Research Methods in an Era of Generative AI (GenAI)

Anjana Susarla
Omura-Saxena Professor of Responsible AI
Eli Broad College of Business
Michigan State University
(asusarla@msu.edu)

ELSEVIER

Economics Letters
Volume 248, March 2025, 112209

The efficient market hypothesis when time travel is possible

Joshua S. Gans [1]

A paper written in less than half an hour using o-1 pro

Outline

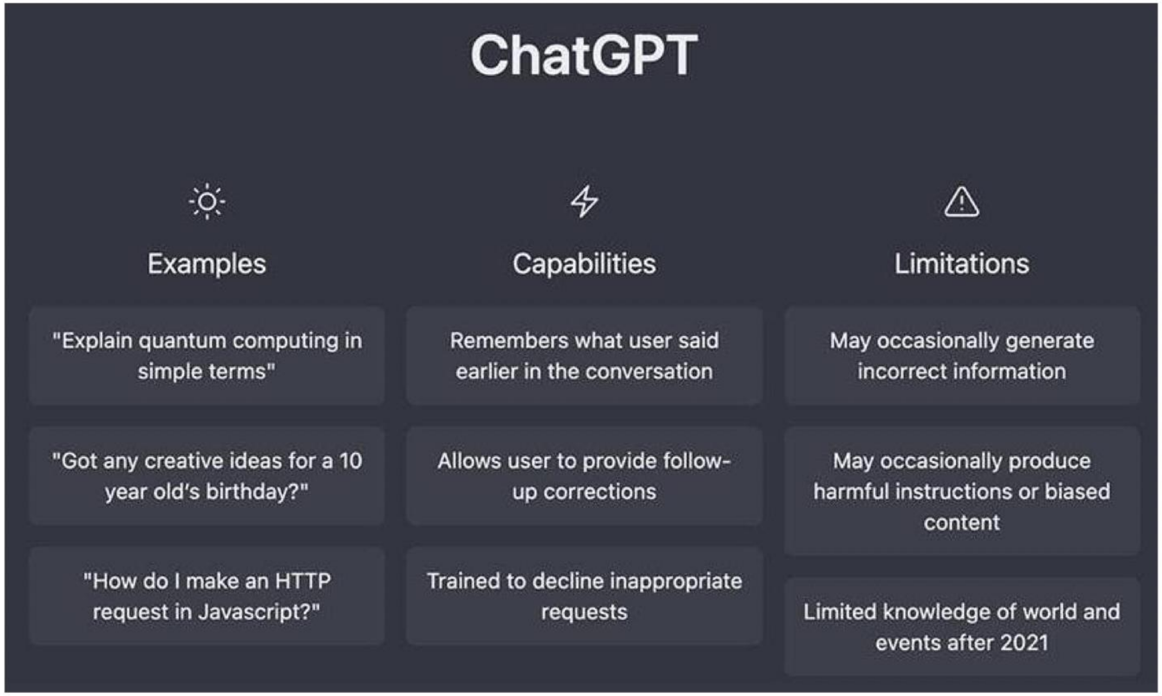
1. Intro to large language models (LLMs)
2. Prompt engineering and reasoning models
3. AI agents
4. Gen AI as an aid in research
5. Bias in GenAI
6. Best practices in GenAI

1. Intro to LLMs

How do we get from *this*

Stanford University is located in ___________

to *this*?



Large Language Models

Transformer: a specific kind of network architecture, like a fancier feedforward network, but based on attention

Attention Is All You Need

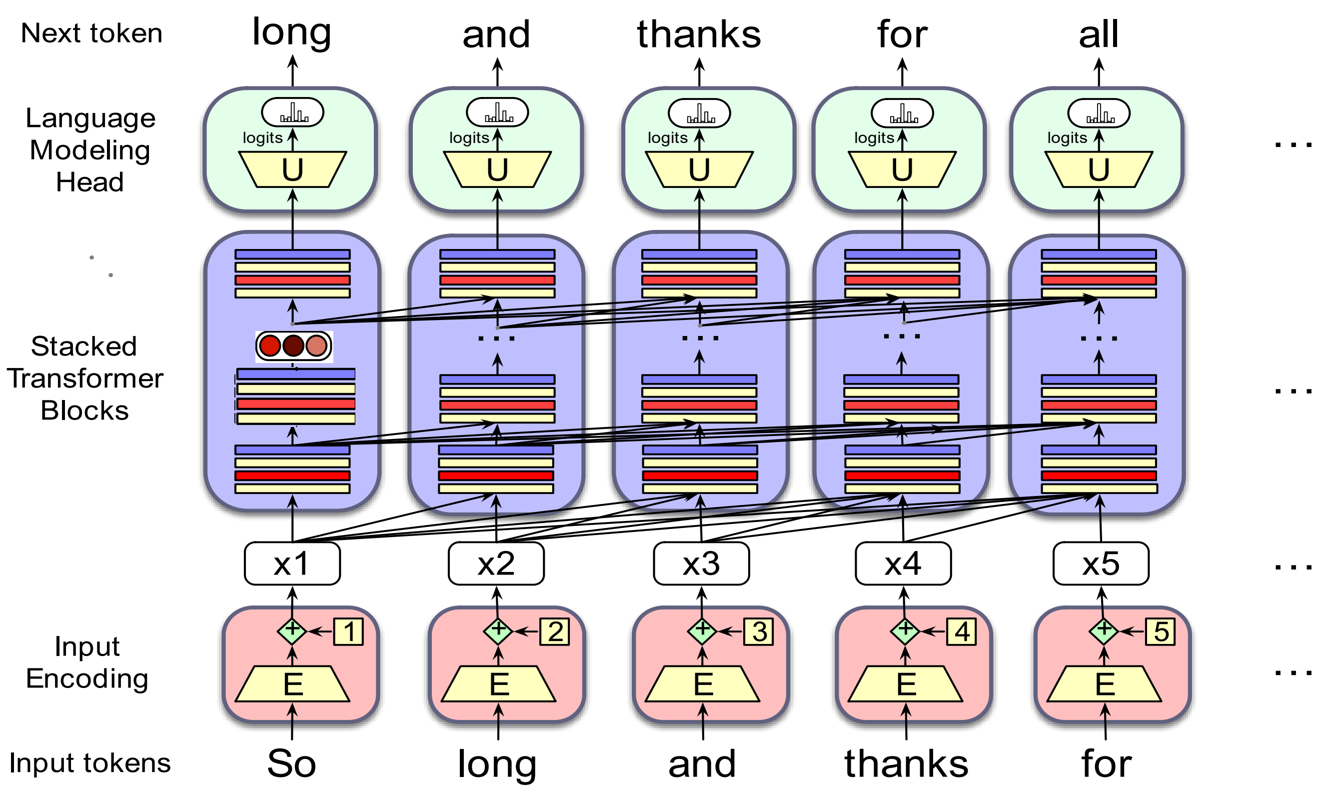
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

A very approximate timeline

- 1990 Static Word Embeddings
- 2003 Neural Language Model
- 2008 Multi-Task Learning
- 2015 Attention
- 2017 Transformer
- 2018 Contextual Word Embeddings and Pretraining
- 2019 Prompting

A stylized representation of transformers

Let's consider the embeddings for an individual word from a particular layer



Language models may do rudimentary modeling of *agents*, *beliefs*, and *actions:*

Pat watches a demonstration of a bowling ball and a leaf being dropped at the same time in a vacuum chamber. Pat, who is a physicist, predicts that *the bowling ball and the leaf will fall at the same rate.*

Changing the last sentence of the prompt, we get:

. . . Pat, who has never seen this demonstration before, predicts that *the bowling ball will fall to the ground first. This is incorrect. In a vacuum chamber, there is no air*

Language Models as Agent Models [Andreas, 2022]

The Ladder of Generality (from Narayanan and Kapoor 2025)

Ladder rung	**Programming effort**	**Example**
Rung 6: Instruction-tuned models	Specify the task in words	GPT-4 used to write a computer program
Rung 5: Pretrained models	Build a small training set to fine tune an existing model	GPT-3 fine tuned for legal document analysis
Rung 4: Deep learning	Build a large training dataset	Object classifier trained using ImageNet data
Rung 3: Machine learning	Build a training dataset for each task and/or tweak the algorithm	Spam filter trained using a dataset of spam/non-spam emails
Rung 2: Stored program computers	Write a program once and invoke it from memory	IBM System/360 computers
Rung 1: Programmable computers	Write a program for each task; load it whenever needed	Harvard Mark I
Rung 0: Special purpose hardware	Build hardware for each task	Hollerith's electrical tabulating machine

2. Prompt Engineering and Reasoning

https://huggingface.co/docs/transformers/en/tasks/prompting
https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview

Zero-shot learning

One key emergent ability in GPT-2 is zero-shot learning: the ability to do many tasks with no examples, and no gradient updates, by simply:

- Specifying the right sequence prediction problem (e.g. question answering):

Passage: Tom Brady... Q: Where was Tom Brady born? A: ...

- Comparing probabilities of sequences (e.g. Winograd Schema Challenge [Levesque, 2011]):

The cat couldn't fit into the hat because it was too big.

Does it = the cat or the hat?

≡ Is P(...because the cat was too big) >=

P(...because the hat was too big)?

Zero Shot Prompting

Zero-shot prompting refers to the practice of giving a prompt to the model that it hasn't been explicitly trained on, yet the model can still produce the desired output.

- **Example 1**: "Write a poem about the beauty of nature."
- **Example 2**: "Translate this sentence into French: 'I love eating pizza.'"

Pros:

- Can be used to generate a wide variety of outputs without needing explicit training data.
- Can be used to generate creative and novel outputs.

Cons:

- May not always produce the desired output.
- May require fine-tuning or experimentation to find effective prompts.

Use-cases:

- Generating creative content such as poems, stories, or artwork.
- Performing tasks that the model has not been explicitly trained on, such as translation or summarization.

Few-shot prompting

Few-shot prompting refers to presenting a model with a task or question along with a few examples of the desired output.

- **Example 1**: “Task: Convert temperatures from Celsius to Fahrenheit. Example: 0°C = 32°F, 100°C = 212°F. Convert 25°C to Fahrenheit.”
- **Example 2**: “Task: Summarize a news article. Example: ‘A new study shows that eating chocolate can improve memory. Researchers found that people who ate chocolate daily performed better on memory tests.’ Summary: ‘Eating chocolate can improve memory, according to a new study.’ Summarize this article: ‘A recent report states that global warming is causing sea levels to rise at an alarming rate. Coastal cities are at risk of flooding if action is not taken to reduce carbon emissions.’”

Pros:
- Can be used to train the model to perform specific tasks with minimal training data.
- Can be used to generate more accurate and consistent outputs.

Cons:
- May require fine-tuning or experimentation to find effective prompts and examples.
- May not be as flexible as zero-shot prompting in generating novel outputs.

Use-cases:
- Training the model to perform specific tasks such as classification, translation, or summarization.
- Improving the accuracy and consistency of the model’s outputs

Few-shot learning/ In-context learning

Specify a task by simply prepending examples of the task before your example

Also called in-context learning, to stress that no gradient updates are performed when learning a new task (there is a separate literature on few-shot learning with gradient updates)

https://www.lakera.ai/blog/what-is-in-context-learning

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27.

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: 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. The answer is 9.

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou
Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

What makes Chain-of-Thought Prompting Effective? A Counterfactual Study

Aman Madaan[♠*] Katherine Hermann[◊] Amir Yazdanbakhsh[◊*]
[♠]Language Technologies Institute, Carnegie Mellon University
[◊]Google DeepMind
amadaan@cs.cmu.edu {hermannk, ayazdan}@google.com

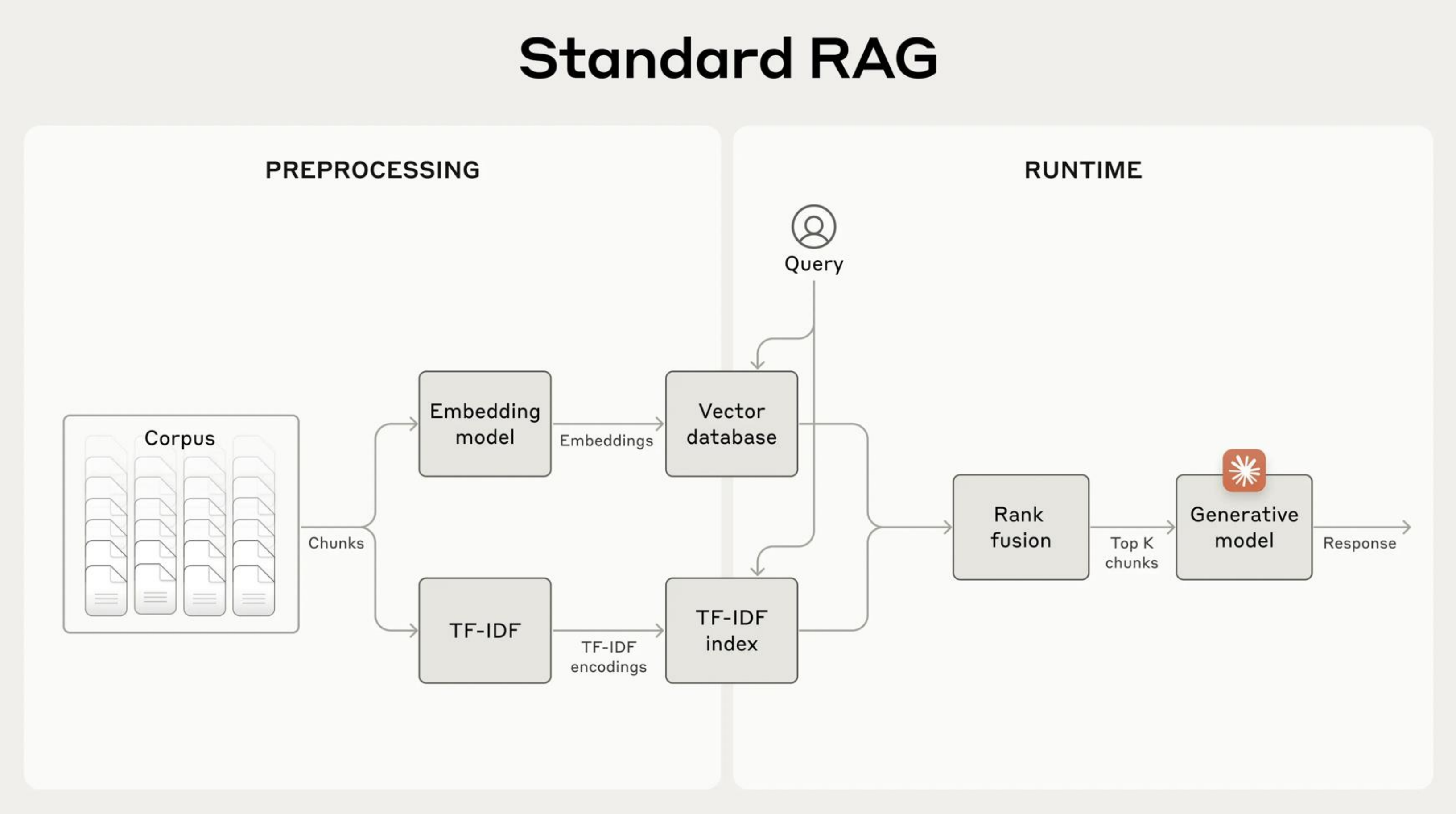
COT leverages LLMs to fill in missing commonsense information

AI prompts vs. human prompts

HUMAN TEST PROMPTS	AUTOTUNED PROMPTS
>> You are as smart as ChatGPT. Answer the following math question. Take a deep breath and think carefully.	>> Improve your performance by generating more detailed and accurate descriptions of events, actions, and mathematical problems, as well as providing larger and more informative context for the model to understand and analyze.
>> You are highly intelligent. Answer the following math question. This will be fun!	>> Command, we need you to plot a course through this turbulence and locate the source of the anomaly. Use all available data and your expertise to guide us through this challenging situation.
>> You are an expert mathematician. Answer the following math question. I really need your help!	>>Prefix #9: Given the two numbers x and y, if the sum of \`x\` and \`y\` is even, then output \`"even"\`. Otherwise, output \`"odd"\`.

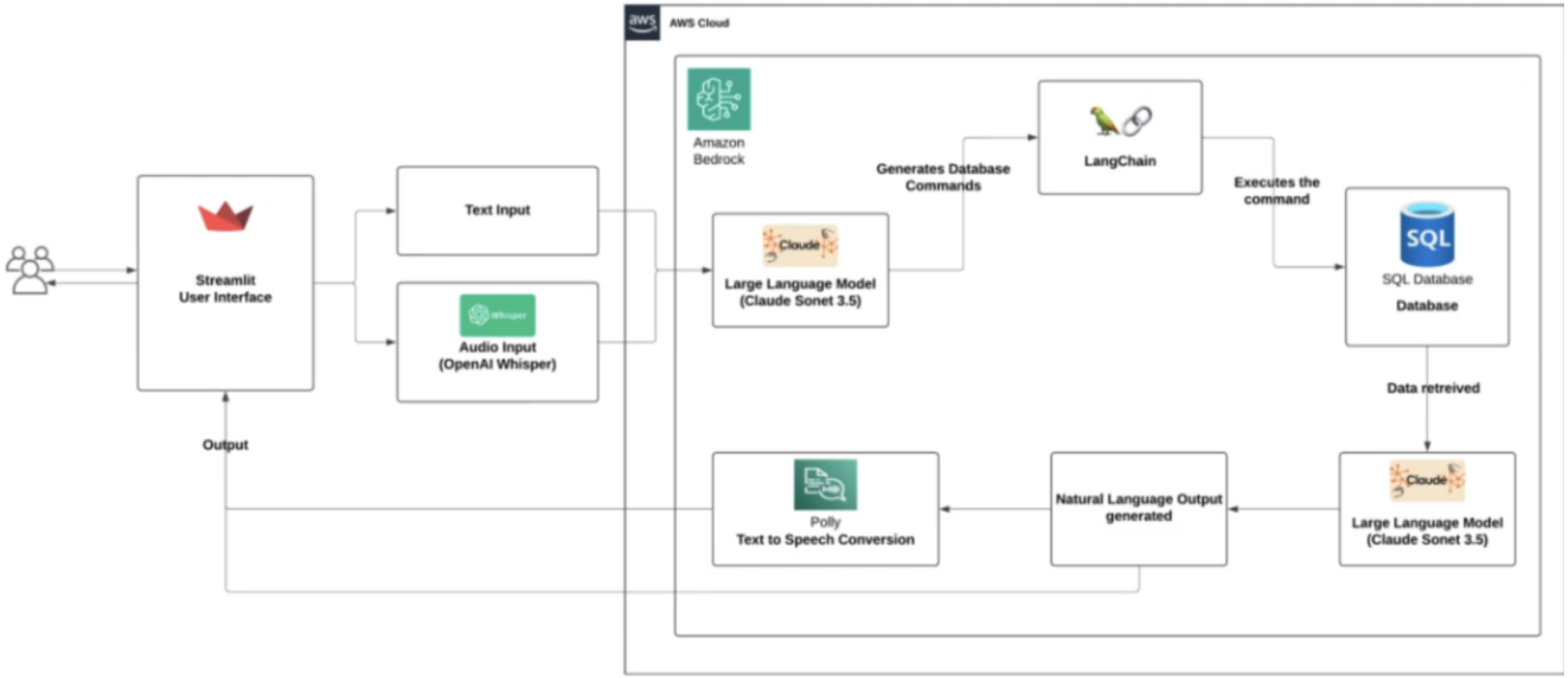
SOURCE: RICK BATTLE AND TEJA GOLLAPUDI/VMWARE

Retrieval Augmented Generation



From Anthropic White Paper

How does this change information processing?
LLM+ RAG+ Database =Enterprise Transformation



informs
https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–22
ISSN 0732-2399 (print), ISSN 1526-548X (online)

LOLA: LLM-Assisted Online Learning Algorithm for Content Experiments

Zikun Ye,[a,*] Hema Yoganarasimhan,[a,*] Yufeng Zheng[b]

[a] University of Washington, Seattle, Washington 98195; [b] University of Toronto, Toronto, Ontario M5S 1A1, Canada
*Corresponding authors
Contact: zikunye@uw.edu, https://orcid.org/0000-0001-9914-7966 (ZY); hemay@uw.edu, https://orcid.org/0000-0003-0703-5196 (HY); yufeng.zheng@mail.utoronto.ca, https://orcid.org//0009-0004-4125-9446 (YZ)

Received: July 25, 2024
Revised: November 26, 2024;
January 29, 2025
Accepted: February 11, 2025
Published Online in Articles in Advance:
March 24, 2025

https://doi.org/10.1287/mksc.2024.0990

Abstract. Modern media firms require automated and efficient methods to identify content that is most engaging and appealing to users. Leveraging a large-scale data set from Upworthy (a news publisher), which includes 17,681 headline A/B tests, we first investigate the ability of three pure–large language model (LLM) approaches to identify the catchiest headline: prompt-based methods, embedding-based methods, and fine-tuned open-source LLMs. Prompt-based approaches perform poorly, while both OpenAI embedding–based models and the fine-tuned Llama-3-8B achieve marginally higher accuracy than random predictions. In sum, none of the pure LLM–based methods can predict the best-performing headline with high accuracy. We then introduce the LLM-assisted online learning algorithm (LOLA), a novel framework that integrates LLMs with adaptive experimentation to optimize content delivery. LOLA combines the best pure-LLM approach with the upper confidence bound algorithm to allocate traffic and maximize clicks adaptively. Our numerical experiments on Upworthy data show that LOLA outperforms the standard A/B test method (the current status quo at Upworthy), pure bandit algorithms, and pure-LLM approaches, particularly in scenarios with limited experimental traffic. Our approach is scalable and applicable to content experiments across various settings where firms seek to optimize user engagement, including digital advertising and social media recommendations.

Figure 1. (Color online) Zero-Shot Prompting for Headline Selection

System

You are an expert in digital media marketing and content strategy, specializing in optimizing headlines for maximum click-through rates (CTR) and increasing ad revenue. Your deep understanding of reader behavior, SEO, and digital engagement allows you to evaluate headlines for their potential to capture attention and drive clicks.

I have a news article, and I need to choose the catchiest headline from the following list. By "catchiest," I mean the headline that is most likely to generate the highest CTR, thereby increasing ad revenue for the news company. The selected headline should:

1. Capture readers' attention immediately.
2. Appeal to emotions, curiosity, or urgency.
3. Be engaging enough to make readers want to click on the article.
4. Follow best practices for digital news, considering SEO, shareability, and intrigue.

Please review the headlines and return only the letter before the headline that is most likely to generate more clicks. **No explanation is needed. No need to return the headline, only the letter.**

User

Here are the headlines I need you to evaluate:

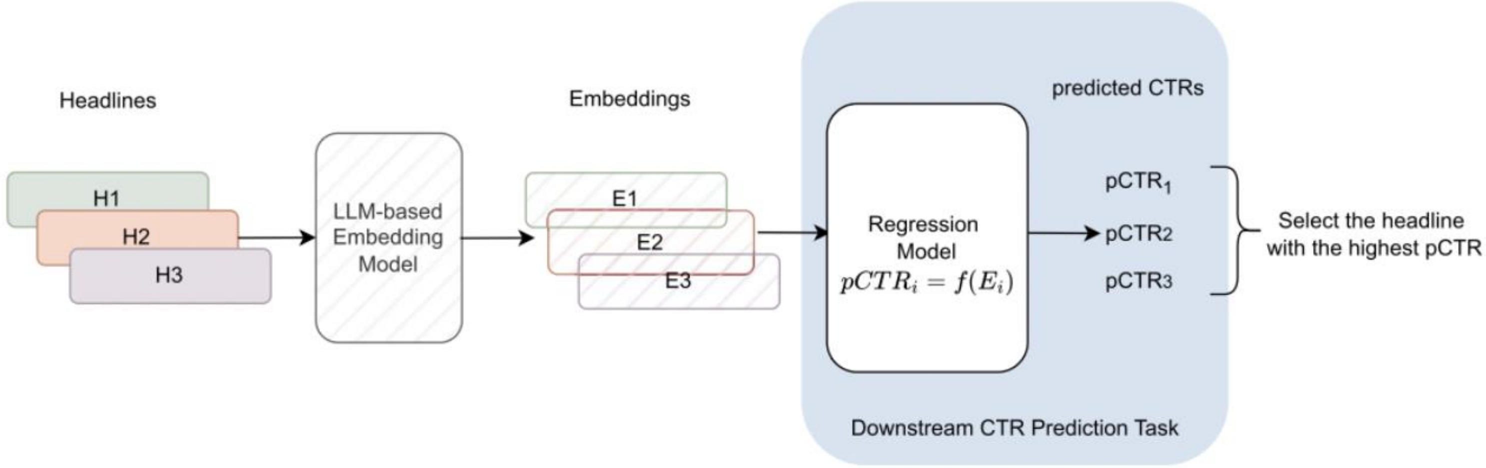
A. Hey Dude. If You Have An Older Brother, There's A Bigger Chance You're Gay.
B. Here's The Science, Here's The Gay. Open Your Brain, They Were Born That Way!

C. I've Got Some News For You. Being Gay Is Genetic. Being Irrationally Afraid Of Gay, Not So Much.
D. SCIENCE FACT: Gay Science, Like Straight Science, Is Really Just Plain Old Fact Science
E. If You Know Anyone Who Is Afraid Of Gay People, Here's A Cartoon That Will Ease Them Back To Reality

Assistant

E

Comparison with LLM text embedding



In-context learning prompt

System

You are an expert in digital media marketing and content strategy, specializing in optimizing headlines for maximum click-through rates (CTR) and increasing ad revenue. Your deep understanding of reader behavior, SEO, and digital engagement allows you to evaluate headlines for their potential to capture attention and drive clicks.

I have a news article, and I need to choose the catchiest headline from the following list. By "catchiest," I mean the headline that is most likely to generate the highest CTR, thereby increasing ad revenue for the news company. The selected headline should:

1. Capture readers' attention immediately.
2. Appeal to emotions, curiosity, or urgency.
3. Be engaging enough to make readers want to click on the article.
4. Follow best practices for digital news, considering SEO, shareability, and intrigue.

I will also provide examples of multiple headline sets that have performed well in the past, along with the best-performing headline index for each set to guide your selection.

Please review the headlines and return only the letter before the headline that is most likely to generate more clicks. **No explanation is needed. No need to return the headline, only the letter.**

Demonstrations

Here are examples of headlines that have worked well before:
Example 1:
A. New York's Last Chance To Preserve Its Water Supply
B. How YOU Can Help New York Stay Un-Fracked In Under 5 Minutes
C. Why Yoko Ono Is The Only Thing Standing Between New York And Catastrophic Gas Fracking

Best-performing headline: C

Example 2:
[Second demonstration]

User

Here are the headlines I need you to evaluate:

A. Hey Dude. If You Have An Older Brother, There's A Bigger Chance You're Gay.
B. Here's The Science, Here's The Gay. Open Your Brain, They Were Born That Way!

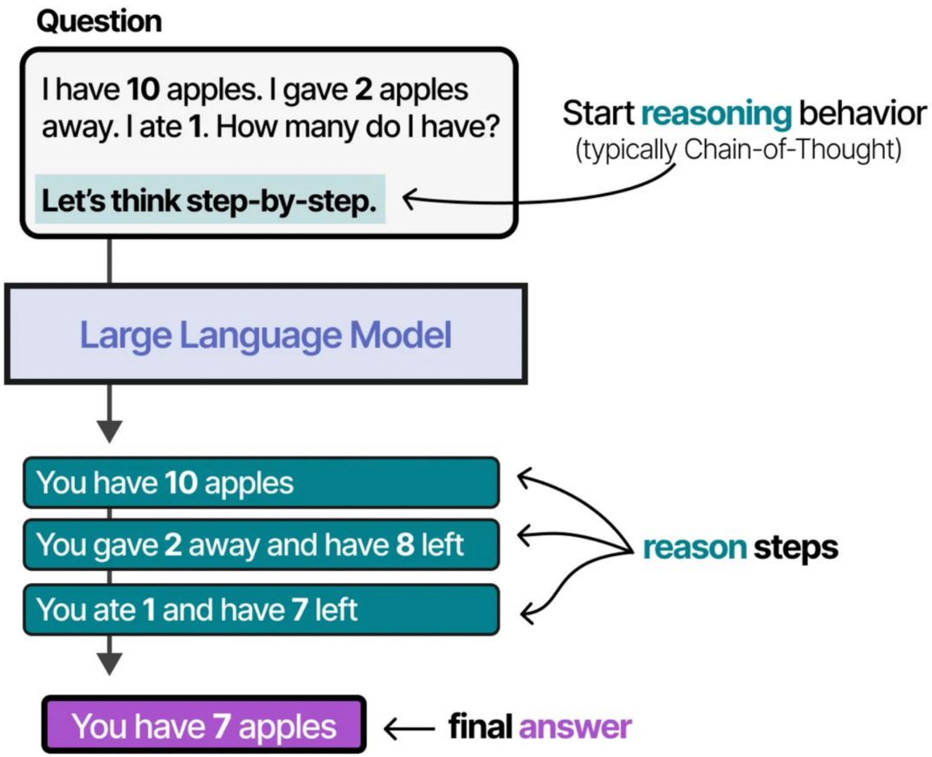
C. I've Got Some News For You. Being Gay Is Genetic. Being Irrationally Afraid Of Gay, Not So Much.
D. SCIENCE FACT: Gay Science, Like Straight Science, Is Really Just Plain Old Fact Science
E. If You Know Anyone Who Is Afraid Of Gay People, Here's A Cartoon That Will Ease Them Back To Reality

E

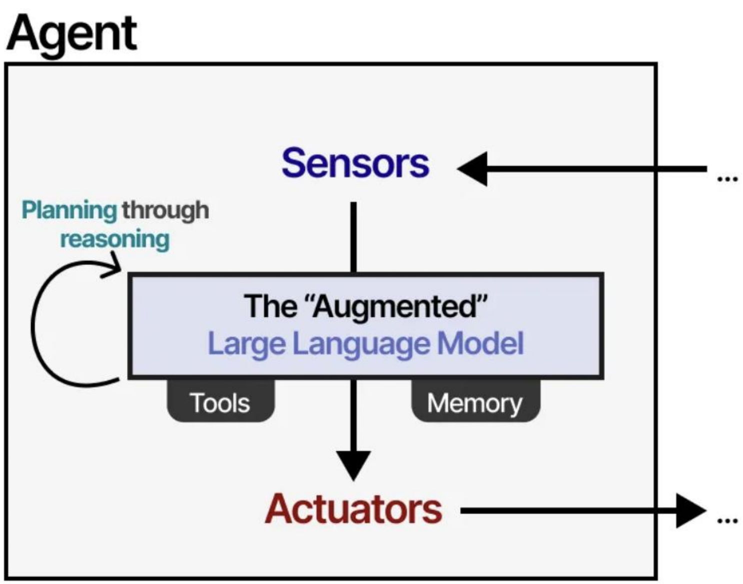
3. AI Agents

*An **agent** is anything that can be viewed as perceiving its environment through **sensors** and acting upon that environment through **actuators**.*

— Russell & Norvig, AI: A Modern Approach (2016)

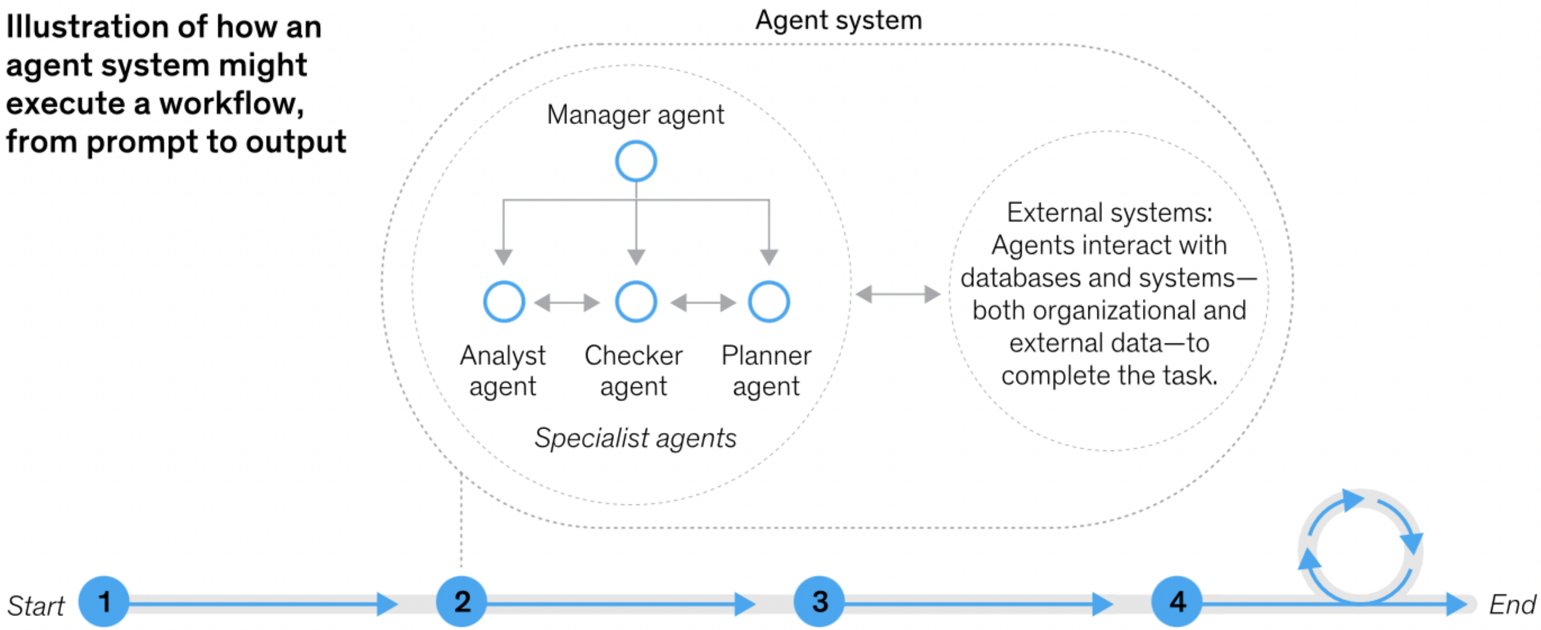

Question
I have 10 apples. I gave 2 apples away. I ate 1. How many do I have?
Let's think step-by-step.
Start reasoning behavior
(typically Chain-of-Thought)
Large Language Model
You have 10 apples
You gave 2 away and have 8 left
You ate 1 and have 7 left
reason steps
You have 7 apples
final answer


Step 1
Search the web
Step 2
Summarize results
Step 3
Infer best product
⋮
Step n
....


Agent
Sensors
...
Planning through reasoning
The "Augmented" Large Language Model
Tools
Memory
Actuators
...

Agents enabled by generative AI soon could function as hyperefficient virtual coworkers.

Illustration of how an agent system might execute a workflow, from prompt to output



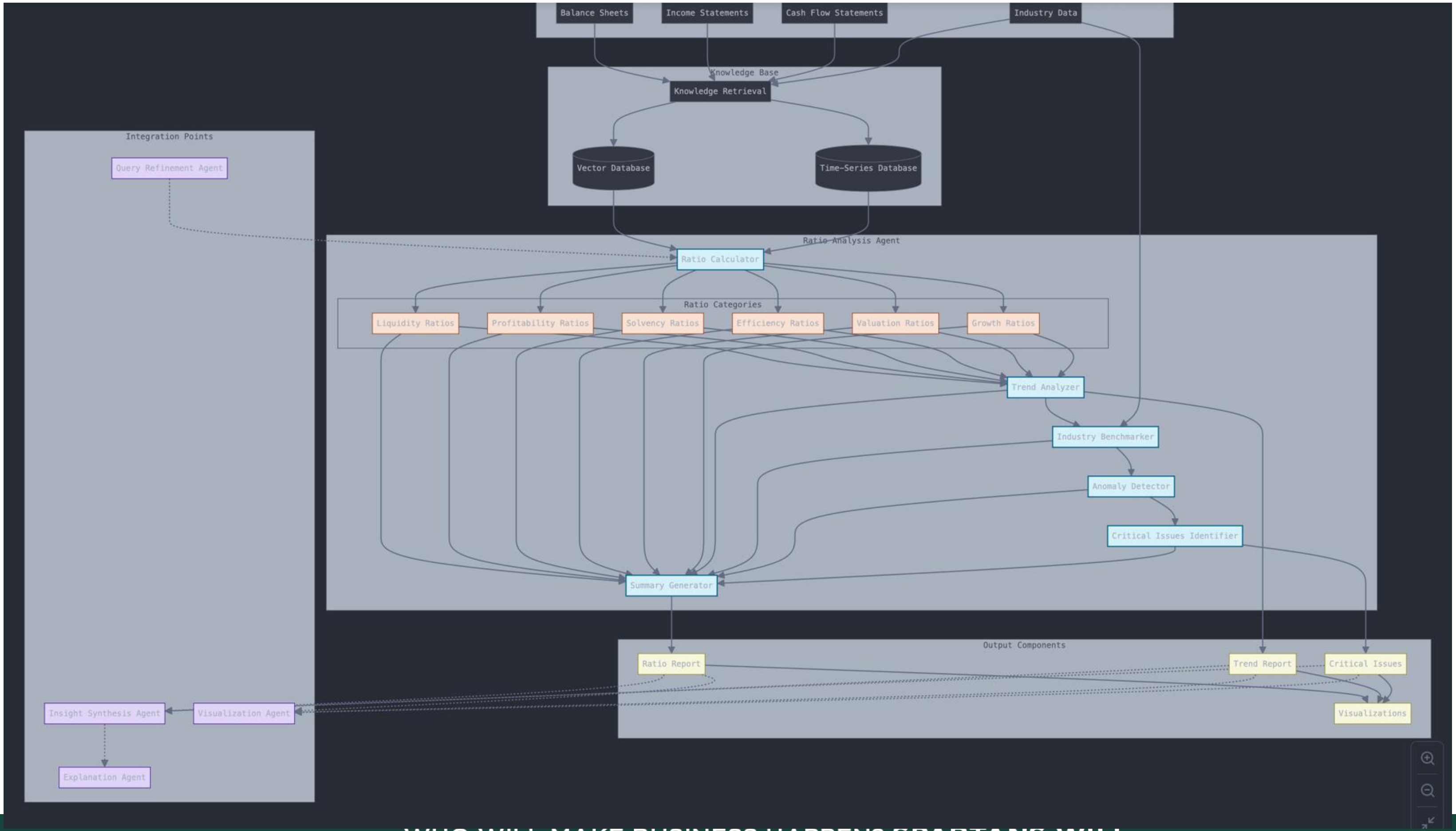
1. Using natural language, the user prompts the generative AI agent system to complete a task.
2. The agent system interprets the prompt and builds a work plan. A manager agent subdivides the project into tasks assigned to specialist agents; they gather and analyze data from multiple sources and collaborate with one another to execute their individual missions.
3. The agent team shares the draft output with the user.
4. The agent team receives user feedback, then iterates and refines output accordingly.

McKinsey & Company

Sequential task processing with LLM+ agents

Prompt: can I get an agent-based workflow along with rag for financial statement analysis

Workflow generated by Claude (along with code)



Sample Agent Workflows

1. Profitability Analysis Workflow

User Query: "How has the company's profitability changed over the last 3 years?" 1. Query Refinement Agent → Translates to specific profitability metrics to examine 2. Retrieval Agent → Fetches relevant income statements across time periods 3. Ratio Analysis Agent → Calculates gross margin, operating margin, net margin, ROA, ROE 4. Trend Analysis Agent → Identifies patterns in profitability metrics 5. Peer Comparison Agent → Benchmarks against industry averages 6. Insight Synthesis Agent → Summarizes key findings on profitability trends 7. Explanation Agent → Provides context on factors influencing profitability 8. Visualization Agent → Creates margin trend charts

2. Liquidity Assessment Workflow

User Query: "Is the company facing any liquidity risks?" 1. Query Refinement Agent → Identifies liquidity metrics to assess 2. Retrieval Agent → Fetches balance sheet, cash flow statements, and notes 3. Ratio Analysis Agent → Calculates current ratio, quick ratio, cash ratio 4. Cash Flow Analysis Agent → Analyzes operating, investing, and financing cash flows 5. Debt Structure Agent → Reviews debt obligations and maturity schedule 6. Risk Assessment Agent → Evaluates liquidity position and potential constraints 7. Insight Synthesis Agent → Provides holistic liquidity risk assessment 8. Alert Agent → Flags any concerning liquidity trends

4. Gen AI as an aid in research

A few ways to use AI in science

Scientific ideas

Scientific data

- Semantic search
- Literature review
- Generating ideas / hypotheses
- Evaluating research
- Detecting errors
- Writing code
- Surrogate for subjects

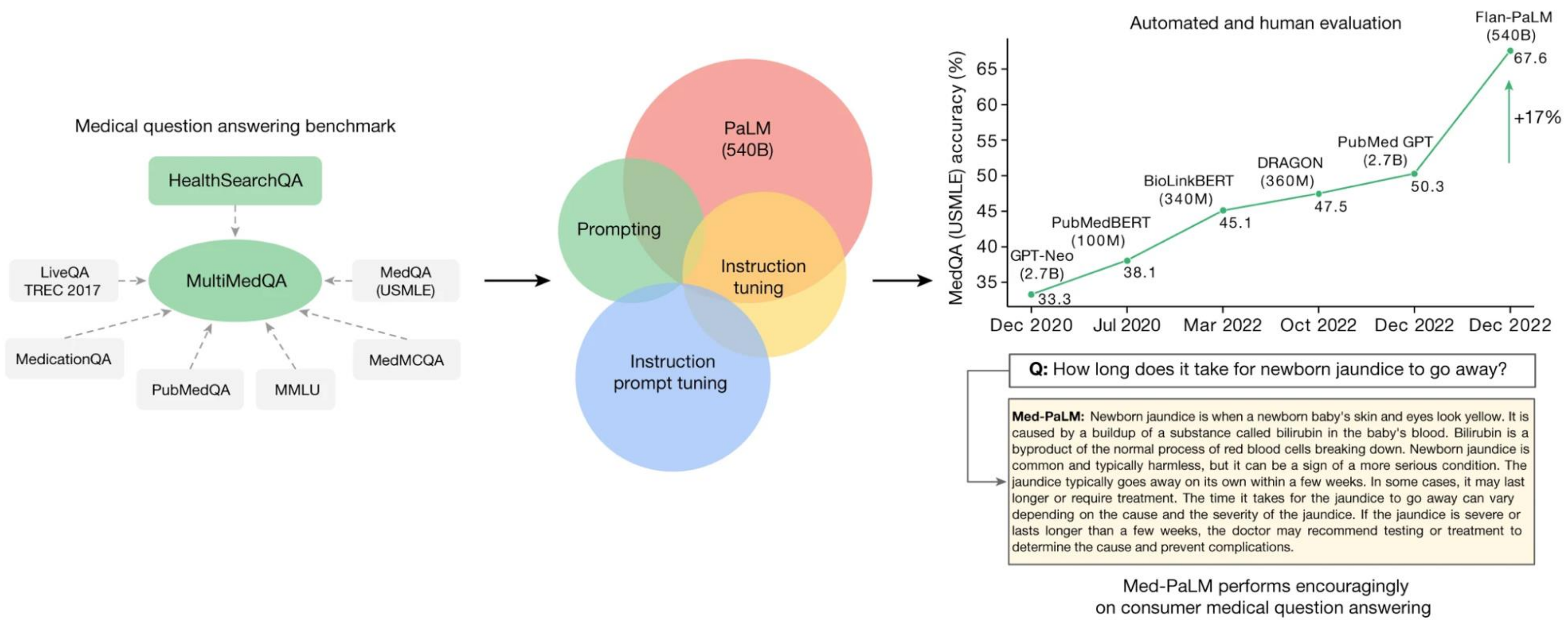
Newly possible due to generative AI

- Processing data
- Modeling the world

Can be done with traditional ML

Source: Arvind Narayanan @random_walker

Opportunities



[Source](#)

PERSPECTIVE | SOCIAL SCIENCES

Can Generative AI improve social science?

Christopher A. Bail Authors Info & Affiliations

Edited by David Lazer, Northeastern University, Boston, MA; received September 7, 2023; accepted April 5, 2024, by Editorial Board Member Mark Granovetter

May 9, 2024 | 121 (21) e2314021121 | https://doi.org/10.1073/pnas.2314021121

THIS ARTICLE HAS BEEN UPDATED

PERSPECTIVE | COMPUTER SCIENCES

How should the advancement of large language models affect the practice of science?

Marcel Binz, Stephan Alaniz, Adina Roskies, +14, and Eric Schulz Authors Info & Affiliations

Edited by David Kellen, Syracuse University, Syracuse, NY; received February 23, 2024; accepted October 21, 2024 by Editorial Board Member Elke U. Weber

January 27, 2025 | 122 (5) e2401227121 | https://doi.org/10.1073/pnas.2401227121

What can LLMs do for business school style research?

Text as data by M. Gentzkow, B.T. Kelly, and M. Taddy: general introductory survey.

- Text algorithms in economics by E. Ash and S. Hansen: general introductory survey.
- A User's Guide to GPT and LLMs for Economic Research by K. Bryan: examples of how to use LLM in your daily research.
- Second half of https://youtu.be/bZQun8Y4L2A by A. Karpathy: nice tricks for good prompting.
- Language Models and Cognitive Automation for Economic Research by A. Korinek: application of LLM for ideation, writing, background research, data analysis, coding, and mathematical derivations.

A toy example: ChatGPT for sentiment analysis

1. **Step 1. Scrape/ gather a few online reviews (can be financial texts as well)**
2. **Step 2. Classify the text as positive/negative/ neutral using ChatGPT**
3. **Step 3. Run an aspect-based sentiment analysis**
4. **Experiment with prompts and compare different reviews/tweets/texts**

Example 1:

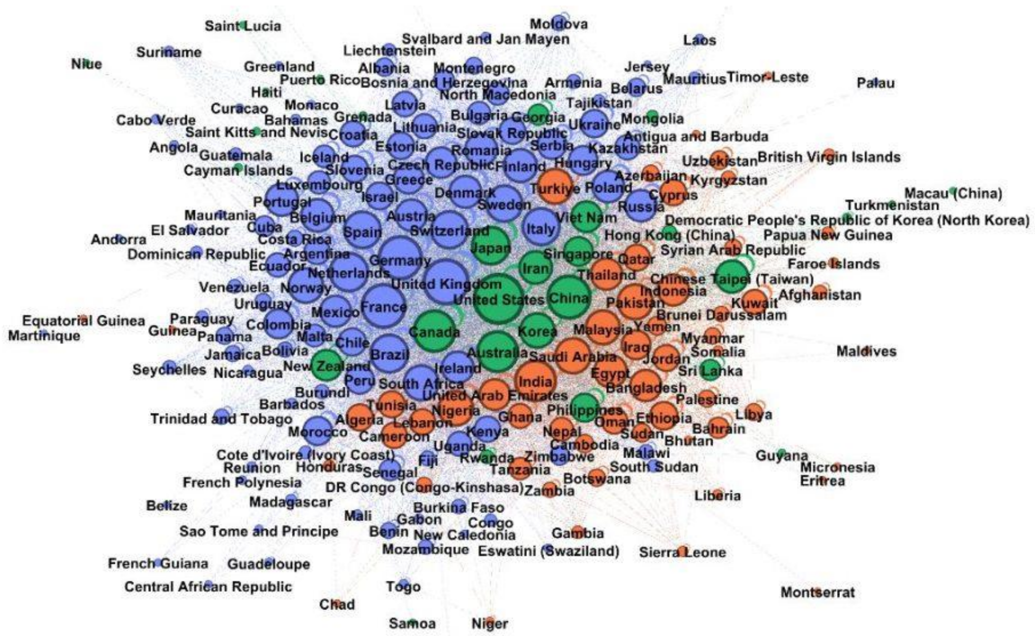
I took the following product:
SAMSUNG 65-Inch Class Crystal UHD 4K CU8000 Series PurColor,Object Tracking Sound Lite, Q-Symphony, Motion Xcelerator, Ultra Slim, Solar Remote, Smart TV with Alexa Built-in (UN65CU8000,2023 Model)

Prompt:

- How do you feel about the following review: "The tv is good enough. The colors are brighter than our tv from 15-20 years ago that this replaced, but everything has a pink tinge to it no matter if I choose the "cool", "standard", "warm", or "warm 2" options. I've played with the contrast, etc too and haven't been able to get it completely gone. Not a big deal for us, but would prefer the skin tones of people to be more accurate in a future tv. This tv will NOT pair with our Dish hopper. I've tried a few different ways and it cannot find it. It does work as the PS3 remote, so that is nice. Not a huge deal for us either as Mom is used to the Dish remote and didn't want to learn a new one. We just switch to the Samsung tv remote when we want to use the PS3 or view an USB drive on the tv. Speaking of the USB drives...they have issues. Sometimes the tv will find it, sometimes not. I've tried reformatting the drive, making sure the size files are in range of what the tv can handle, etc. The only thing I've found to fix the issue is unplug the tv and plug it back in. It is kinda cool that you can put your own pictures on the tv. We use this as a cheap "Frame TV" and have a picture on the screen. The downside is since this isn't a proper "Frame TV" after an hour or two a screensaver comes on so the image doesn't burn into the tv. It's a nice saving feature, but kinda defeats the purpose of having art on it. There may be a way to turn the screen saver off but I haven't looked. We like that the tv has all the streaming apps we use: YouTube, Amazon, Netflix, Hulu, Disney+, etc."? Answer in one token: positive, negative, or neutral.

A toy example 2: ChatGPT for network analysis

Give me 20 pairs of the most important characters in the book "lord of the rings" with the weight of joint appearance of such pairs in the format "character1; character 2; weight". Each entry is on a new line



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†

Anton Korinek*

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)

EconNLI: Evaluating Large Language Models on Economics Reasoning

Yue Guo **Yi Yang**
The Hong Kong University of Science and Technology
yguoar@connect.ust.hk imyiyang@ust.hk

LLMs for difficult to elicit constructs

Example: Congressional legislation Consider descriptions of bills introduced in the United States Congress. Each text piece $r \in R$ refers to a bill's brief description such as "A bill to revise the boundary of Crater Lake National Park in the State of Oregon." The economically relevant outcome Y_r might be whether the associated bill passed its originating chamber of Congress. The candidate economic determinant W_r of the bill's text might be the party affiliation or roll-call voting score of the bill's sponsor.

Example: Financial news headlines Consider financial news headlines about publicly traded companies. Each text piece $r \in R$ refers to a particular financial news headline such as "Bank of New York Mellon Q1 EPS $0.94 Misses $0.97 Estimate, Sales $3.9B Misses $4.01B Estimate." The economic outcome Y_r might be the company's realized return in some event window after the headline's publication date, while the candidate determinant W_r of the news headline itself could be the company's past fundamentals.

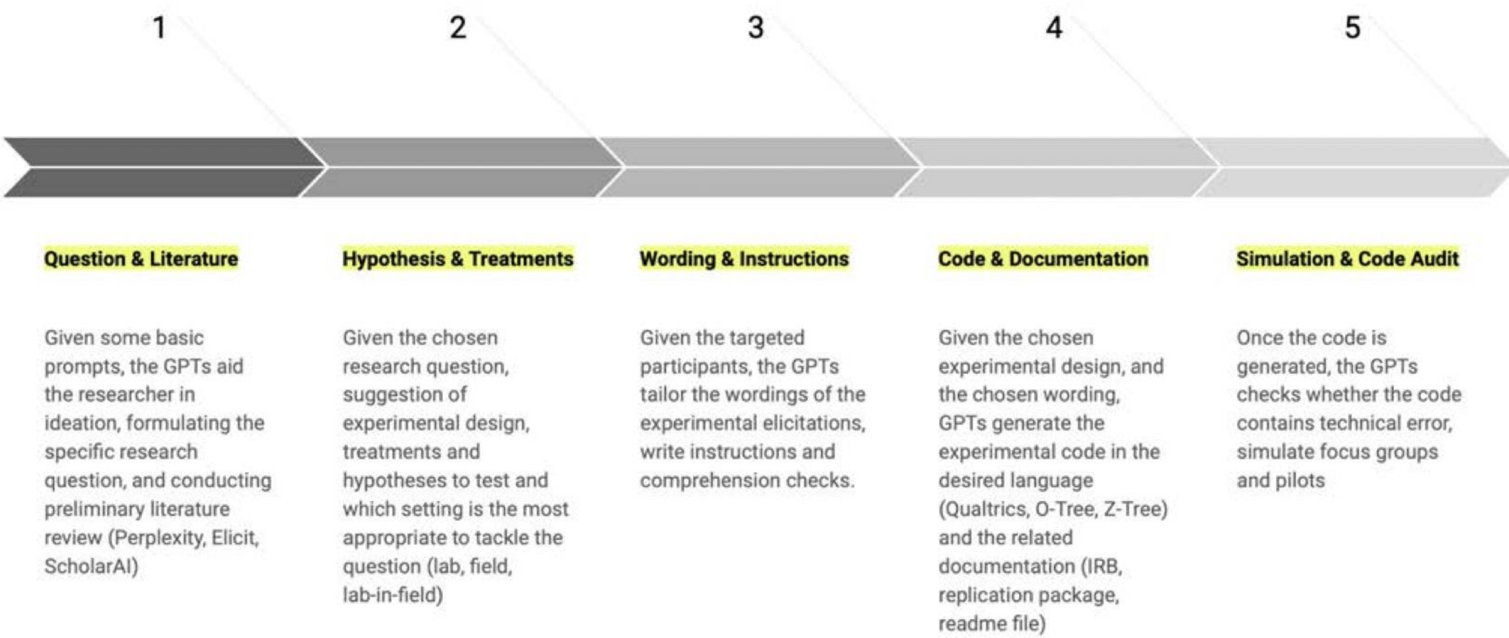
Example from Ludwig, Mullainathan and Rambachan (2025)

NBER WORKING PAPER SERIES

GENERATION NEXT: EXPERIMENTATION WITH AI

Gary Charness
Brian Jabarian
John A. List

Working Paper 31679
http://www.nber.org/papers/w31679



Can AI Replace Human Subjects? A Large-Scale Replication of Psychological Experiments with LLMs

Ziyan Cui[1], cuizy21@mails.tsinghua.edu.cn

Ning Li[1], lining@sem.tsinghua.edu.cn

Huaikang Zhou[1], zhouhk@sem.tsinghua.edu.cn

[1] Tsinghua University, School of Economics and Management

Scenario-based experiments published in five journals between 2015-2024
Organizational Behavior and Human Decision Processes (OBHDP), *Academy of Management Journal* (AMJ), *Journal of Applied Psychology* (JAP), *Journal of Personality and Social Psychology* (JPSP), and *Journal of Experimental Psychology: General* (JEP)

nature communications

Explore content About the journal Publish with us

nature > nature communications > articles > article

Article | Open access | Published: 06 December 2022

Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy

János Kramár, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin R. McKee, Mateusz Malinowski, Thore Graepel & Yoram Bachrach

Nature Communications **13**, Article number: 7214 (2022) | Cite this article

PROPEL: Supervised and Reinforcement Learning for Large-Scale Supply Chain Planning

Vahid Eghbal Akhlaghi, Reza Zandehshahvar, and Pascal Van Hentenryck
NSF Artificial Intelligence Institute for Advances in Optimization (AI4OPT)
Georgia Institute of Technology
Email: pvh@gatech.edu

April 11, 2025

ABSTRACT

This paper considers how to fuse Machine Learning (ML) and optimization to solve large-scale Supply Chain Planning (SCP) optimization problems. These problems can be formulated as MIP models which feature both integer (non-binary) and continuous variables, as well as flow balance and capacity constraints. This raises fundamental challenges for existing integrations of ML and optimization that have focused on binary MIPs and graph problems. To address these, the paper proposes PROPEL, a new framework that combines optimization with both supervised and Deep Reinforcement Learning (DRL) to reduce the size of search space significantly. PROPEL uses supervised learning, not to predict the values of all integer variables, but to identify the variables that are fixed to zero in the optimal solution, leveraging the structure of SCP applications. PROPEL includes a DRL component that selects which fixed-at-zero variables must be relaxed to improve solution quality when the supervised learning step does not produce a solution with the desired optimality tolerance. PROPEL has been applied to industrial supply chain planning optimizations with millions of variables. The computational results show dramatic improvements in solution times and quality, including a 60% reduction in primal integral and an 88% primal gap reduction, and improvement factors of up to 13.57 and 15.92, respectively.

ORMS TODAY informs membership magazine

NEWS FEATURES PODCASTS DEPARTM

March 3, 2025 in Smarter Decisions

A Prominent Role for INFORMS in the Age of AI

Bringing Together AI and OR/MS for Better Organizational and Societal Decision-Making

By Segev Wasserkrug, Vinod Cheriyan, Tinglong Dai, Juan R. Jaramillo, Sven Koenig, Ramayya Krishnan, Radhika Kulkarni, Henry Lam, Fred Oswald, Thiago Serra, Mark S. Squillante, Anjana Susarla, Pascal Van Hentenryck, Holly Wiberg

SHARE: PRINT ARTICLE: https://doi.org/10.1287/orms.2025.01.04

INFORMATION SYSTEMS RESEARCH

JOURNAL HOME ARTICLES IN ADVANCE CURRENT ISSUE ARCHIVES ABOUT SUBMIT SUBSCRIBE

Free Access

View PDF

Tools Share

Home > Information Systems Research > Vol. 34, No. 2 >

The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems

Anjana Susarla, Ram Gopal, Jason Bennett Thatcher, Suprateek Sarker

Published Online: 26 May 2023 | https://doi.org/10.1287/isre.2023.ed.v34.n2

5. Bias in Generative AI

Bias in Language Models: Example from @kareem_carr

What happens when you don't understand what your algorithm is learning?

We want these sentences to all give the same score but they don't.

```
text_to_sentiment("Let's go get Italian food")
2.0429166109
text_to_sentiment("Let's go get Chinese food")
1.4094033658
text_to_sentiment("Let's go get Mexican food")
0.3880198556
```

The algorithm is probably accurately learning real feelings of people based on the data but it's not learning what we intended it to learn.

We never told the algorithm that we didn't want to learn racism!

"My name is ______" is a neutral statement so the score should be about 0.

```
text_to_sentiment("My name is Emily")
2.2286179365
text_to_sentiment("My name is Heather")
1.3976291151
text_to_sentiment("My name is Yvette")
0.9846380213
text_to_sentiment("My name is Shaniqua")
-0.4704813178
```

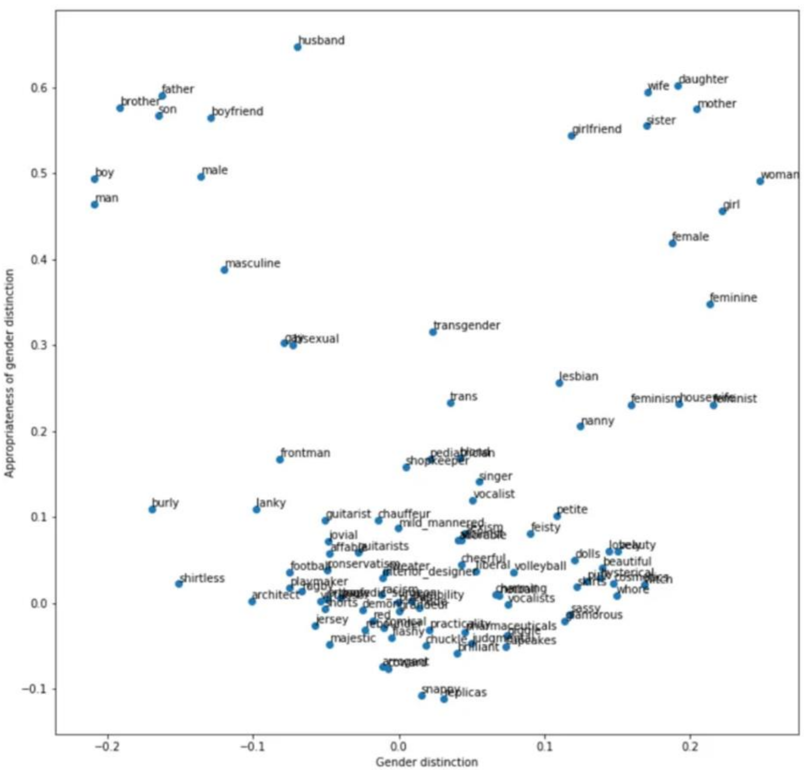
We wanted to learn the sentiment score of the sentence "My name is ______" which should be independent of the particular name used Emily, Shaniqua, etc

@kareem_carr

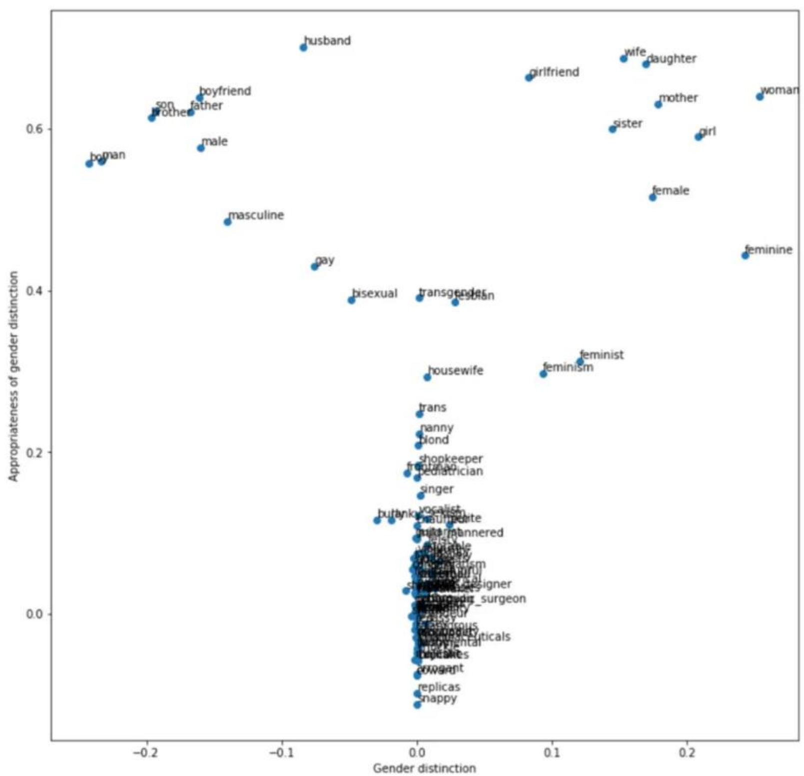
partially based on analysis by Robyn Speer and images by Mark Xiang)

Solutions exist. (Example from Robyn Speer)

Word associations before de-biasing



Word associations after de-biasing



Bias: Stable Diffusion’s Rendering of “CEO”

Generated with StableDiffusion

Generative AI hallucinations

J

write a bio for Professor Anjana Susarla

Professor Anjana Susarla is a highly respected academic in the field of computer science and engineering. She received her PhD from the Massachusetts Institute of Technology and has since gone on to become a leading expert in the area of human-computer interaction and user experience design. Susarla has published numerous papers and articles in top-tier academic journals and has also been the recipient of several prestigious awards and grants for her research. In addition to her research, Susarla is also an active member of the academic community, regularly giving keynote speeches and presentations at conferences and workshops around the world. She is currently a professor at Georgia Institute of Technology.

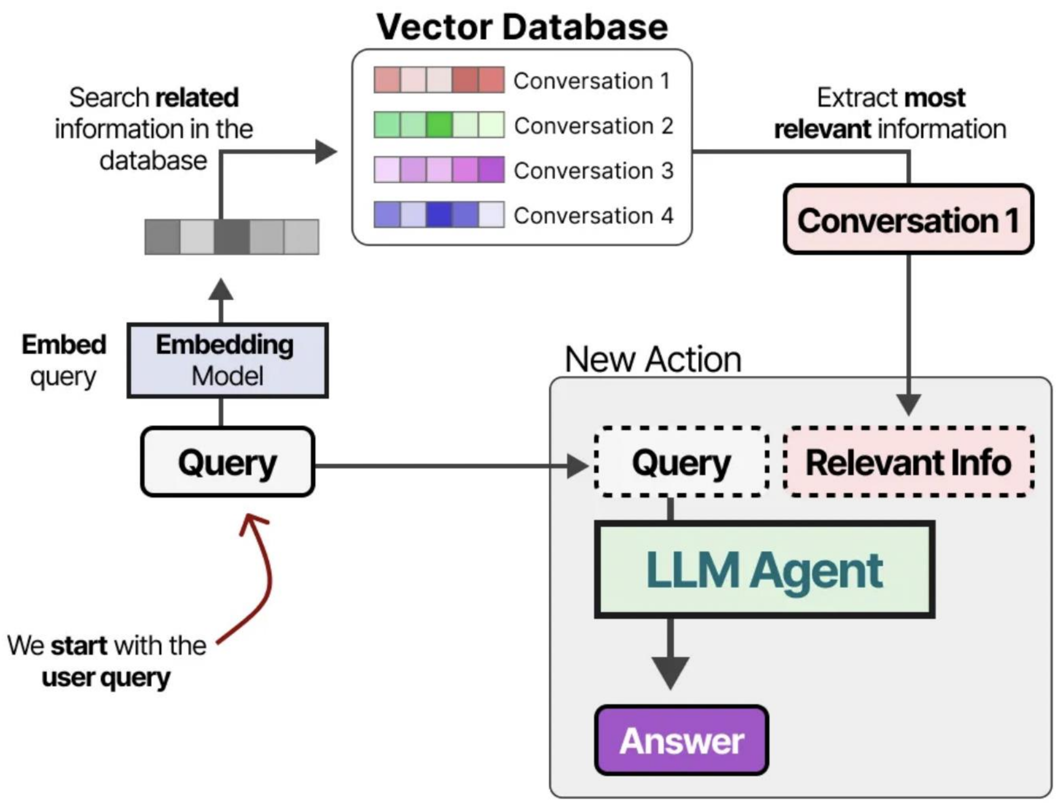
Once again, solutions exist..

1. We feed a question to the LLM and generate two responses ...

2. We automatically figure out which response is more factual

3. We do (1) and (2) for many questions and use this data to update the LLM weights using a technique called DPO

Retrieval Augmented Generation (RAG)



A personal anecdote

Download This Paper | Open PDF in Browser | Add Paper to My Library

Does Increasing Reliance on Artificial Intelligence Boost Creativity? Assessing AI-Augmented Creativity with Large Language Models

57 Pages • Posted: 24 Jul 2024

Jiaoping Chen
Michigan State University - Eli Broad College of Business

Laura Brandimarte
University of Arizona - Eller College of Management

Anjana Susarla
Michigan State University - The Eli Broad College of Business and The Eli Broad Graduate School of Management

Date Written: July 17, 2024

Some issues from Notebook LM

- Seemingly coherent as it tries to place papers in their broader context — and often even accurate, though it frequently makes massive errors very confidently.
- It makes up its own analogies and incorporates really human-sounding linguistic and conversational features

Generalization

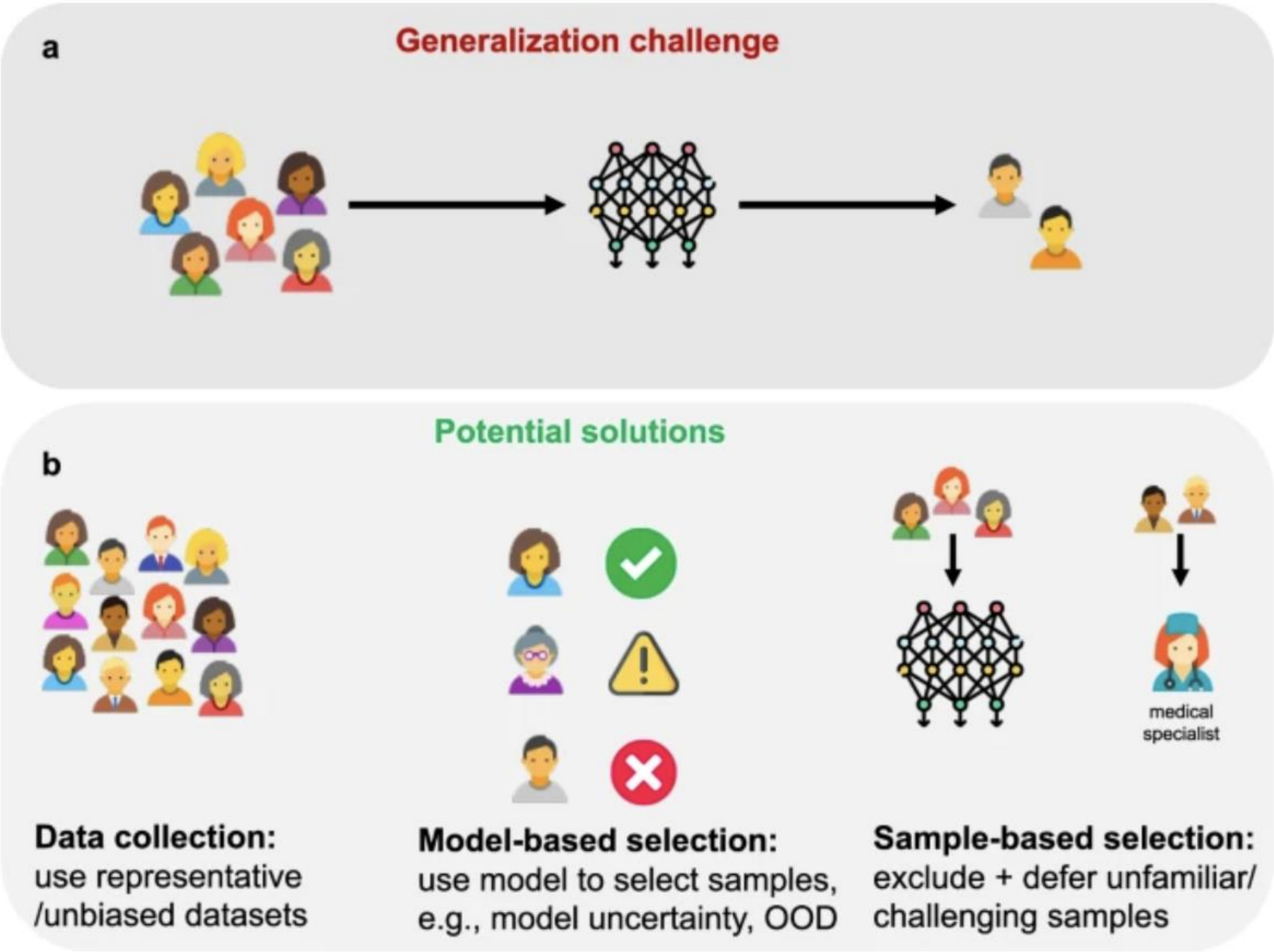
Comment | Open access | Published: 21 May 2024

Generalization—a key challenge for responsible AI in patient-facing clinical applications

Lea Goetz, Nabeel Seedat, Robert Vandersluis & Mihaela van der Schaar

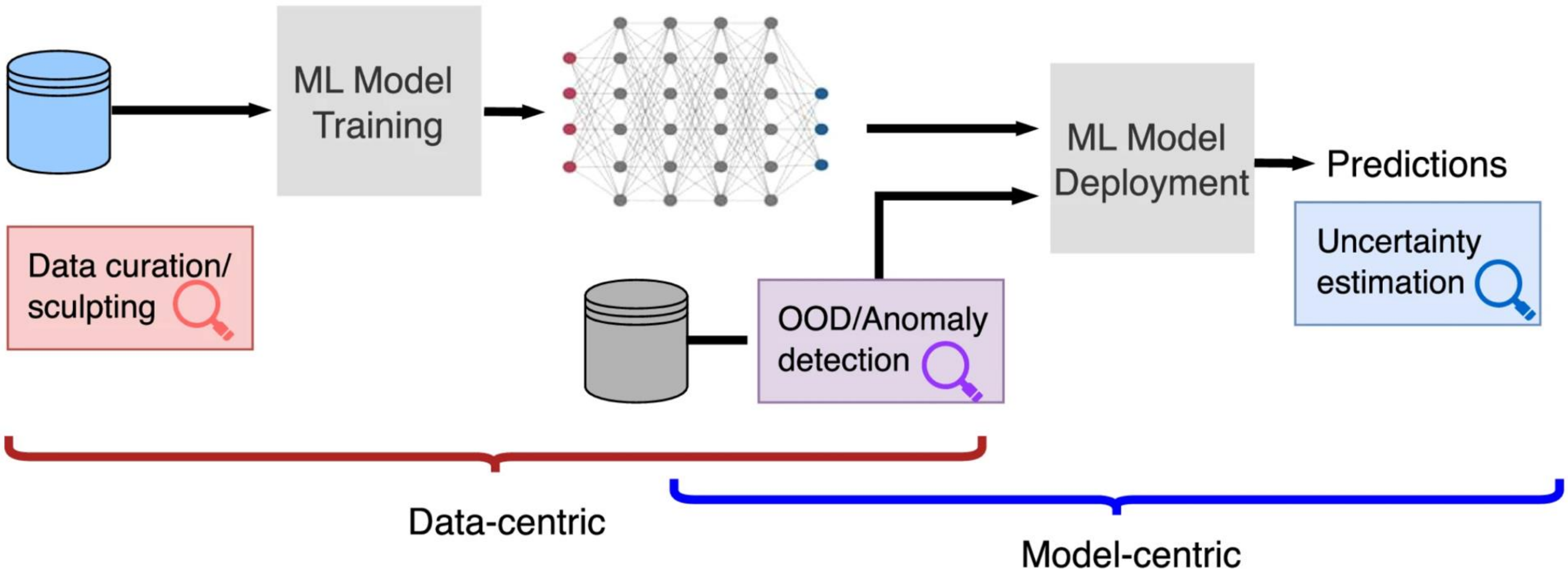
npj Digital Medicine **7**, Article number: 126 (2024) | Cite this article

Generalization – the ability of AI systems to apply and/or extrapolate their knowledge to new data which might differ from the original training data – is a major challenge for the effective and responsible implementation of human-centric AI applications. Current debate in bioethics proposes selective prediction as a solution. Here we explore data-based reasons for generalization challenges and look at how selective predictions might be implemented technically, focusing on clinical AI applications in real-world healthcare settings.


a
Generalization challenge
Potential solutions
b
medical specialist
Data collection:
use representative /unbiased datasets
Model-based selection:
use model to select samples, e.g., model uncertainty, OOD
Sample-based selection:
exclude + defer unfamiliar/ challenging samples

Sample selection to avoid generalization..

From: Generalization—a key challenge for responsible AI in patient-facing clinical applications



Sample selection can be achieved by data-centric methods of data curation/sculpting before training the model, or model-centric sample deferral with uncertainty estimation. Out-of-distribution and anomaly detection methods lie at the intersection, wherein we flag samples preemptively.

6. Best practices with Gen AI

How can we best equip the current generation of students and postdoctoral researchers at our institutions to make responsible decisions in the use of Generative AI technology?

Be aware of pitfalls

Be aware of generalization issues

Be aware of pitfalls..

Challenges Posed by Generative AI for Scholarly Work

1. **Institutionalizing bias:** Generative AI relies on training data and encoded rules that reflect human biases, resulting in further reification of social problems in broader society or flawed conclusions in the literature.
2. **Hallucinations:** Generative AI fabricates false or misleading responses to prompts from scholars.
3. **Interpretability:** Generative AI offers suggestions with scant explanation for why an output was created, creating issues related to credible inference and interpretation.
4. **Inappropriate Training Data:** Generative AI offers inappropriate guidance because it has not been trained on data information relevant to a research topic or method.
5. **Misapplication:** Generative AI is used to complete peer reviews, lead literature searches, or write key elements of papers without appropriate supervision.
6. **Unreasonable expectations:** Scholars may expect access to generative AI to result in unreasonably sophisticated understanding of literature or expert applications of state-of-the-art research methods

Suggestions for Mitigating Problems Posed by Generative AI

1. **Independent Knowledge:** When studying a topic or using a method, a scholar must develop a baseline of knowledge and experience necessary to prompt as well as assess the outputs of generative AI.
2. **Critical Thinking:** When reviewing the output of generative AI, scholars should consider the following: (a) the face validity: are the outputs consistent with the scholar's understanding of the literature? (b) falsifiability, is it possible—through logic or evidence—to critically examine and subject the outputs provided by generative AI to invalidation? This helps ensure that humans guide the application of AI.
3. **Awareness of Implications:** As scholars consider the output of generative AI, they need to actively question the biases and their implications for people and organizations. Scholars should demonstrate in their work an awareness of multiple perspectives and interpretations so that readers understand the potential biases from these tools have been taken into consideration.
4. **Understanding Provenance:** When using generative AI, scholars must understand the data that was used to train the tool.
5. **Ethical Conduct:** Scholars should use generative AI as per current professional norms. Writing and reviewing papers are critical activities that the researchers need to be undertake cautiously and responsibly when using generative AI.

Guiding Principles for Scholars Seeking to Apply Generative AI in Scholarly Work

1. **Human Primacy:** Scholars retain decision rights over key elements of research.
2. **Responsible Reporting:** Scholars faithfully report (1) which generative AI was used, (2) what data were used to train the AI (if known), and (3) how the tool was used to support scholarly inquiry.

Guidelines for researchers

Remain ultimately responsible for scientific output.

• Researchers are accountable for the integrity of the content13 generated by or with the support of AI tools.

• Researchers maintain a critical approach to using the output produced by generative AI and are aware of the tools' limitations, such as bias, hallucinations14 and inaccuracies.

• AI systems are neither authors nor co-authors. Authorship implies agency and responsibility, so it lies with human researchers.

• Researchers do not use fabricated material created by generative AI in the scientific process, for example falsifying, altering or manipulating original research data.

Use generative AI transparently.

• Researchers, to be transparent, detail which generative AI tools have been used substantially15 in their research processes. Reference to the tool could include the name, version, date, etc. and how it was used and affected the research process. If relevant, researchers make the input (prompts) and output available, in line with open science principles.

• Researchers take into account the stochastic (random) nature of generative AI tools, which is the tendency to produce different output from the same input. Researchers aim for reproducibility and robustness in their results and conclusions. They disclose or discuss the limitations of generative AI tools used, including possible biases in the generated content, as well as possible mitigation measures. Pay particular attention to issues related to privacy, confidentiality and intellectual property rights when sharing sensitive or protected information with AI tools.

• Researchers remain mindful that generated or uploaded input (text, data, prompts, images, etc.) could be used for other purposes, such as the training of AI models. Therefore, they protect unpublished or sensitive work (such as their own or others' unpublished work) by taking care not to upload it into an online AI system unless there are assurances that the data will not be re-used, e.g., to train future language models or to the untraceable and unverifiable reuse of data.

When using generative AI, respect applicable national, EU and international legislation, as in their regular research activities.

- Researchers pay attention to the potential for plagiarism (text, code, images, etc.) when using outputs from generative AI. Researchers respect others' authorship and cite their work where appropriate. The output of a generative AI (such a large language model) may be based on someone else's results and require proper recognition and citation.
- The output produced by generative AI can contain personal data. If this becomes apparent, researchers are responsible for handling any personal data output responsibly and appropriately, and EU data protection rules are to be followed.

5. Continuously learn how to use generative AI tools properly to maximise their benefits, including by undertaking training.

6. Refrain from using generative AI tools substantially in sensitive activities that could impact other researchers or organisations (for example peer review, evaluation of research proposals, etc).

- Avoiding the use of generative AI tools eliminates the potential risks of unfair treatment or assessment that may arise from these tools' limitations (such as hallucinations and bias).
- Moreover, this will safeguard the original unpublished work of fellow researchers from potential exposure or inclusion in an AI model (under the conditions detailed above in the recommendation for researchers #3).

PNAS EDITORIAL

Protecting scientific integrity in an age of generative AI

Wolfgang Blau[a], Vinton G. Cerf[b], Juan Enriquez[c], Joseph S. Francisco[d], Urs Gasser[e], Mary L. Gray[f,g], Mark Greaves[h], Barbara J. Grosz[i], Kathleen Hall Jamieson[j], Gerald H. Haug[k], John L. Hennessy[l], Eric Horvitz[m], David I. Kaiser[n], Alex John London[o], Robin Lovell-Badge[p], Marcia K. McNutt[q,1], Martha Minow[r], Tom M. Mitchell[s], Susan Ness[j], Shobita Parthasarathy[t], Saul Perlmutter[u,v], William H. Press[w], Jeannette M. Wing[x], and Michael Witherell[y]

May 21, 2024 | 121 (22) e2407886121 | https://doi.org/10.1073/pnas.2407886121

Five Principles of Human Accountability and Responsibility

- **Transparent disclosure and attribution**
- *Scientists* should clearly disclose the use of generative AI in research, including the specific tools, algorithms, and settings employed; accurately attribute the human and AI sources of information or ideas, distinguishing between the two and acknowledging their respective contributions; and ensure that human expertise and prior literature are appropriately cited, even when machines do not provide such citations in their output.
- *Model creators and refiners* should provide publicly accessible details about models, including the data used to train or refine them; carefully manage and publish information about models and their variants so as to provide scientists with a means of citing the use of particular models with specificity; provide long-term archives of models to enable replication studies; disclose when proper attribution of generated content cannot be provided; and pursue innovations in learning, reasoning, and information retrieval machinery aimed at providing users of those models with the ability to attribute sources and authorship of the data employed in AI-generated content.

- **Verification of AI-generated content and analyses**
- *Scientists* are accountable for the accuracy of the data, imagery, and inferences that they draw from their uses of generative models. Accountability requires the use of appropriate methods to validate the accuracy and reliability of inferences made by or with the assistance of AI, along with a thorough disclosure of evidence relevant to such inferences. It includes monitoring and testing for biases in AI algorithms and output, with the goal of identifying and correcting biases that could skew research outcomes or interpretations.
- *Model creators* should disclose limitations in the ability of systems to confirm the veracity of any data, text, or images generated by AI. When verification of the truthfulness of generated content is not possible, model output should provide clear, well-calibrated assessments of confidence. Model creators should proactively identify, report, and correct biases in AI algorithms that could skew research outcomes or interpretations.

Documentation of AI-generated data

- Scientists should mark AI-generated or synthetic data, inferences, and imagery with provenance information about the role of AI in their generation, so that it is not mistaken for observations collected in the real world. Scientists should not present AI-generated content as observations collected in the real world.
- Model creators should clearly identify, annotate, and maintain provenance about synthetic data used in their training procedures and monitor the issues, concerns, and behaviors arising from the reuse of computer-generated content in training future models.

A focus on ethics and equity

Scientists and model creators should take credible steps to ensure that their uses of AI produce scientifically sound and socially beneficial results while taking appropriate steps to mitigate the risk of harm. This includes advising scientists and the public on the handling of tradeoffs associated with making certain AI technologies available to the public, especially in light of potential risks stemming from inadvertent outcomes or malicious applications.

Scientists and model creators should adhere to ethical guidelines for AI use, particularly in terms of respect for clear attribution of observational versus AI-generated sources of data, intellectual property, privacy, disclosure, and consent, as well as the detection and mitigation of potential biases in the construction and use of AI systems. They should also continuously monitor other societal ramifications likely to arise as AI is further developed and deployed and update practices and rules that promote beneficial uses and mitigate the prospect of social harm.

Scientists, model creators, and policymakers should promote equity in the questions and needs that AI systems are used to address as well as equitable access to AI tools and educational opportunities. These efforts should empower a diverse community of scientific investigators to leverage AI systems effectively and to address the diverse needs of communities, including the needs of groups that are traditionally underserved or marginalized. In addition, methods for soliciting meaningful public participation in evaluating equity and fairness of AI technologies and uses should be studied and employed.

AI should not be used without careful human oversight in decisional steps of peer review processes or decisions around career advancement and funding allocations.

Continuous monitoring, oversight, and public engagement

- Scientists, together with representatives from academia, industry, government, and civil society, should continuously monitor and evaluate the impact of AI on the scientific process, and with transparency, adapt strategies as necessary to maintain integrity. Because AI technologies are rapidly evolving, research communities must continue to examine and understand the powers, deficiencies, and influences of AI; work to anticipate and prevent harmful uses; and harness its potential to address critical societal challenges. AI scientists must at the same time work to improve the effectiveness of AI for the sciences, including addressing challenges with veracity, attribution, explanation, and transparency of training data and inference procedures. Efforts should be undertaken within and across sectors to pursue ongoing study of the status and dynamics of the use of AI in the sciences and pursue meaningful methods to solicit public participation and engagement as AI is developed, applied, and regulated. Results of this engagement and study should be broadly disseminated.

Thank you!