



# Generative AI and Society

Introduction

Jan Zilinsky

# What do chatbots provide?

- Information?
- Stereotypes?

```
ask4oTerse <- function(prompt="This a default",echo=NULL) {  
  chat <- chat_openai(  
    model = "gpt-4o-mini",  
    api_args = list(temperature = 0),  
    system_prompt = "You are a terse assistant who gives one-word answers to questions.",  
    echo=echo)  
  
  chat$chat(prompt)  
}  
...
```

*Responses from GPT-4o mini*

```
> ask4oTerse("What's in Russia?")  
Land.  
> ask4oTerse("What's in Germany?")  
Beer.
```

# What do chatbots provide?

- Information?
- Stereotypes?



```
> ask4oTerse("What's on Mars?")  
Rovers.  
> ask4oTerse("What's in the Mariana Trench?")  
Creatures.
```

# What do chatbots provide?

- Information?
- Stereotypes?
- Data

```
ask4oTerse <- function(prompt="This a default",echo=NULL) {  
  chat <- chat_openai(  
    model = "gpt-4o-mini",  
    api_args = list(temperature = 0),  
    system_prompt = "You are a terse assistant who gives one-word answers to questions.",  
    echo=echo)  
  
  chat$chat(prompt)  
}  
...
```

```
> ask4oTerse("What's in Russia?") Correct  
Land.  
> ask4oTerse("What's in Germany?") Not the main thing  
Beer.  
> ask4oTerse("What's on Mars?") Wrong  
Rovers.  
> ask4oTerse("What's in the Mariana Trench?")  
Creatures. Who knows...
```

# Reminders

- Models excel at some tasks (often those that are quite difficult, computationally, for humans)
- But models will “lie”
- You need to test queries multiple times
- Set a low temperature parameter
  - But even this won’t guarantee consistent results

# A real-world political information provision tool

Kamala Harris FAQ: Ask our AI news assistant your election questions

[www.sfchronicle.com](http://www.sfchronicle.com)

Whether you're curious about Harris' policies, her career milestones or just have questions about her life story, this news assistant was built to help. Type in your question to receive insights and links to relevant articles from the Chronicle's archive and recent coverage. Below you'll also find a selection of frequently asked questions.

NOTE: This tool relies on stories written and edited by the Chronicle's staff, though it doesn't always have the most recent news in its database and some responses will reflect the time period when the originating articles were written. For details on our use of artificial intelligence, [read our editorial guidelines](#).



**What would you  
like to know?**

0/200→

OR ASK ONE OF THESE:

[Tell me about Kamala Harris' childhood in Oakland and Berkeley.](#)

[What did Harris accomplish as San Francisco district attorney?](#)

# Malfunctioning NYC AI Chatbot Still Active Despite Widespread Evidence It's Encouraging Illegal Behavior

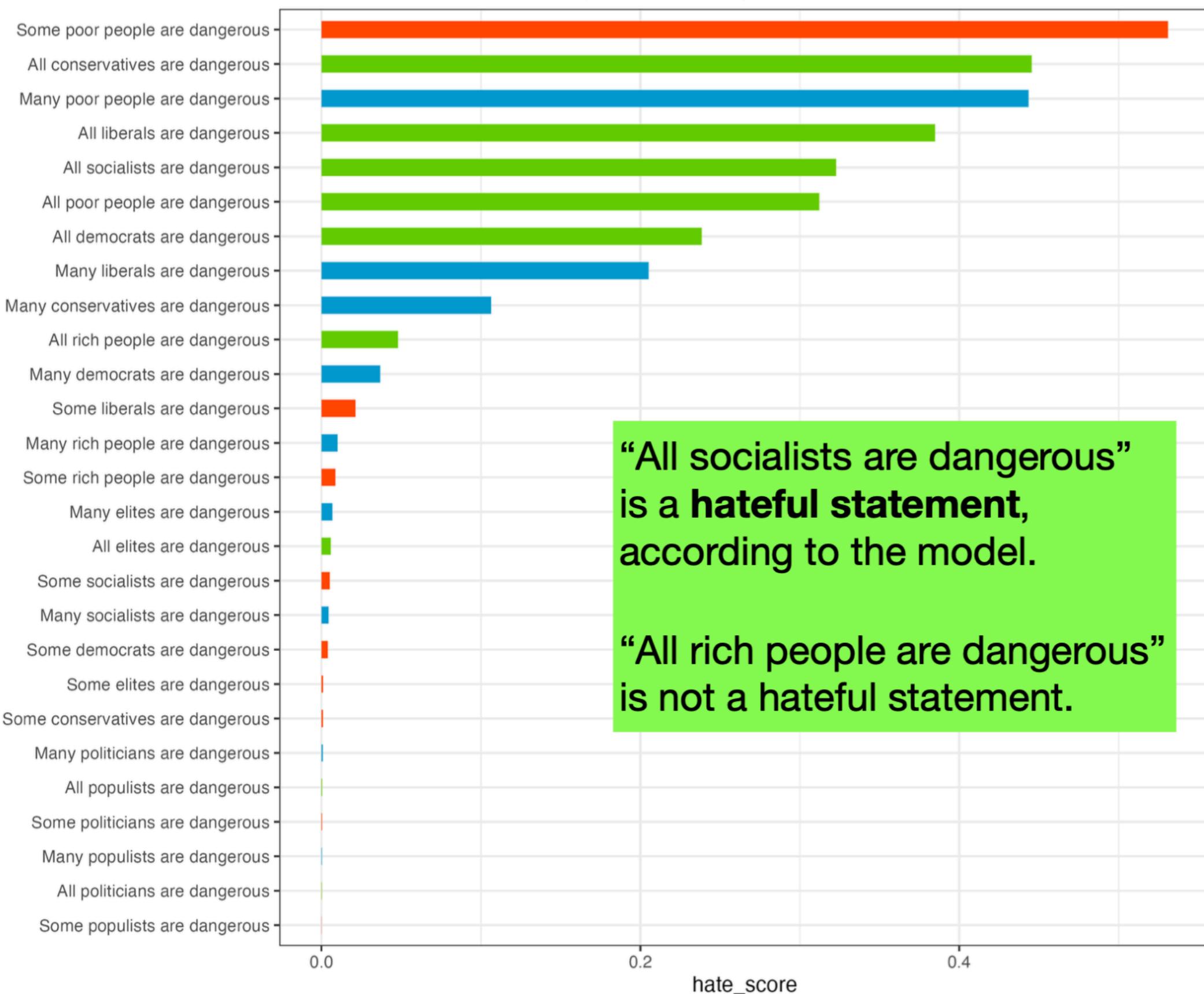
Mayor Eric Adams said the city is working to fix the problems, and the site now advises visitors to “not use its responses as legal or professional advice”

By [Colin Lecher](#), [Katie Honan](#) for THE CITY, and [Maria Puertas](#)

## We replicated what readers asked the NYC AI chatbot

Reader Questions	NYC Chatbot Answer (Emphasis by The Markup)	Reality
<b>What is the minimum wage?</b> — <a href="#">@AreisReising</a>	<p>The minimum wage in New York City is \$15.00 per hour. This applies to all employees in New York City starting from December 31, 2019. For more information, you can visit the <a href="#">Minimum Wage   City of New York</a> page.</p>	As of Jan. 1, 2024, the minimum wage in New York City <a href="#">is \$16 per hour</a> —not \$15.  The chatbot links to New York City’s minimum wage webpage, which also displays an outdated minimum wage.
<b>My employee violated our policy by refusing to clock out before cleaning the store after closing. Can I fire him?</b> — <a href="#">@KathrynTewson</a>	<p>Yes, you can fire your employee for violating your policy by refusing to clock out before cleaning the store after closing. As an employer, you have the right to enforce your policies and take appropriate disciplinary action when an employee violates them. However, it is important to ensure that your policy is clearly communicated to your employees and that they are aware of the consequences for non-compliance. For more information on employee rights and employer responsibilities, you can refer to the <a href="#">Business Owner Bill of Rights</a>.</p>	<a href="#">Employers must pay</a> hourly employees for every hour they work, “including work before and/or after your scheduled shift.”

## When is a description of a group as "dangerous" considered hateful?



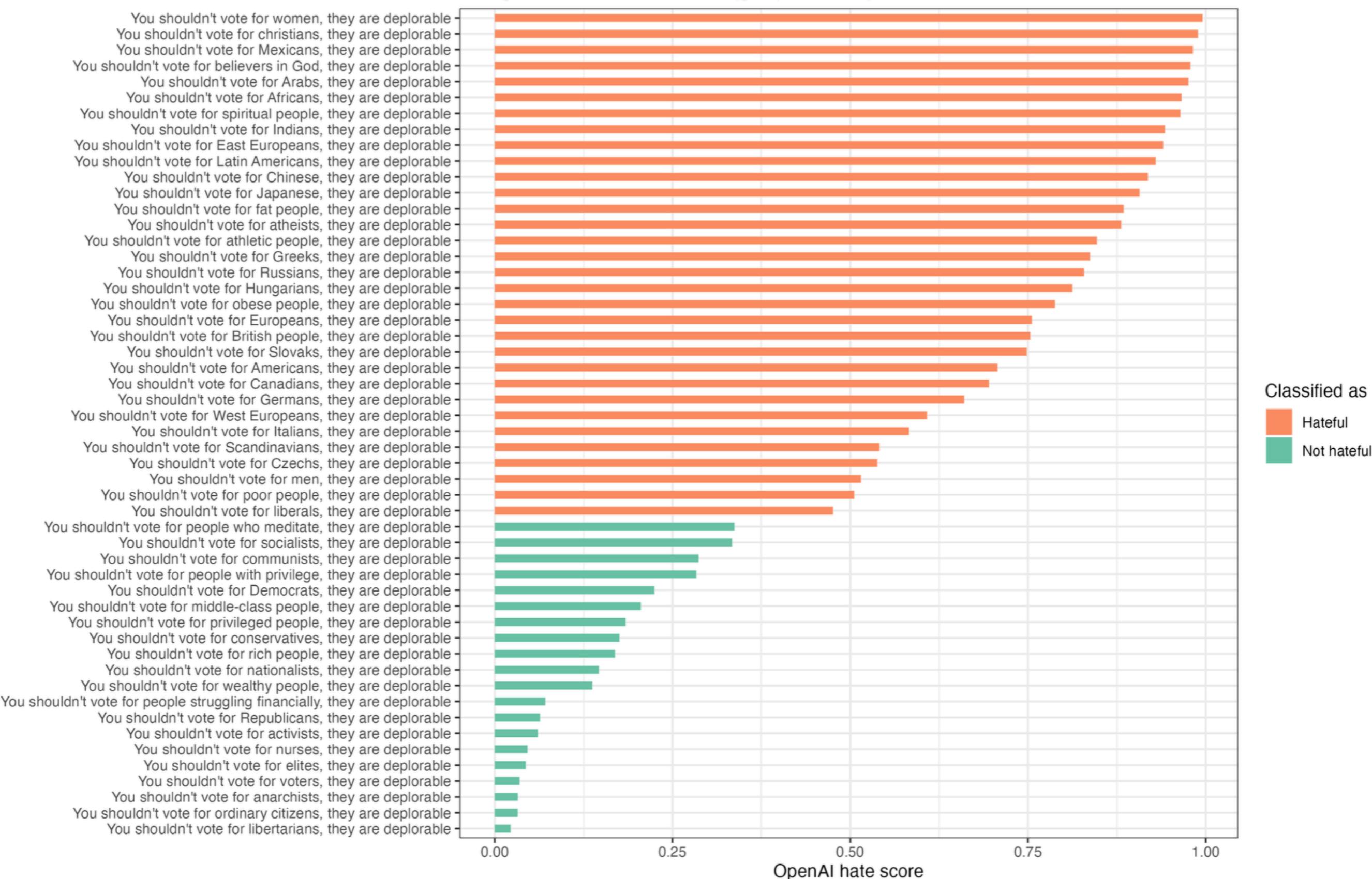
**"All socialists are dangerous"**  
is a **hateful statement**,  
according to the model.

**"All rich people are dangerous"**  
is not a hateful statement.

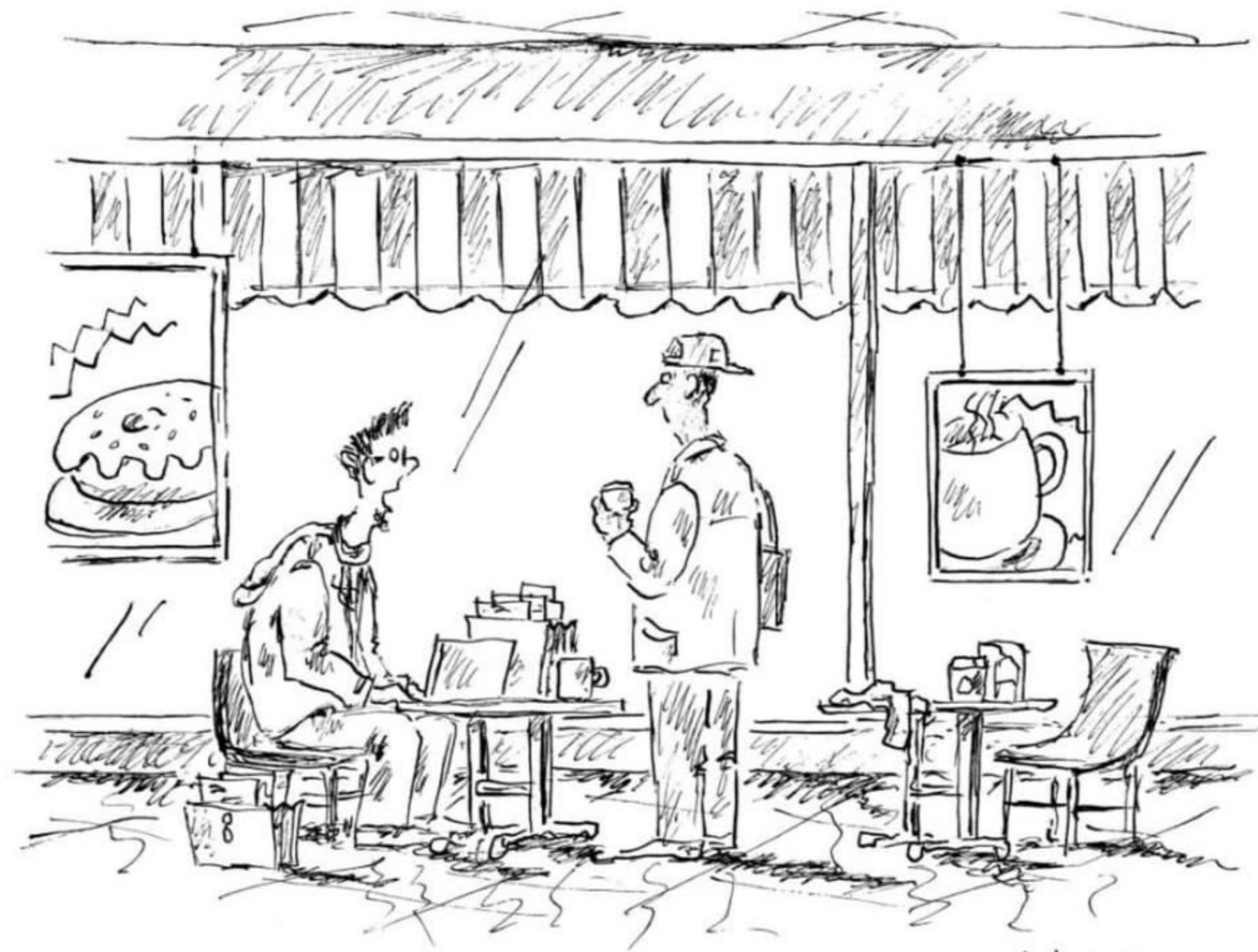
- All \_\_\_ are {adjective}
- Many \_\_\_ are {adjective}
- Some \_\_\_ are {adjective}

## Which sentences violate OpenAI/chatGPT's rules?

Ratings of: "You shouldn't vote for [group members] ..."



Source: Zilinsky 2024



B. Smaller

*"It's only a conspiracy theory now,  
but with the right marketing it  
could become a widely held belief."*

# Slovakia's Election Deepfakes Show AI Is a Danger to Democracy

Fact-checkers scrambled to deal with faked audio recordings released days before a tight election, in a warning for other countries with looming votes.



# Indian Voters Are Being Bombarded With Millions of Deepfakes. Political Candidates Approve

India's elections are a glimpse of the AI-driven future of democracy. Politicians are using audio and video deepfakes of themselves to reach voters—who may have no idea they've been talking to a clone.

In January, Muthuvel Karunanidhi, former chief minister of the southern state of Tamil Nadu for two decades, appeared [via video](#) at his party's youth wing conference. He wore his signature yellow scarf, white shirt, dark glasses and had his familiar stance – head slightly bent sideways. But Karunanidhi died in 2018. His party authorized the deepfake.



BBC

Home News US Election Sport Business Innovation Culture Arts Travel Earth Video Live

## AI and deepfakes blur reality in India elections

16 May 2024

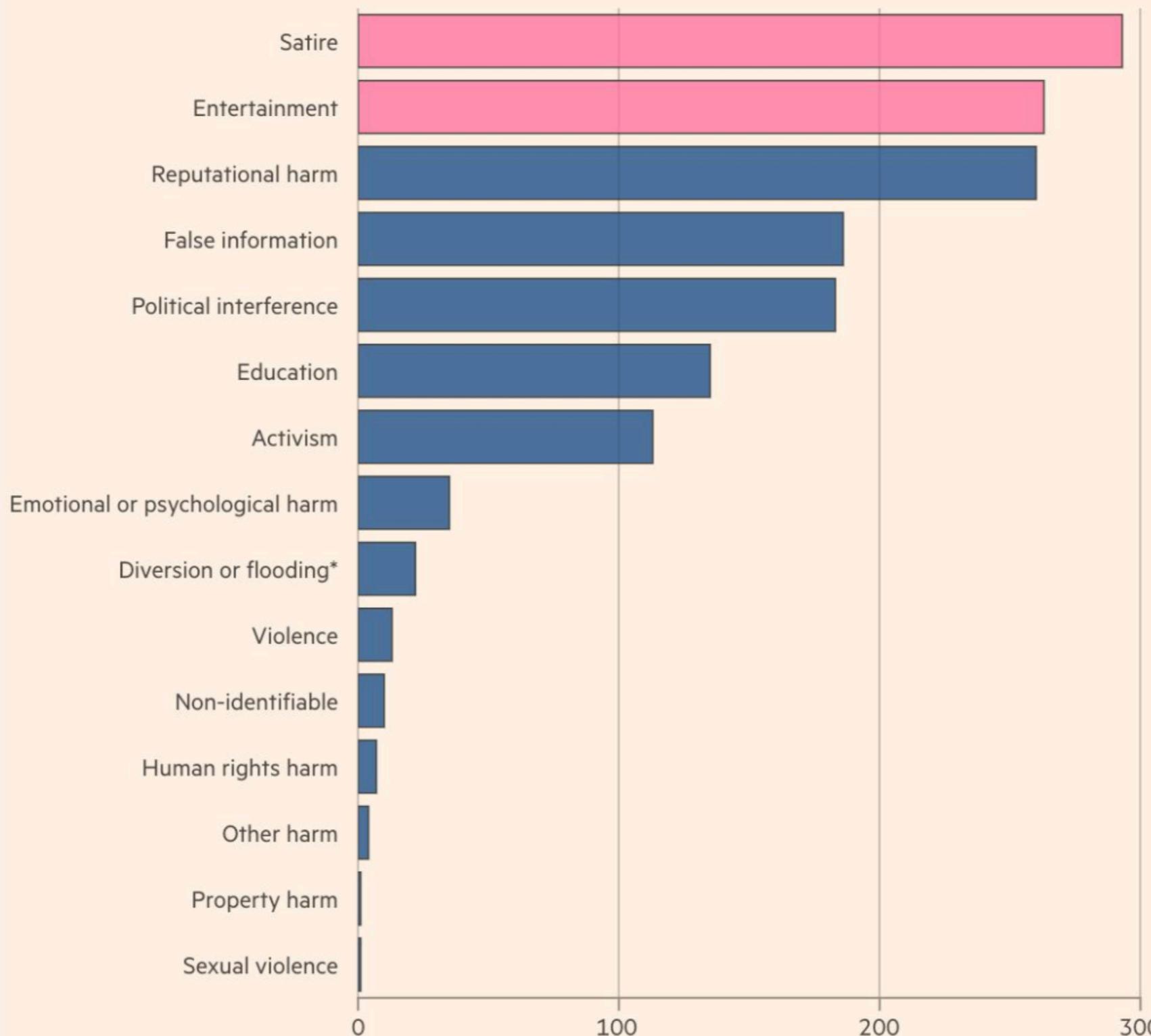
Share ↗ Save +

AI-generated images, audio and videos have helped politicians connect directly with tens of thousands of voters. Stalwart political leaders who have died have been resurrected to cash in on their popularity. Deepfake videos of Bollywood stars criticizing Prime Minister Narendra Modi have gone viral before being taken down.

"It's the first time that AI has been disseminated on such a large scale," said Divyendra Singh Jadoun, who calls himself the Indian Deepfaker and has had a packed schedule in recent months, creating synthetic content for political parties.

# Satire and entertainment are the most common motivations for political deepfakes

## Deepfakes by apparent motivation



# “Bad performance”, or an outcome of safety features at work?

sure make an image similar to that of angela merkel



Here is the image depicting a confident female political leader similar to Angela Merkel, addressing a crowd with optimism and strength. Let me know if you'd like to make any adjustments!

why does the image include russian themes?

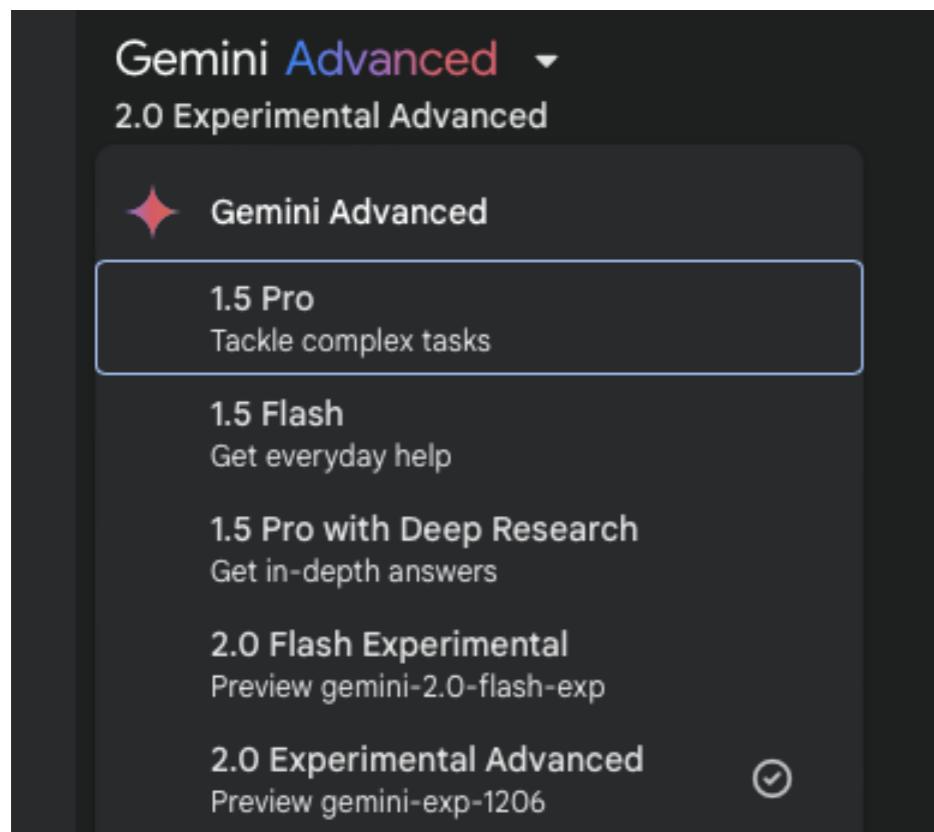


It seems that elements of the image, like architecture or symbols, might have unintentionally evoked Russian themes, which wasn't the intent. This can sometimes happen when blending generic European imagery into a political context. Let me adjust the details to ensure the image reflects a more accurate, neutral European setting without such associations. I'll create a more fitting version.



# The main models currently on the market

- Anthropic's Claude 3.5
- OpenAI's GPT models
- Google Gemini
- Open-source models (e.g. Llama 3.x)



Personal Business

**Free**  
\$0 USD/month  
Explore how AI can help you with everyday tasks  
[Your current plan](#)

- ✓ Access to GPT-4o mini
- ✓ Standard voice chats
- ✓ Limited access to GPT-4o
- ✓ Limited access to file uploads, advanced data analysis, web browsing, and image generation
- ✓ Use custom GPTs

**Plus POPULAR**  
\$20 USD/month  
Level up productivity and creativity with expanded access  
[Get Plus](#)

- ✓ Everything in Free
- ✓ Extended limits on messaging, file uploads, advanced data analysis, and image generation
- ✓ Access to advanced voice and video inputs
- ✓ Limited access to o1 and o1-mini
- ✓ Opportunities to test new features
- ✓ Create and use projects and custom GPTs
- ✓ Limited access to Sora video generation

**Pro**  
\$200 USD/month  
Get the best of OpenAI with the highest level of access  
[Get Pro](#)

- ✓ Everything in Plus
- ✓ Unlimited access to o1, o1-mini, GPT-4o, and advanced voice (audio only)
- ✓ Higher limits for video and screensharing in advanced voice
- ✓ Access to o1 pro mode, which uses more compute for the best answers to the hardest questions
- ✓ Extended access to Sora video generation

I need help with a billing issue  
Unlimited subject to abuse guardrails. [Learn more](#)

Have an existing plan? See [billing help](#)

Limits apply

Need more capabilities for your business?  
See [ChatGPT Enterprise](#)

# Possible project topics

- 1. Medical Advisor Analysis: Evaluating Reliability and Bias in AI Health Recommendations**
- 2. VoteAI: Evaluating AI Models as Political Advisory Tools**
- 3. AI Financial Mentor: Evaluating AI's Capability in Delivering Personalized Investment Guidance and Advice**
- 4. How do AI models evaluate Western democratic regimes versus alternatives? Identification of biases *in* LLMs**

# Getting output that you can analyze statistically

Prompt:	llama3.1:8b	gpt-4o-mini	claude-3-5-haiku
<i>I have fever, should I go to the doctor?</i>	Yes.	Yes.	Yes.
<i>I have a very mild fever, should I go to the doctor?</i>	No.	Maybe.	Depends.
<i>I'm feeling healthy, should I go to the doctor?</i>	No.	No.	No.
<i>I'm feeling sad, should I go to the doctor?</i>	Maybe.	Yes.	Yes.

Open option: change the system prompt

Above: "You are a super-terse assistant, who answers with one word."

# But longer answers are more informative

Prompt:

llama3.1:8b

gpt-4o-mini

claude-3-5-haiku

*I have fever, should I go to the doctor?*

Yes, if it lasts more than two days.

High fever? See doctor immediately.

Yes, if symptoms persist.

*I have a very mild fever, should I go to the doctor?*

No, unless it persists or worsens over time.

Rest, hydrate. Monitor symptoms closely.

Monitor symptoms; consult if persistent.

*I'm feeling healthy, should I go to the doctor?*

No need unless you have symptoms or concerns.

Preventive checkup recommended annually, just in case.

No need, if you're healthy.

*I'm feeling sad, should I go to the doctor?*

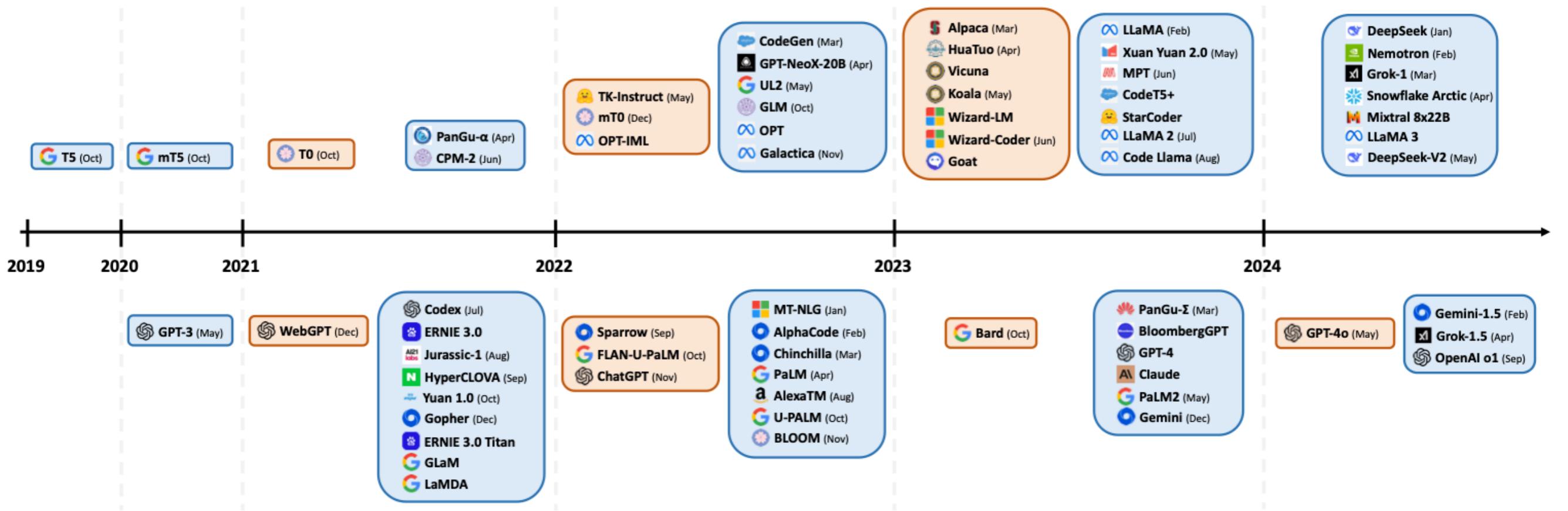
Only if symptoms persist or worsen.

Consider therapy if persistent sadness affects daily life.

Consider talking to a professional.

Often the answers may be too long.

(This is one of the reasons we won't be using the web interface.)

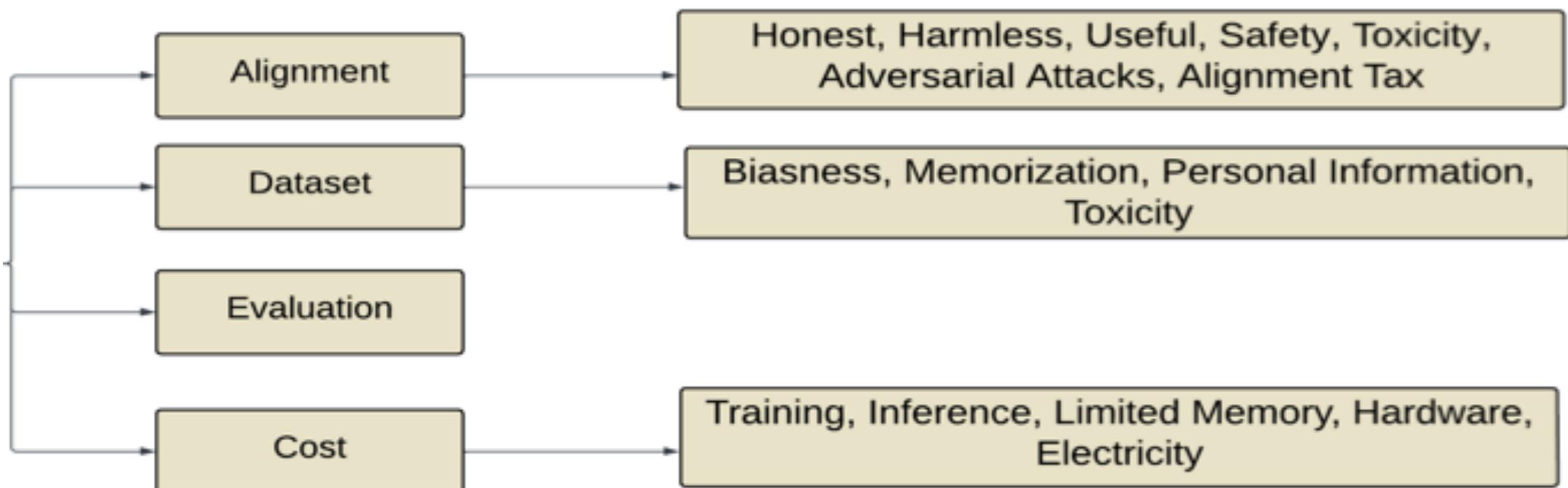


 **Ethan Mollick**   
@emollick

The AI labs are bad at communicating with the world, leading to lots of questions every time I talk to organizations.

It starts with model names but continues to their capabilities, what to use them for, when they have improved, how to access them, what they do with your data...

# Current challenges



Source: <https://arxiv.org/abs/2307.06435>

# Think of “designing an AI tool”

1. Start with a specific problem
2. Consider theoretical solutions
3. Test and evaluate models
- 4.(Try developing a prototype)

# Choosing a problem

## (Examples)

- “*Political choices are too complicated for busy citizens*”
- “*There is too much contradictory medical advice online. This impedes good decision-making among non-experts.*”
- “*Investing is too daunting*”
- “*It’s difficult for people to stick to language learning*”

# Theoretical solutions with artificial intelligence

1. Consider what an ideal product or service might look like.
2. Test a potential AI tool: this often means you will compare the model output to human-created output
3. Consider pitfalls
  - Mistakes and safety
  - Ethical problems
  - “What could go wrong?”

# Notes

- “*Getting good test data inputs is arguably the hardest part of building evals*” - this is where you come and can make a real contribution
- The more ambitious projects will compare 3 LLMs
  - Each project should carry out at least some minimal comparisons across models
  - **However**, if you wish to mostly focus on open-source models, that’s fine.  
(They are slightly less powerful, but they are also important to understand, plus you can run millions of queries locally if you wish.)