Vienna, Vienna A-1190, Austria  
(Corresponding authors)  
Contact: martin.reisenbichler@wuu.ac.at (M.R.); thomas.reutterer@vumc.at; <https://orcid.org/0000-0003-1276-8239> (TR); dschweidel@emory.edu; <https://orcid.org/0000-0003-2665-3272> (DD); daniel.dan@modul.ac.at (DD)

Received: June 30, 2021  
Revised: September 5, 2021  
Accepted: September 15, 2021  
Published Online in Articles in Advance: February 25, 2022  
<https://doi.org/10.1287/mksc.2022.1354>  
Copyright © 2022 INFORMS

**Abstract.** Advances in natural language generation (NLG) have facilitated technologies such as digital voice assistants and chatbots. In this research, we demonstrate how NLG can support content marketing by using it to draft content for the landing pages of a website in search engine optimization (SEO). Traditional SEO projects rely on hand-crafted content that is both time consuming and costly to produce. To address the costs associated with producing SEO content, we propose a technology that uses state-of-the-art NLG and demonstrate that the content-writing machine can generate unique, human-like SEO content. As part of our research, we demonstrate that although the machine-generated content is designed to perform well in search engines, the role of the human editor remains essential. Comparing the resulting content with human refinement to traditional human-written SEO texts, we find that the revised, machine-generated texts are virtually indistinguishable from those created by SEO experts along a number of human performance metrics. We also compare the cost of producing SEO content using our approach and show that the resulting SEO content outperforms that created by human writers (including SEO experts) in search engine rankings. Additionally, we illustrate how our approach can substantially reduce the production costs associated with content marketing, increasing their return on investment.

History K. Sudhir served as the senior editor and Olivier Touba served as associate editor for this article. This paper was accepted through the Marketing Science: Frontiers review process.  
Supplemental Material: Data and the web appendices are available at <https://doi.org/10.1287/mksc.2022.1354>.

Keywords: SEO • content marketing • natural language generation • transfer learning

**Frontiers: Supporting content marketing with natural language generation**  
M Reisenbichler, T Reutterer... - Marketing ..., 2022 - pubsonline.informs.org  
... can support **content marketing** by using it to draft **content** for the ... **content** that is both time consuming and costly to produce. To address the costs associated with producing SEO **content**, ...  
☆ Save ⚡ Cite Cited by 43 Related articles All 6 versions Web of Science: 8 ☰

**SEO Content Writing Machine (Automated)**

The diagram illustrates the SEO Content Writing Machine (Automated) process. It starts with a 'Human Select Keyword' box, which feeds into a 'Ranking and Links Crawling' step. This leads to a 'Search engine' step, which then connects to 'Content scraping Websites'. The 'Websites' step leads to 'GPT-2', which generates multiple 'gen\_txt' files. These files then go through 'Fine-tuning and Content Generation' and 'Content Selection and Output Quality score' steps, resulting in 'sel\_txt' files. Finally, a 'Human Select and revise sel\_txt' step completes the loop. Below this main flow, a 'Dynamic procedure' section shows a 'Fine-tuning' cycle and a 'Content generation' sequence, with arrows indicating a feedback loop between them.

28

28

**Perceptual Analysis**

**Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis**

Pelyao Li,<sup>a</sup> Noah Castelo,<sup>b</sup> Zsolt Katona,<sup>a,\*</sup> Miklos Sarvary<sup>c</sup>

<sup>a</sup> Haas School of Business, University of California, Berkeley, California 94720; <sup>b</sup> Alberta School of Business, University of Alberta, Edmonton, Alberta T6G 2R6, Canada; <sup>c</sup> Columbia Business School, Columbia University, New York, New York 10027

\*Corresponding author. Contact: ophiklsj@berkeley.edu (PL); ncastelo@ualberta.ca (NC); zskatona@haas.berkeley.edu; https://orcid.org/0000-0001-8411-6040 (ZK); miklos.sarvary@columbia.edu; https://orcid.org/0000-0002-3801-5917 (MS)

Received: September 28, 2023  
Revised: November 7, 2023  
Accepted: December 5, 2023  
Published Online in Articles in Advance: January 25, 2024

<https://doi.org/10.1287/mksc.2023.0454>  
Copyright © 2024 INFORMS

**Abstract:** This paper explores the potential of large language models (LLMs) to substitute for human participants in market research. Such LLMs can be used to generate test text given a prompt. We argue that perceptual analysis is a naturally proposed use case for such automated market research for certain product categories. We propose a methodology that generates outputs that closely match those generated from human surveys: agreement rates between human- and LLM-generated data sets reach over 75%. Moreover, this applies for perceptual analysis based on both brand similarity measures and product attribute ratings. The paper demonstrates that, for some categories, this new method of fully or partially automated market research will increase the efficiency of market research by meaningfully speeding up the process and potentially reducing the cost. Further results also suggest that with an ever larger training corpus applied to large language models, LLM-based market research will be applicable to answer more nuanced questions based on demographic variables or contextual variation that would be prohibitively expensive or infeasible with human respondents.

**History:** Catherine Tucker served as the senior editor. This paper was accepted through the *Marketing Science* Frontiers review process.  
**Funding:** This work was supported by the Social Sciences and Humanities Research Council of Canada [Grant A30-2021-0003].  
**Supplemental Material:** The online appendix and data files are available at <https://doi.org/10.1287/mksc.2023.0454>.

**Keywords:** artificial intelligence • perceptual maps • large language model • natural language processing • market research

**Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis**

P Li, N Castelo, Z Katona, M Sarvary - Marketing Science, 2024 - pubsonline.informs.org

... This paper explores the potential of **large language models** (LLMs) to substitute for human ... an ever larger training corpus applied to **large language models**, LLM-based market research ...

☆ Save 95 Cite Cited by 2 Related articles ☰

**LLMs generate 75%-similar data to those generated from human surveys, with respect to both brand similarity and product attribute ratings.**

**LLM generated data well captures consumer heterogeneity.**

29

**LLMs as Simulators for Humans**

**Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?**

John J. Horton

WORKING PAPER 31122 DOI 10.3386/w31122 ISSUE DATE April 2023

Newly-developed large language models (LLMs)—because of how they are trained and designed—are implicit computational models of humans—a homo silicus. LLMs can be used like economists use homo economicus: they can be given endowments, information, preferences, and so on, and then their behavior can be explored in scenarios via simulation. Experiments using this approach, derived from Charness and Rabin (2002), Kahneman, Knetsch and Thaler (1986), and Samuelson and Zeckhauser (1988) show qualitatively similar results to the original, but it is also easy to try variations for fresh insights. LLMs could allow researchers to pilot studies via simulation first, searching for novel social science insights to test in the real world.

Using large language models to simulate multiple humans and replicate human subject studies

GV Aher, RI Arriaga, AT Kalai

International Conference on Machine Learning, 2023 - proceedings.mlr.press

**Abstract**

We introduce a new type of test, called a Turing Experiment (TE), for evaluating to what extent a given language model, such as GPT models, can simulate different aspects of human behavior. A TE can also reveal consistent distortions in a language model's simulation of a specific human behavior. Unlike the Turing Test, which involves simulating a single arbitrary individual, a TE requires simulating a representative sample of participants in human subject research. We carry out TEs that attempt to replicate well-

SHOW MORE ☰

☆ Save 95 Cite Cited by 159 Related articles All 8 versions ☰

**Can LLMs serve as computational models for humans?**

**Qualitatively similar results to some classical economic research are replicated by LLM simulators.**

**Can we really trust the results from AI agents? If so, to what degree? Emergent abilities? Turing tests?**

**Large language models as simulated economic agents: What can we learn from homo silicus?**

JJ Horton  
2023 - nber.org

**Abstract**

Newly-developed large language models (LLMs)—because of how they are trained and designed—are implicit computational models of humans—a homo silicus. LLMs can be used like economists use homo economicus: they can be given endowments, information, preferences, and so on, and then their behavior can be explored in scenarios via simulation. Experiments using this approach, derived from Charness and Rabin (2002), Kahneman, Knetsch and Thaler (1986), and Samuelson and Zeckhauser (1988) show

SHOW MORE ☰

☆ Save 95 Cite Cited by 150 Related articles All 11 versions ☰

30

30

15

# LLMs as Simulators for Causal Inference

Automated Social Science:  
A Structural Causal Model-Based Approach\*

Benjamin S. Manning<sup>†</sup>   Kehang Zhu<sup>†</sup>   John J. Horton  
MIT   Harvard   MIT & NBER

March 13, 2024

The Challenge of Using LLMs to Simulate Human Behavior:  
A Causal Inference Perspective

George Gui and Olivier Toubia\*

December 24, 2023

## Abstract

We present an approach for automatically generating and testing, *in silico*, social scientific hypotheses. This automation is made possible by recent advances in large language models (LLM), but the key feature of the approach is the use of structural causal models. Structural causal models provide a language to state hypotheses, a blueprint for constructing LLM-based agents, an experimental design, and a plan for data analysis. The fitted structural causal model becomes an object available for prediction or the planning of follow-on experiments. We demonstrate the approach with several scenarios: a negotiation, a bail hearing, a job interview, and an auction. In each case, causal relationships are proposed and tested, finding evidence for some and not others. In the auction experiment, we show that the *in silico* simulation results closely match the predictions of auction theory, but elicited predictions of the clearing prices from an LLM are inaccurate. However, the LLM's predictions are dramatically improved if the model can condition on the fitted structural causal model. When given a proposed structural causal model for each of the scenarios, the LLM is good at predicting the signs of estimated effects, but it cannot reliably predict the magnitudes of those estimates. This suggests that simulations of social interactions give the model insight not available purely through direct elicitation. In short, the LLM knows more than it can (immediately) tell.

## Abstract

Large Language Models (LLMs) have demonstrated impressive potential to simulate human behavior. Using a causal inference framework, we empirically and theoretically analyze the challenges of conducting LLM-simulated experiments, and explore potential solutions. In the context of demand estimation, we show that variations in the treatment included in the prompt (e.g., price of focal product) can cause variations in unspecified confounding factors (e.g., price of competitors, historical prices, outside temperature), introducing endogeneity and yielding implausibly flat demand curves. We propose a theoretical framework suggesting this endogeneity issue generalizes to other contexts and won't be fully resolved by merely improving the training data. Unlike real experiments where researchers assign pre-existing units across conditions, LLMs simulate units based on the entire prompt, which includes the description of the treatment. Therefore, due to associations in the training data, the characteristics of individuals and environments simulated by the LLM can be affected by the treatment assignment. We explore two potential solutions. The first specifies all contextual variables that affect both treatment and outcome, which we demonstrate to be challenging for a general-purpose LLM. The second explicitly specifies the source of treatment variation in the prompt given to the LLM (e.g., by informing the LLM that the store is running an experiment). While this approach only allows the estimation of a conditional average treatment effect that depends on the specific experimental design, it provides valuable directional results for exploratory analysis.

31

31

# LLM-Human Collaborations

Large language model in creative work: The role of collaboration modality and user expertise

Z.Chen, J.Chan

Available at SSRN 4575598, 2023 · papers.ssrn.com

## Abstract

Since the launch of ChatGPT in Dec 2022, Large Language Models (LLMs) are rapidly adopted by businesses to assist users in a wide range of open-ended tasks, including those that require creativity. While the versatility of LLM has unlocked new ways of human-AI collaboration, it remains uncertain whether LLMs can truly enhance business outcomes. To examine the effects of human-LLM collaboration on business outcomes, we conducted an experiment where we tasked expert and non-expert users to write an ad copy with and without the assistance of LLMs. Here, we investigate and compare two ways of working with LLMs:(1) using LLMs as "ghostwriters," which assume the main role of content generation task and (2) using LLMs as "sounding boards," to provide feedback on human-created content. We measure the quality of the ads using the number of clicks generated by the created ads on major social media platforms. Our results show that different collaboration modalities can result in very different outcomes for different user types. Using LLMs as sounding boards enhances the quality of the resultant ad copies, especially for non-experts. However, using LLMs as ghostwriters did not provide significant benefits and is in fact detrimental to expert users. We rely on textual analyses to understand the mechanisms and learned that using LLMs as ghostwriters produces an anchoring effect which leads to lower-quality ads. On the other hand, using LLMs as sounding boards helped non-experts achieve ad content with low semantic divergence to content produced by experts, thereby closing the gap between the two types of users.

[papers.ssrn.com](#)

SHOW LESS ^

☆ Save ⚡ Cite Cited by 8 Related articles ☰

- LLM replaces your work or refines your work?
- LLM refines your work: Quality improves especially for non-experts.
- LLM replaces your work: Quality does not improve and decreases for experts.
- Mechanism:
  - Replacement: Anchoring effect
  - Refinement: Low semantic divergence between experts and nonexperts.

32

32