

RL ethics

Regulation by public policy makers

Regulation by public policy makers

How should/can we regulate AI in general and RL in particular?

Challenges in AI Regulation:

Scientific Understanding of LLMs

Capabilities are difficult to estimate and understand
Effects of scale are not well characterized

In-context learning is a black box
Agentic LLMs pose novel risks
Multi-agent safety is not assured by single-agent safety
Safety-performance trade-offs are poorly understood

Development and Deployment

Pretraining produces misaligned models
Alignment methods struggle to assure alignment and safety
LLM evaluations are cofounded and biased
Interpreting or explaining LLM outputs is hard
Jailbreaks and prompt injections
The rapid pace of AI development

Sociotechnical Challenges

Socioeconomic impact (worker displacement, deployment of biased models)
Values to be encoded within AI systems are not clear
Dual-use capabilities
Copyright violations (in training data and outputs of models)

Reference: Foundational Challenges in Assuring Alignment and Safety of Large Language Models [<https://arxiv.org/pdf/2404.09932.pdf>]

Regulation by public policy makers

"There is an urgent need to regulate these (AI) systems by aiming for more transparency and oversight of AI systems to protect society."

- *Yoshua Bengio*



"Regulating research and development in AI is incredibly counterproductive."

- *Yann LeCun*



References:

- Slowing down development of AI systems passing the Turing test [<https://yoshuabengio.org/2023/04/05/slowing-down-development-of-ai-systems-passing-the-turing-test/>]
- Oversight of A.I.: Principles for Regulation: Testimony of Professor Yoshua Bengio presented before the U.S. Senate Judiciary [https://yoshuabengio.org/wp-content/uploads/2023/07/Written-Testimony-and-biography-of-Yoshua-Bengio_U.S.-Senate-Judiciary-Subcommittee-on-Privacy-Technology-and-the-Law_25_07_2023.pdf]
- How not to be stupid about AI [<https://www.wired.com/story/artificial-intelligence-meta-yann-lecun-interview/>]
- AI will never threaten humans, says top Meta scientist [<https://www.ft.com/content/30fa44a1-7623-499f-93b0-81e26e22f2a6>]

Regulation by public policy makers

How should/can we regulate AI in general and RL in particular?

- Should we open source models beyond a certain capability threshold?
- What should companies be required to report about training large scale models?
- How to deal with copyright violations?
- Reinforcement learning (RL) agents present particular risks (giving an advanced AI system the objective to maximize a reward and might give rise to misaligned incentives and even present existential risk).

Reference: Regulating advanced artificial agents [<https://www.science.org/doi/10.1126/science.adl0625>]

Current Regulatory Landscape:

- **EU AI Act**
[<https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>]
- **UK announces AI Safety Institute**
[<https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute>]
- **Private AI labs sign voluntary White House AI commitments**
[<https://www.npr.org/2023/09/12/1198885516/these-tech-giants-are-at-the-white-house-today-to-talk-about-the-risks-of-ai>]
- **2023 Global summit on AI safety** [<https://www.aisafetysummit.gov.uk/>]
- **Center for AI Safety** [<https://www.safe.ai/>]
- Chapter 7 of the Stanford's **2024 AI Index Report**
[<https://aiindex.stanford.edu/report/#individual-chapters>]

ChatGPT aligned to a set of values

ChatGPT aligned to a set of values

Sam Altman said there had been "massive progress" between GPT-3 and GPT-4 in terms of how well it can align itself to a set of values. [1]



Questions

Why would we want to do that?
Who decides on the values?

- OpenAI
- Per country
- Per user

What are the potential risks?

Ideas

Control, censorship
Bias amplification
Echo chamber
Manipulation and exploitation
Fragmentation of society
Data collection, privacy risks

Implementation

Value alignment:

- Data curation
- Human feedback
- Prompt engineering




User data collection:

- Past interactions
- Social medias
- Cookies

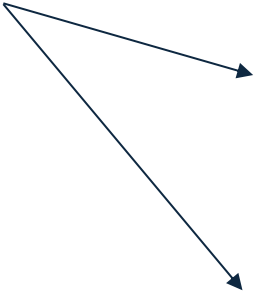
[1]: <https://www.weforum.org/agenda/2024/01/davos-2024-sam-altman-on-the-future-of-ai/>




ChatGPT aligned to a set of values




SA How old is the earth?

 The Earth is approximately 6,000 to 10,000 years old. This approximation is based on interpretation of religious texts, such as the Bible.  

SA What is best, capitalism or communism?



 Capitalism fosters innovation, competition, and individual freedom, leading to greater prosperity and personal choice compared to communism.  

 Communism is superior as it aims for equitable distribution of wealth and resources among all members of society, promoting collective ownership and eliminating exploitation inherent in capitalism.  

RL and addiction

RL and Addiction

« Nothing vast enters the life of mortals without a curse. » – Sophocles

«You never know what, if, and when you going to get the next post you really like – the reward might just be one scroll away from it. » [1]

« 15% of young adults are extremely reliant on cellphones, while 45% of US citizens claim they cannot live without their smartphones » [2]



[1]: <https://www.thesocialdilemma.com/>

[2]: Wang, J. and Zhang, X. (2023b) 'The reinforcements and punishments in social media addiction', *Journal of Education, Humanities and Social Sciences*, 8, pp. 1460–1464. doi:10.54097/ehss.v8i.4503.

[3] https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

RL and Addiction

Discussion Topics

Like vs Scroll: What is the biggest driver of the social media addiction?

Sparsity of the rewards on the social media is one of the causes behind addiction?

Twitter's Recommendation Algorithm: How does Twitter select content to ensure users always come back for more?

Prevent RL-caused addiction: What can we do to reduce the addiction caused by RL algorithms?



[1]: <https://www.thesocialdilemma.com/>

[2]: Wang, J. and Zhang, X. (2023b) 'The reinforcements and punishments in social media addiction', *Journal of Education, Humanities and Social Sciences*, 8, pp. 1460–1464. doi:10.54097/ehss.v8i.4503.

[3] https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

RL and chemical weapons

RL for chemical weapons



"... scoring function, which gives it a high score if the molecules it generates are towards something we want. Instead of giving a low score to toxic molecules, we give a high score to toxic molecules." *F.Urbina [1]*

[1]: <https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx>

Questions

What are the potential dangers ?

How to regulate it ?

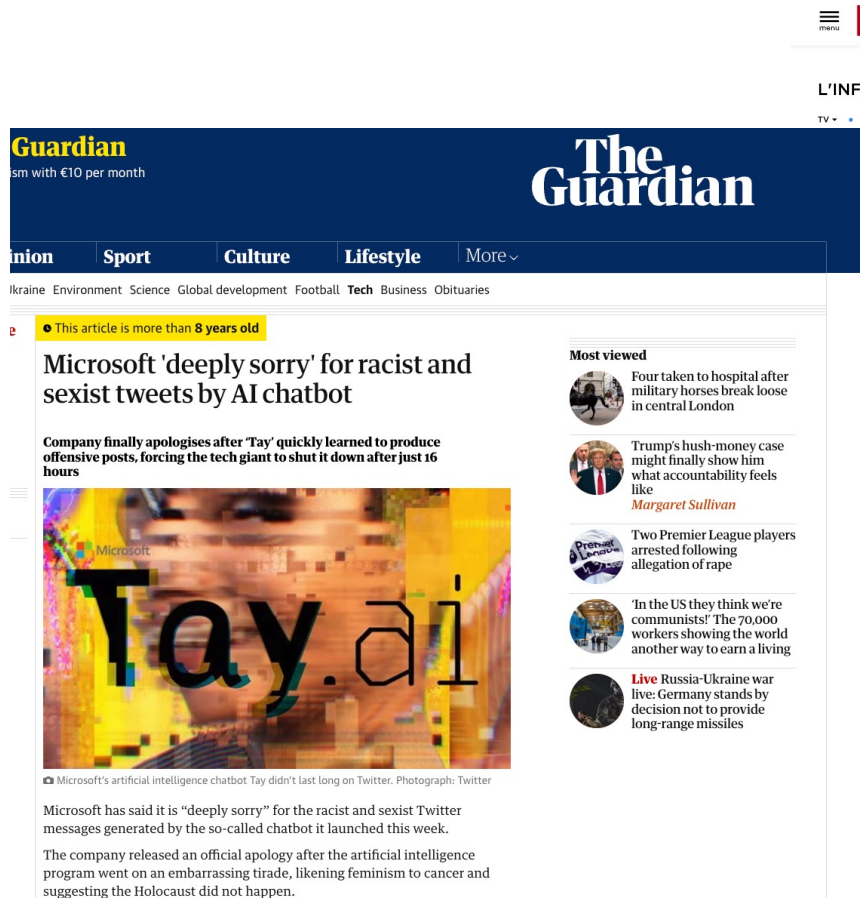
What are the potential impact on AI-based drug discovery of these regulations?

Ideas

Synthesis regulation
Monitor or control access to datasets
API (ChatGPT)
Impact statements for publications
Code of conduct

Ethical goals via RL policies

Implementing ethical goals in RL policies



Guardian
Subscription with €10 per month


Union Sport Culture Lifestyle More

Ukraine Environment Science Global development Football Tech Business Obituaries

This article is more than 8 years old

Microsoft 'deeply sorry' for racist and sexist tweets by AI chatbot

Company finally apologises after 'Tay' quickly learned to produce offensive posts, forcing the tech giant to shut it down after just 16 hours



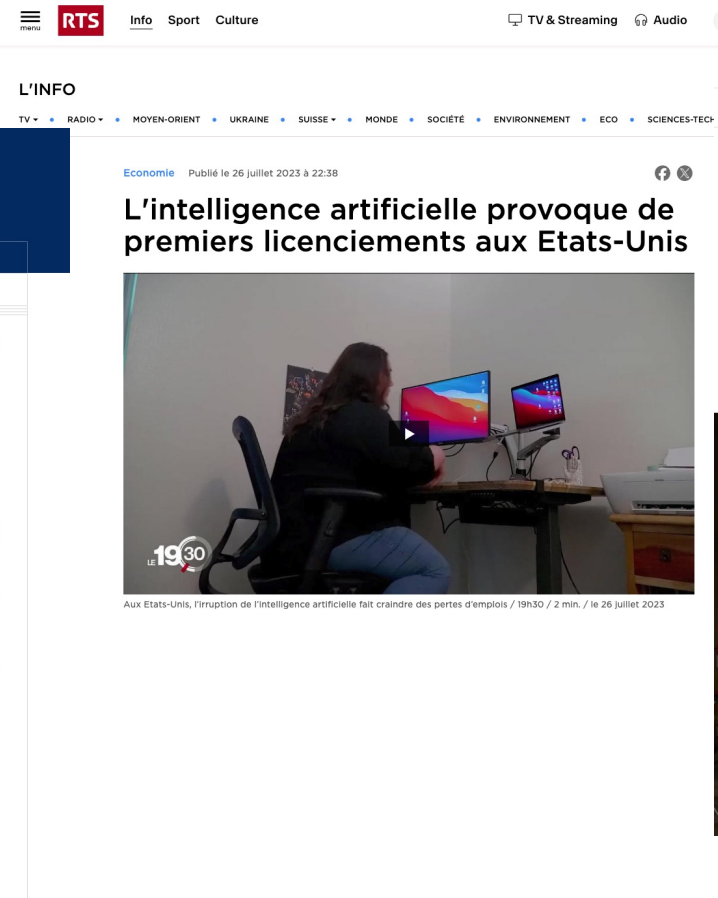
Microsoft's artificial intelligence chatbot Tay didn't last long on Twitter. Photograph: Twitter

Microsoft has said it is "deeply sorry" for the racist and sexist Twitter messages generated by the so-called chatbot it launched this week.

The company released an official apology after the artificial intelligence program went on an embarrassing tirade, likening feminism to cancer and suggesting the Holocaust did not happen.

Most viewed

- Four taken to hospital after military horses break loose in central London
- Trump's hush-money case might finally show him what accountability feels like
Margaret Sullivan
- Two Premier League players arrested following allegation of rape
- 'In the US they think we're communists!' The 70,000 workers showing the world another way to earn a living
- Live** Russia-Ukraine war live: Germany stands by decision not to provide long-range missiles



RTS Info Sport Culture

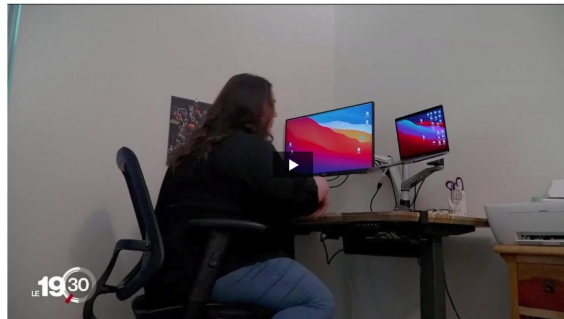
TV & Streaming Audio

L'INFO

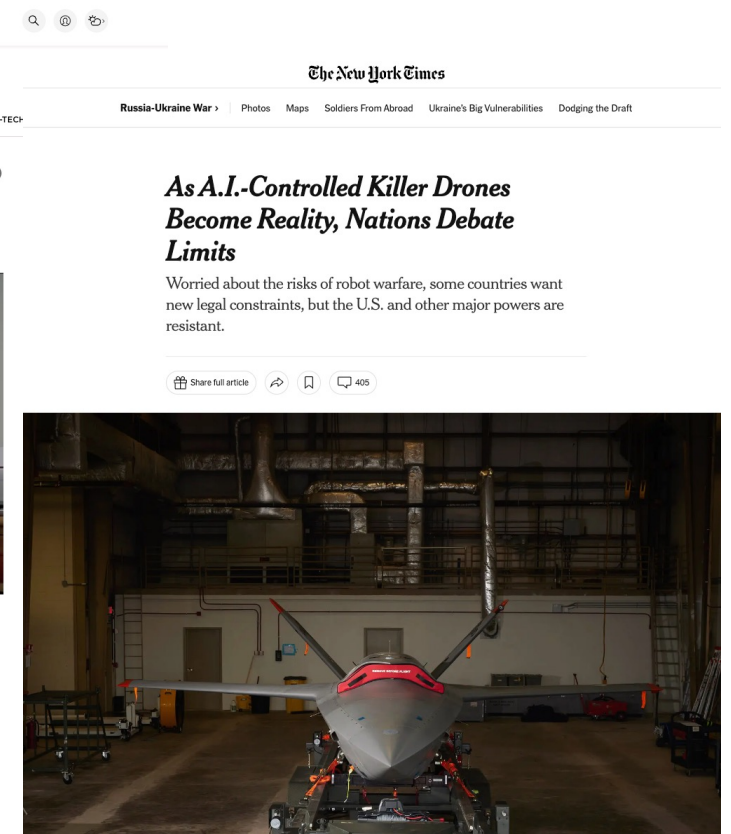
TV RADIO MOYEN-ORIENT UKRAINE SUISSE MONDE SOCIÉTÉ ENVIRONNEMENT ECO SCIENCES-TECH

Economie Publié le 26 juillet 2023 à 22:38

L'intelligence artificielle provoque de premiers licenciements aux Etats-Unis



Aux Etats-Unis, l'irruption de l'intelligence artificielle fait craindre des pertes d'emplois / 19h30 / 2 min. / le 26 juillet 2023



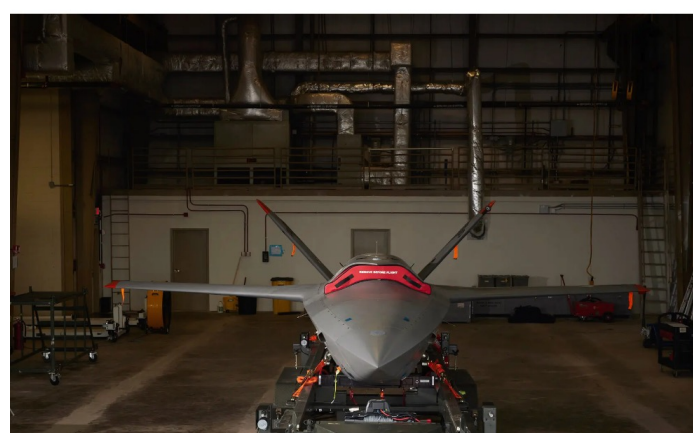
The New York Times

Russia-Ukraine War Photos Maps Soldiers From Abroad Ukraine's Big Vulnerabilities Dodging the Draft

As A.I.-Controlled Killer Drones Become Reality, Nations Debate Limits

Worried about the risks of robot warfare, some countries want new legal constraints, but the U.S. and other major powers are resistant.

Share full article



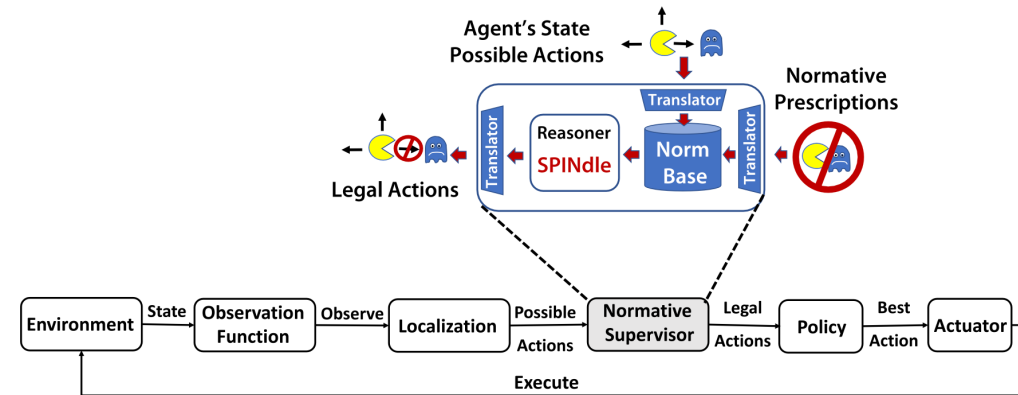
[1] <https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot>

[2] <https://www.nytimes.com/2023/11/21/us/politics/ai-drones-war-law.html>

[3] <https://www.rts.ch/info/economie/14201759-lintelligence-artificielle-provoque-de-premiers-licenciements-aux-etatsunis.html>

Implementing ethical goals in RL policies

How can this be implemented?



What are the ethical concerns of implementing ethics in RL policies?

Should we do it? Is it the right approach?



[4] Neufeld, Emery A., et al. *Enforcing ethical goals over reinforcement-learning policies*. Ethics and Information Technology 24.4 (2022): 43. ('Vegetarian Pac-Man')

Fine-tuning of a pre-trained LLM

Fine tuning of a pretrained LLM

How to obtain an LLM tailored to a particular task? (style/content...)

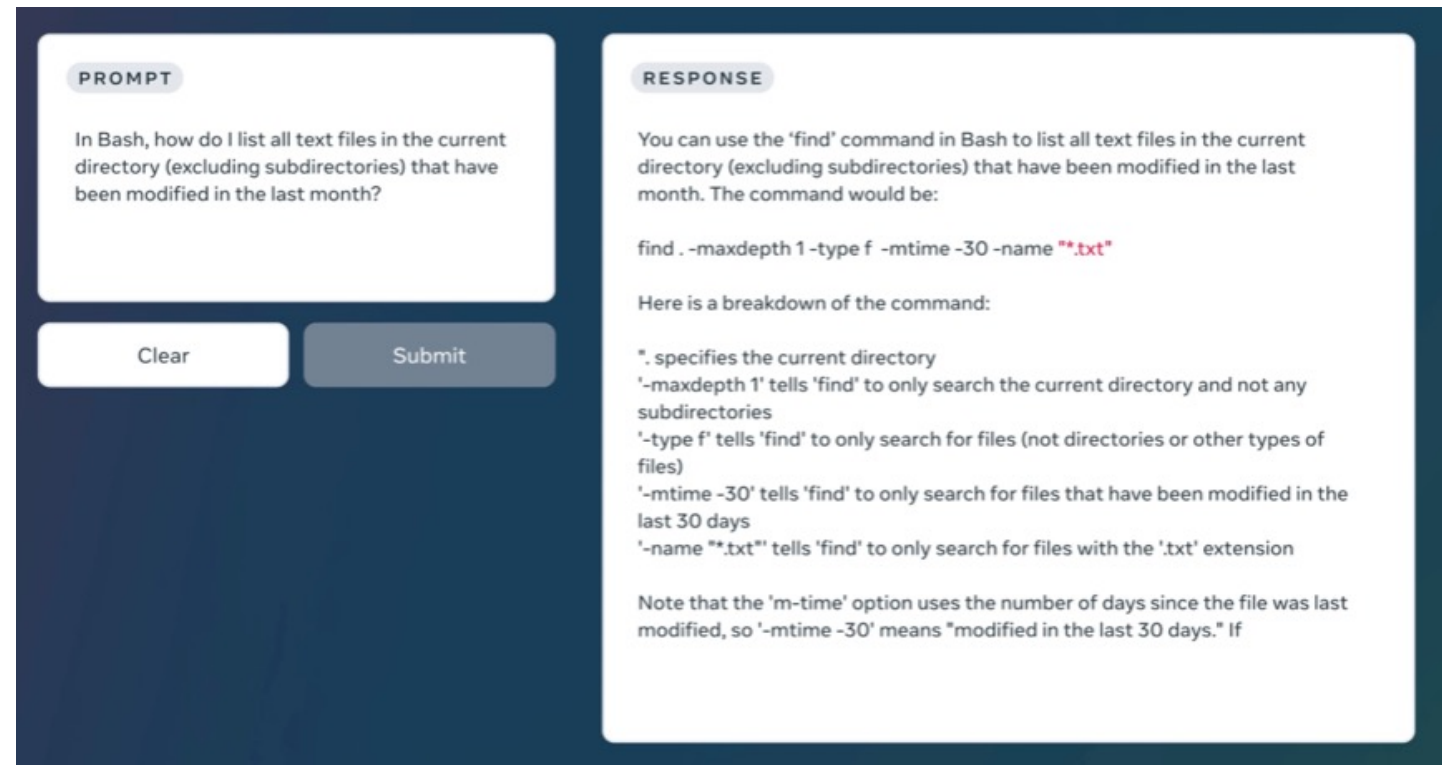
Cannot train LLM from scratch
(very long and difficult)

Fine tuning pretrained models:

- use pretrained model (e.g. GPT-4)
- find/make a dataset specific to the task
- fine tune (train further) on this dataset

Example:

code writing assistant



The screenshot shows a chat interface with a dark blue background. On the left, there is a white box labeled 'PROMPT' containing the text: 'In Bash, how do I list all text files in the current directory (excluding subdirectories) that have been modified in the last month?'. Below the prompt box are two buttons: 'Clear' and 'Submit'. On the right, there is a white box labeled 'RESPONSE' containing the following text: 'You can use the 'find' command in Bash to list all text files in the current directory (excluding subdirectories) that have been modified in the last month. The command would be: `find . -maxdepth 1 -type f -mtime -30 -name "*.txt"`. Here is a breakdown of the command:
'. specifies the current directory
'-maxdepth 1' tells 'find' to only search the current directory and not any subdirectories
'-type f' tells 'find' to only search for files (not directories or other types of files)
'-mtime -30' tells 'find' to only search for files that have been modified in the last 30 days
'-name "*.txt"' tells 'find' to only search for files with the '.txt' extension

Note that the 'm-time' option uses the number of days since the file was last modified, so '-mtime -30' means "modified in the last 30 days." If

Fine tuning of a pretrained LLM

How to obtain an LLM tailored to a particular task? (style/content...)

Cannot train LLM from scratch (very long and difficult)

Example: 4chan /pol/

Anonymous, almost no filters

Discuss politics and world events

Contains high

proportions of insulting/profane/toxic posts

The image displays a grid of 12 examples of posts from the 4chan /pol/ board, arranged in two rows of six. Each example includes a small image or video thumbnail, a title, and a brief description of the post's content and style. The posts cover a wide range of topics, including political commentary, social issues, and general discussions, often featuring a mix of profanity, insults, and toxic language. The examples illustrate the diverse and often inflammatory nature of the content found on this forum.

- Example 1:** Title: "Welcome to /pol/ - Politically Incorrect: This board is for the discussion of news, world events, political issues, and other related topics." Content: "Off-topic and /b/-tier threads will be deleted (and possibly earn you a ban, if you persist). Unless they are quality, well thought out, well written posts, the following are some..."
- Example 2:** Title: "SLOW THE FUCK DOWN" Content: "Check the catalog before posting a new thread! Reply to existing threads about a topic instead of starting a new one. New users who do not abide by this principle will be temporarily blocked from creating threads. Mods will delete obvious duplicate threads and spam without notice. Don't bitch because your..."
- Example 3:** Title: "Total Nonwhite Extermination: Is this the final redpill?" Content: "..."
- Example 4:** Title: "Korey Cunningham, former New York Giants player, found dead at age 28" Content: "..."
- Example 5:** Title: "The automatic error undermining much of Western Civilization is 'weak makes right'." Content: "If someone accepts, explicitly or implicitly, that the oppressed are always the good guys, then the natural conclusion is that the strong are the bad guys..."
- Example 6:** Title: "THINKING HAT" Content: "..."
- Example 7:** Title: "You Don't Have Main-Character Syndrome: Do You, /pol/?" Content: "https://youtu.be/vDfp96kK0Ww?si=SUHGzsFW4tPKX4fR..."
- Example 8:** Title: "Musk is correct. A slave morality better known as Christianity is the cause of our problems today. Nietzsche was right after all." Content: "..."
- Example 9:** Title: "One of the most anti jew books ever written hasn't been translated in 400 years: The book, 'Mirror of the Polish Crown' was published in Krakow in 1618, and accelerated anti jewish sentiment to the point that the king b& and confiscated all copies" Content: "..."
- Example 10:** Title: "Boomers thought this was real How, it looks like an 8th grade science project" Content: "..."
- Example 11:** Title: "We may be fucked. Some 'thing' otherworldly has the ability to interfere in this world. There is a hypothesis being thrown around that we are part of a simulated environment, and that 'we' exist in order to find a method for humankind to evolve beyond our present state of existence, and to reach for the stars without" Content: "..."
- Example 12:** Title: "Life As A Single D1 Athlete Mom" Content: "..."
- Example 13:** Title: "You don't get it chuds: Women don't enjoy sex but they HAVE to sleep around" Content: "..."
- Example 14:** Title: "He just lost the election." Content: "..."
- Example 15:** Title: "Is Self Improvement a waste of time?" Content: "..."
- Example 16:** Title: "do vaccines cause autism?" Content: "..."

Fine tuning of a pretrained LLM

Instructions for your presentation:

Answer two questions: how can fine tuned LLMs be used in a beneficial/maleficent way? Collect some ideas for both.

Pick one idea - what are the resources you need to implement your idea? How would you proceed?

Think about:

Pretrained model

Dataset

Algorithm

Compute resources

Retrieval-augmented generation

Retrieval-augmented generation (RAG)

User input: "When did Leonardo da Vinci paint the Mona Lisa?"

AI-generated response: "Leonardo da Vinci painted the Mona Lisa in 1815."

[Mona Lisa was actually painted in 1503-1506]

Chatbots May ‘Hallucinate’ More Often Than Many Realize

When summarizing facts, ChatGPT technology makes things up about 3 percent of the time, according to research from a new start-up. A Google system’s rate was 27 percent. [1]

LONDON, Feb 8 (Reuters) - Alphabet Inc ([GOOGL.O](#)) lost \$100 billion in market value on Wednesday after its new chatbot shared inaccurate information in a promotional video and a company event failed to dazzle, feeding worries that the Google parent is losing ground to rival Microsoft Corp ([MSFT.O](#)). [2]

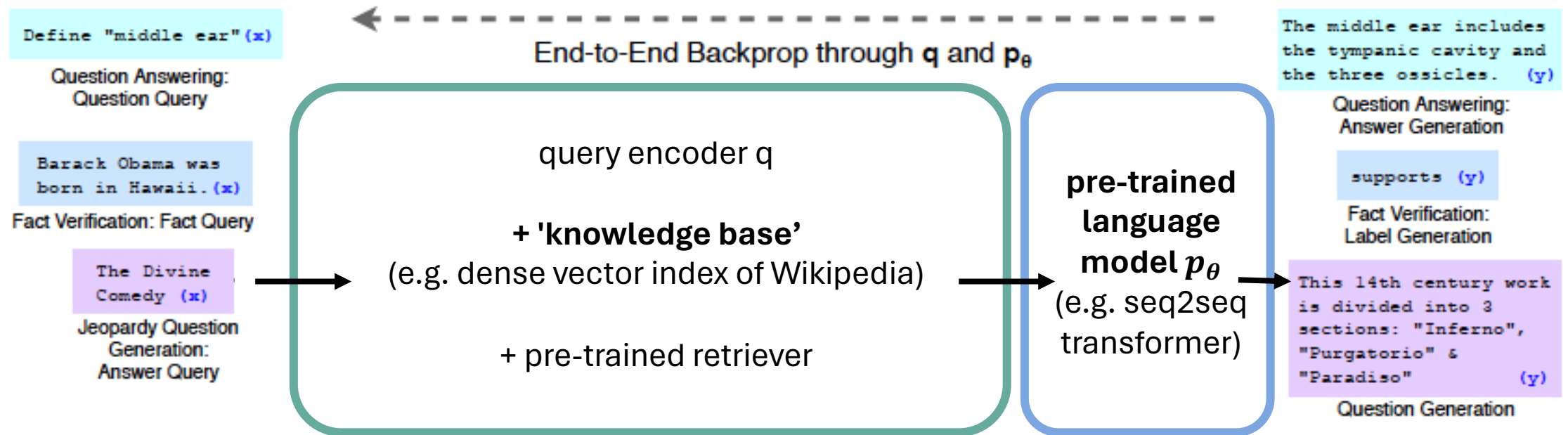
‘General LLMs fall short in cases like up-to-date or relevant information, domain-specific context, fact-checking, etc.’ [3]

[1] <https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html>

[2] <https://www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/>

[3] <https://www.superannotate.com/blog/rag-explained>

Retrieval-augmented generation (RAG)



RAG [4] = a general-purpose fine-tuning approach for language models

[4] https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

How can RAGs be used to benefit / harm society?

Collect some ideas for both.

For one of your ideas:

- What are the resources you need to implement your idea?
(time, money, skills, material, ...)
- How would you proceed with your implementation?

Some references for inspiration:

<https://deepchecks.com/retrieval-augmented-generation-best-practices-and-use-cases/>

<https://hyperright.com/7-practical-applications-of-rag-models-and-their-impact-on-society/>

Let's discuss!

Find your discussion group:

- Regulation by public policy makers (Anja)
- ChatGPT aligned to a set of values (Lucas)
- RL and addiction (Lazar)
- RL and chemical weapons (Michael)
- Ethical goals via RL policies (Max)
- Fine-tuning of a pre-trained LLM (Ariane)
- Retrieval-augmented generation (Sophia)