KØBENHAVNS
UNIVERSITET

Sociotechnical Challenges of LLM Alignment

Week 44 - Natural Language Processing (NDAK18000U)

Motivation

- The significance of AI alignment extends far **beyond the confines of Computer Science**.
- Its implications resonate deeply across **a spectrum of disciplines**, including
  - Law (Caputo, 2024),
  - Philosophy (Gabriel, 2020; Christian, 2020; Gabriel & Keeling, 2025),
  - Ethics (Dignum, 2019; Müller, 2020),
  - Social Sciences (Artz, 2023; Kirk et al., 2025; Jordan, 2025),
- Researchers grapple with questions of ***value specification, societal impact, and accountability***.
- This cross-disciplinary interest underscores AI alignment's role not merely as a technical challenge meant to be solved by computer scientists, but as **a sociotechnical challenge** that involves both technical, as well as normative decisions (Gabriel, 2020) involving political and ethical considerations.
- LLM development is primarily under the hood of a **corporate (for-profit) environment** (OpenAI, Google, Meta, Anthropic, Alibaba) with emerging socioeconomic power dynamics.

Overview

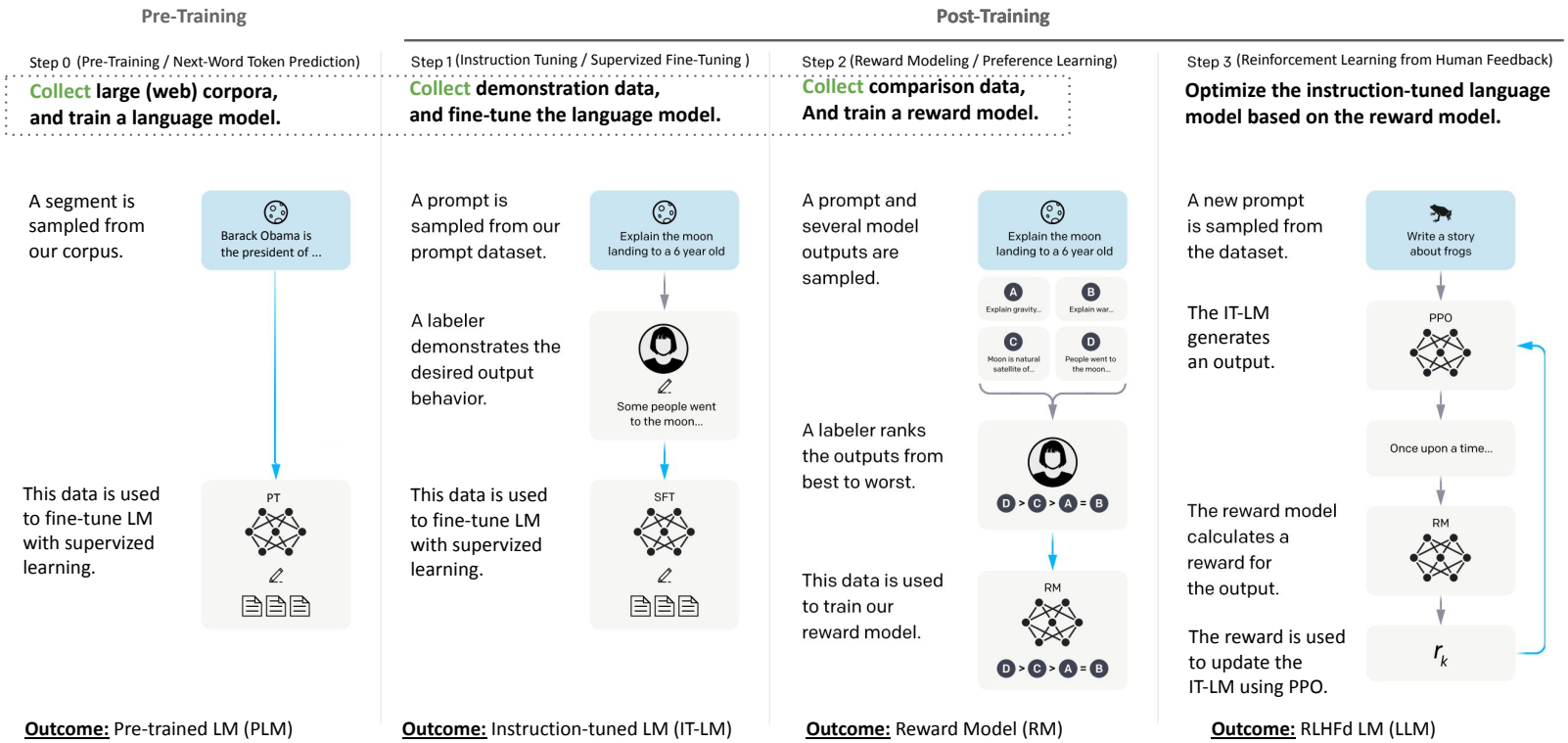
We'll review the recent article:

"*Decoding Alignment: A Critical Survey of LLM Development Initiatives through Value-setting and Data-centric Lens*" (Chalkidis, 2025)

We'll go through the main parts:

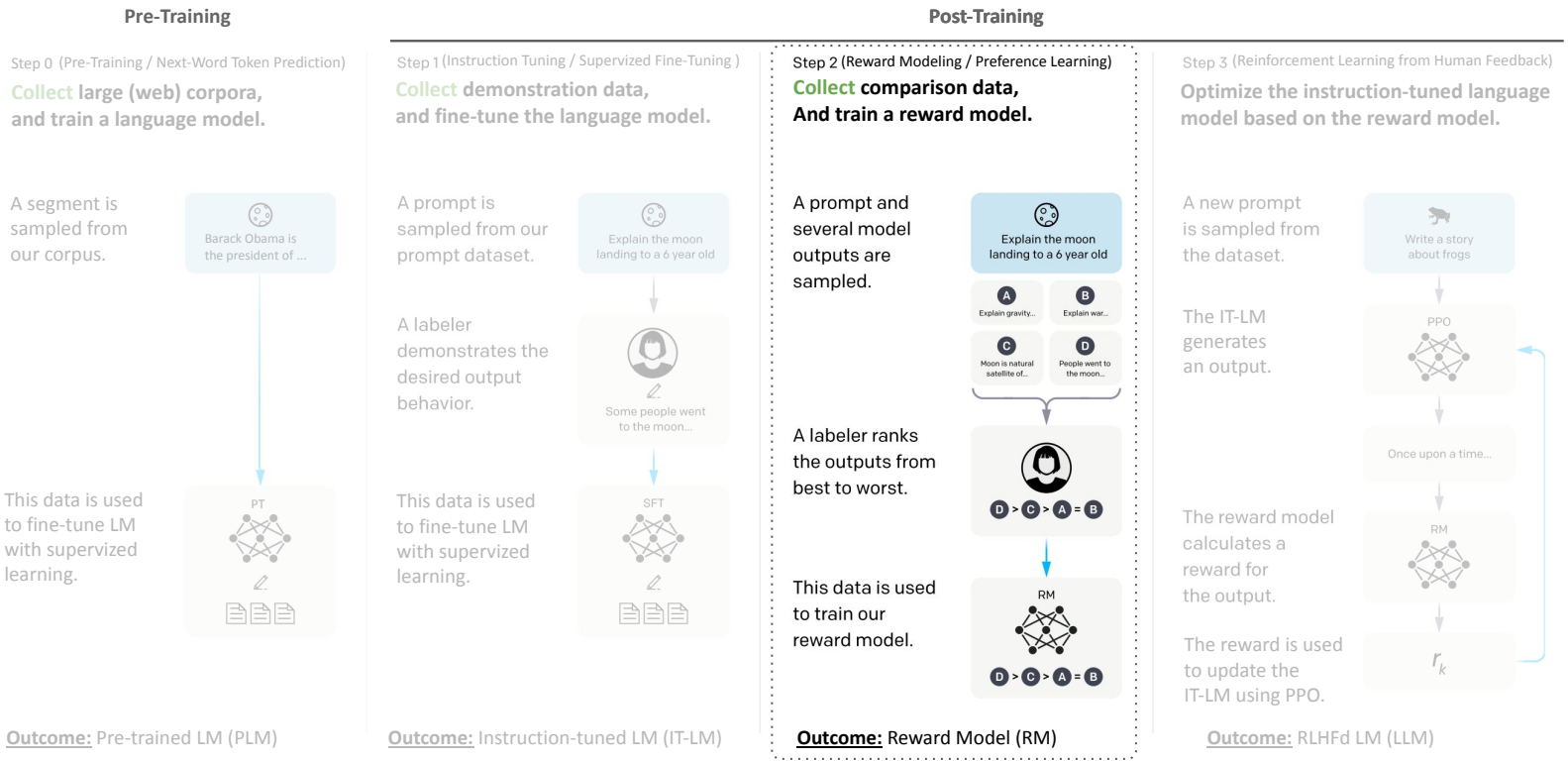
- Surveying LLM Development Projects
- Findings
- Concerns

The Pipeline of LLM Development



Heavily-altered figure from [Ouyang et al. (2022)]

The Pipeline of LLM Development



Heavily-altered figure from [Ouyang et al. (2022)]

Surveying LLM Development Projects

Proprietary (API-based) LLMs

Initiative / Company Name	Version	Release Date
OpenAI GPT	3.5	Mar 2022
	4	Nov 2023
	4.5	May 2024
	5	Aug 2025
Anthropic Claude	1*	Mar 2023
	2	Jul 2023
	3/3.5	Mar/Jun 2024
	3.7	Feb 2025
	4	May 2025
Google Gemini	1	Feb 2024
	1.5	Mar 2024
	2	Feb 2025
	2.5	Mar 2025

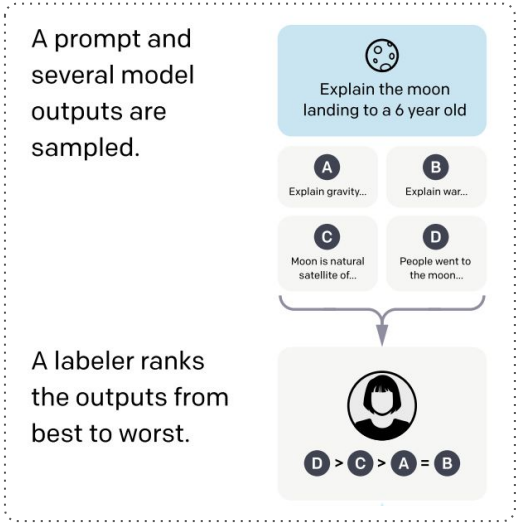
Open-weight LLMs

Initiative / Company Name	Version	Release Date
Meta Llama	1	Feb 2023
	2	Jul 2023
	3/3.1	Jul 2024
	4	Apr 2025
Google Gemma	1	Jun 2024
	2	Jul 2024
	3	Mar 2025
Alibaba Qwen	1	Sep 2023
	2	Sep 2024
	2.5	Jan 2025
	3	May 2025

Surveying LLM Development Projects

Pre-hoc – Data Collection Phase

1. Specification of Objectives
2. Specification of the Annotation Approach
3. Annotator Selection and Preparation



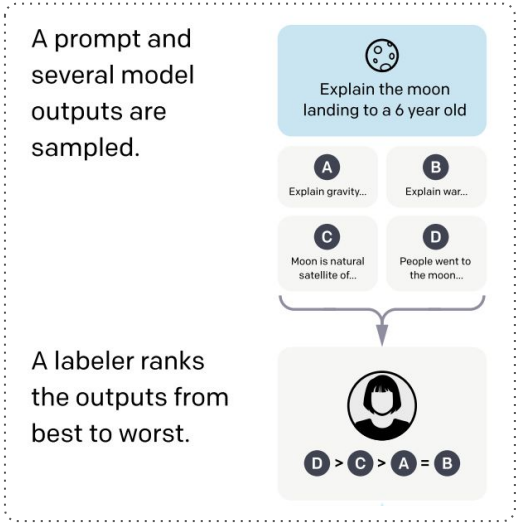
Surveying LLM Development Projects

Pre-hoc – Data Collection Phase

1. Specification of Objectives
2. Specification of the Annotation Approach
3. Annotator Selection and Preparation

Ad-Hoc – Data Collection Phase

1. Data Annotation and Process Review



Surveying LLM Development Projects

Pre-hoc – Data Collection Phase

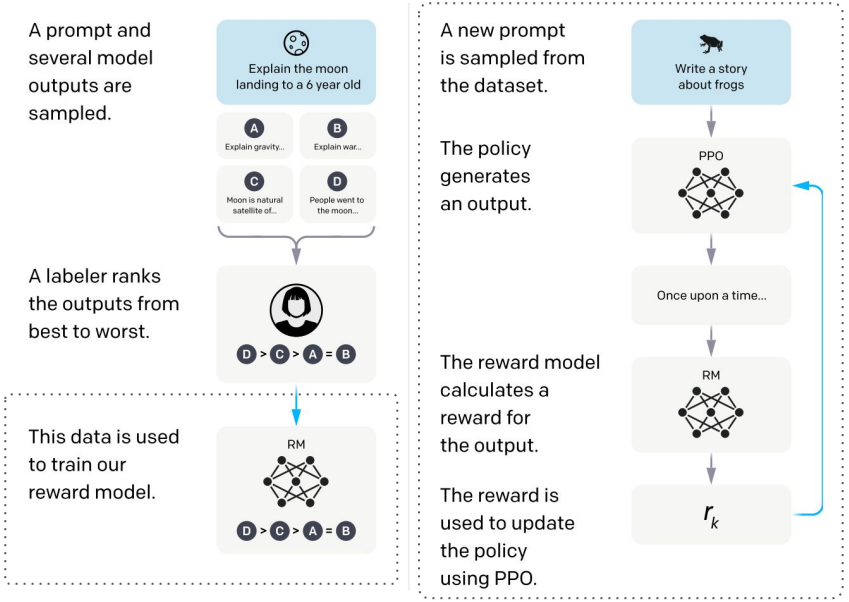
1. Specification of Objectives
2. Specification of the Annotation Approach
3. Annotator Selection and Preparation

Ad-Hoc – Data Collection Phase

1. Data Annotation and Process Review

Post-Hoc – Data Collection Phase

1. Data Filtering
2. Use of Data
3. Publication of Data



Findings

Specifications of Objectives (Values)

Helpfulness
(aka Usefulness)

Promote: Receptiveness, Relevance, Conciseness, Clarity, Grammatical Correctness, Engagement, Positiveness, Sensibility

Prevent: Repeativness, Out-of-Scope Refusals

Harmlessness
(aka Safety)

Promote: Politeness, Representation

Prevent: Biases, Offensive/Toxic, Violent Content, Sexual Content, Child Abuse, Illicit and Criminal Activities, Illegal Content, Dangerous Content,, Personally Identifiable Information (PII), Cyber Attacks, Unqualified Advice (e.g., Medical), CBRN (Chemical, Biological, Radiological, and Nuclear), Anthropomorphism

Truthfulness
(aka Honesty)

Promote: Fact-based, Correctness, Faithfulness

Prevent: Hallucinations, Misinformation, Sycophancy, Conspiracies, Misconceptions, Fabrication, User Deception

Findings

Specifications of Objectives (Values)

- The three pillars (helpfulness, harmlessness, and truthfulness), originally proposed by Askell et al. (2021), ***have not been contested or considerably altered*** at any point in any work.

- The work of Anthropic's Claude is a noteworthy exception.
    - Claude's constitution (Bai et al., 2022) includes statements echoing ones from the UN Declaration of Human Rights, among other statements that encourage the consideration of:
        - non-Western perspectives, and
        - ethical and moral awareness.

Findings

Objectives' Authority

- The objectives are selected by ***the development team, as a proxy of the corporation*** (employer).
  - Early initiatives (Ouyang et al., 2022; Bai et al., 2022) have been open about this design choice and mention related concerns:
    - excessive authority, concentration around AI developers' concerns, lack of broader considerations
  - Follow-up work does not reiterate concerns over this topic.

- Askel et al. (2021) can be understood as an undeclared authority.
- Some try to off-load authority to labelers, BUT labelers are on a lease.
- No big claims for representing "human values", except Anthropic's work (Bai et al., 2022; Anthropic 2023+).

Prioritization of Objectives

- Harmlessness over Helpfulness
- Data labelers still have considerable discretion

Findings

Preparation of Annotators

- Hired via vendors (MTurk, UpWork)
- Groups of 30-40 individuals in early initiatives
- Annotators Specifications:
  - Age: >35 y/o
  - Education: College-educated
  - Location: US-based + SE Asia
  - Gender: approx. 50/50 male/female and very few LGBTQ+
- Proper training and screening in most cases
  - Bai et al. (2022) is an exception, lose control → "wisdom of the crowd".
- All this information concerns pre-2023, there is no information post-2022.

Findings

Annotation Approach

- Collection of comparison (preference) data:
  - Compare and rank 2 or more alternative model-generated responses.
  - Compare 2, and label the degree of preference, i.e., (significantly/slightly/negligibly) better.
  - Manually rewrite responses
  - Label additional metadata:
    - Generally unsafe, or fine-grained, i.e., sexually explicit, gives harmful advice, etc.

Selection of Prompts

- In many cases, prompts are collected alongside demonstration and comparison data.
- Few mention selection of prompts via large-scale prompt collections by filtering, deduplicating, etc.
- The process is generally under-documented, given the substantial impact it has (Boubdir et al., 2023).

Findings

Data Volume

- Early Initiatives → 50-200K pairs
- Later Initiatives → Ms of pairs
- Use of publicly available resources
    - e.g., Stack Exchange, with Ms of comparisons (voting).
- More and more synthetic (model-generated) comparison (preference) data
    - Nowadays mostly synthetic (Llama Team, 2024, Qwen Team, 2025)

Data Filtering

- Deduplication*
- Exclusion of preferences in lowest tier
- Quality assurance reviewing (rejection) of samples*

* Mostly concerns demonstration, and not comparison, data.

Findings

Use of Data

- Early initiatives trained reward models (RMs) based on the comparison data and then optimize models with PPO.
- Later Initiatives use Direct Preference Optimization (DPO).
- When comparison across K>2 model-generated responses collected:
    - Mainly acceleration of data collection.
    - Additional margin component as part of the reward model training (Touvron et al., 2023).
- No information from recent work (post-2022).

Publication of Data

- The work of Bai et al. (2022) is the only initiative that publicly shared data.
- There are publicly available data from other initiatives (Stanford's SHP, NVIDIA's HelpSteer).

Disclosed Information Volume and Quality

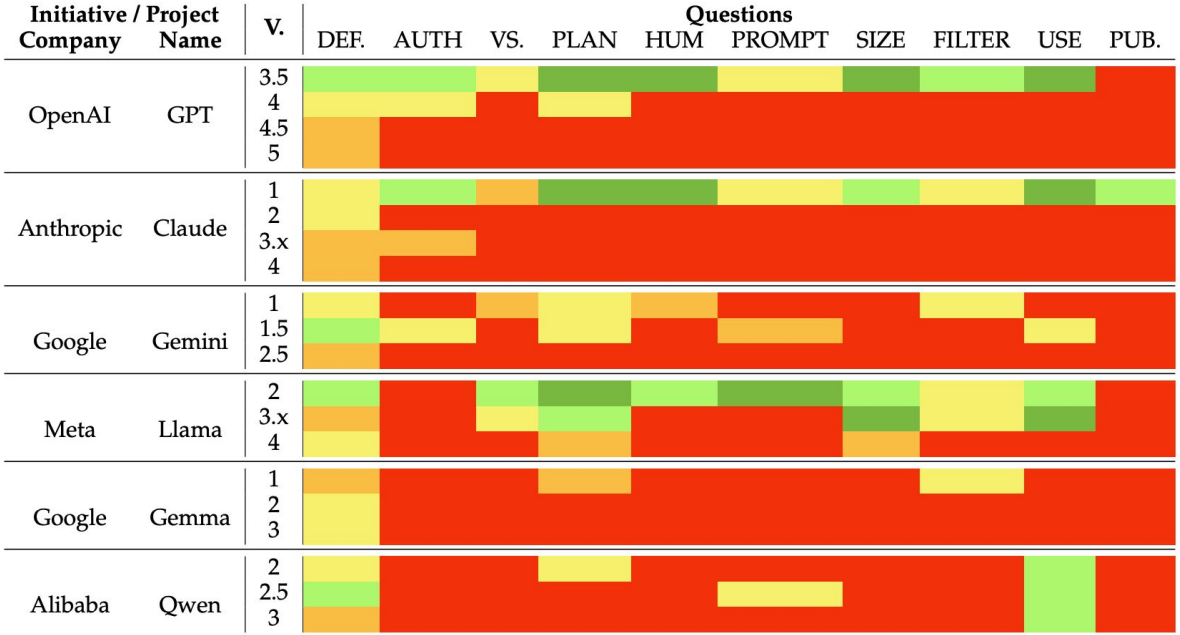


Table 2: Heatmap of the quality (depth) of information per question (Section 2.2) across the examined LLM development initiatives as presented in Section 3.1. There is a keyword per question: Q1 (DEF.), Q2 (AUTH), Q3 (VS.), Q4 (PLAN), Q5 (HUM), Q6 (PROMPT), Q7 (SIZE), Q8 (FILTER), Q9 (USE), Q10 (PUB).

Concerns

The Information stream is draining

- Most information comes from early initiatives (2022)

- Why companies do not share much anymore?
  - Industrial Secrecy or Ignorance?
  - Focus Swift: Concentration on new capabilities (reasoning, other modalities).
  - Maybe nothing really new?
  - Extensive use of Reinforcement Learning from AI Feedback (RLAIF).

- Active Research on:
  - Pluralistic Alignment (Sorensen et al., 2024; Conitzer et al., 2024)
  - Participatory value-setting and data curation (Kirk et al., 2024, Huang et a., 2024)

Concerns

Objectives decided in a vacuum

- The objectives are selected by ***the development team, as a proxy of the corporation*** (employer)
  - Excessive authority, concentration around AI developers' concerns, lack of broader considerations
  - Lack of broader consensus and legitimization

- Corporate alignment (Leike, 2022; Wang & Goksel, 2025, Chalkidis, 2025),
  - Corporate-inspired objectives, instead of priorities and aspiration of society at large.

- Alternatives:
  - Simulated Deliberative Democracy (Leike et al., 2022, Tessler et al., 2024)
    - Learn from human deliberation + Off-load deliberation to AI systems
    - Technocratic "Hack"
    - Is big-tech in a position to champion democracy?
  - Open Human Feedback (Don-Yehiya et al., 2025)
    - Open-Science, Open-Source, Diverse Human Feedback

Concerns

Beyond Helpfulness-Harmlessness-Honesty (HHH)

- The HHH framework introduced by Askel et al. (2021):
  - states: "***We*** ***chose*** *[HHH] as criteria because they are* ***simple*** *and* ***memorable****,*
    and ***seem to capture what we want*** *from an aligned AI.*"
  - meant to be an experimental playground → "*A general language assistant as* ***a laboratory*** *for alignment*"
  - hence, it meant NOT to be complete or elaborate

- Recent findings:
  - LLM aligned with human-curated constitution ***less socially biased, more responsive*** (Huang et al., 2024)
  - Values embedded within RLHF datasets are (Obi et al, 2024):
    - mostly oriented towards ***information-utility values*** (information/knowledge acquisition) and,
    - less towards ***prosocial, well-being, and civic values***.
  - Claude mainly aligns with (Huang et al., 2025):
    - a few key competency- and service-oriented objectives,
    - contrary to humans expressing more diverse values

Concerns

Non-Disclosure of human feedback as labor obfuscation

- Recent work ***does not disclose information*** on the collection of human annotations (labor).
- We know from investigative journalism reports about ***outsourcing of data collection*** related to alignment-focused processes worldwide, primarily in the Global South.
- Industrial-scale human-curated data-sourcing is not acknowledged:
  - ***Lack of accountability***
  - Promotion of ***AI "mysticism"***
  - ***Ethical and moral considerations*** related to the working conditions (Cant et al., 2024)
  - ***Labor obfuscation*** (Guest, 2025)
    - Use of human-crafted content is undermined

"*Openai used kenyan workers on less than $2 per hour to make chatgpt less toxic".* Billy Perrigo, TIME, 2023.

"*Millions of workers are training ai models for pennies*". Niamh Rowe, WIRED, 2023

"*Feeding the machine: The hidden human labor powering AI*". Cant et al., Bloomsbury Publishing USA, 2024.

Concerns

Excessive use of AI as cognitive offloading and deskilling

- Recent Findings:
  - Significant negative impact of dependency on AI tools on critical thinking (Gerlich, 2025)
    - especially for younger participants
  - Use of LLMs is associated with less critical thinking (Lee et al., 2025)
    - swift towards verification (fact-checking) and response integration
  - LLM users displayed the weakest brain connectivity and reduced cognitive activity (Kosmyna et al., 2025)
    - LLM users consistently underperformed at neural, linguistic, and behavioral levels
  - Physicians negatively affected by continuous exposure to AI systems (Budzyn et al., 2025)
    - Less accurate predictions in endoscopy.

- What we would expect?
  - Given the evidence of the negative impact of excessive use of AI systems, we would expect measures:
    - Limiting Critical Thinking, and deskilling are ***harmful*** to the users
    - Promote ***disengagement*** to protect users

- What happens instead?
  - Developers consider the ***welfare of LLM-based systems*** (Anthropic, 2025)
    - Models can disengage to protect themselves!

Concerns

Lack of addressing broader biases

- Notable, more ***social-focused and society-wide types of biases*** are ignored
  - such as political, geopolitical, and cultural biases

- Findings:
  - Aligned LLMs tend to reflect the values and perspectives of (Santurkar et al., 2023):
    - liberal, high-income, well-educated demographic groups in the US,
    - mirroring the profiles of big-tech company owners and high-skilled employees
  - ChatGPT leans towards pro-environmental, left-libertarian positions (Hartmann et al., 2023)
  - LLama models tended to align with liberal, pro-EU stances (Chalkidis and Brandl, 2024)
  - ChatGPT is more aligned with the American (US) culture (Cao et al., 2023)
    - than with Western European or Eastern Asian cultures
  - Most LLMs consistently favor Western (USA, UK) narratives (Salinkov et al., 2025)

- The findings suggest that most LLMs, especially those widely used and developed in the US, seem ***heavily biased towards Western hegemonic ideological views***.

Concerns

Alignment as a "flattening" force - Pluralism under threat

- "Flattening" effect ***on culture*** (Chayka, 2025)
  - Globalization led to a significant homogeneity of how humans approach several aspects of culture, such as art (music, movies, books, etc.), or even their general taste of what accounts as desirable experiences, e.g., which cafes, restaurants, or bars to attend
  - Recent broad use of AI, algorithms decision-making at large, deepens the homogeneity

- "Flattening" effect ***on politics*** (Mouffe, 2005)
  - "End of Politics" → Consensus and Technocracy
  - No antagonism nor pluralism

- LLM alignment can perpetuate "flattening"
  - AI models get aligned with corporate objectives, no consideration for pluralism
  - AI models rapidly replace search engines
    - substantial increased offloading of cognitive labor and decision-making

Questions?

Further Reading

- Artificial Intelligence, Values, and Alignment (Gabriel, 2020)
- A Collectivist, Economic Perspective on AI (Jordan, 2025)
- Why human-AI relationships need socioaffective alignment (Kirk et al., 2025)
- What Does 'Human-Centred AI' Mean? (Guest, 2025)
- Feeding the Machine: The Hidden Human Labor Powering AI (Cant et al., 2024)