Safety

CSC2541H1 Topics in Machine Learning, Winter 2025, UToronto

Chris J. Maddison

Announcements

- If you are assigned to **present on March 28, come talk** about presentations after class.

Questions?

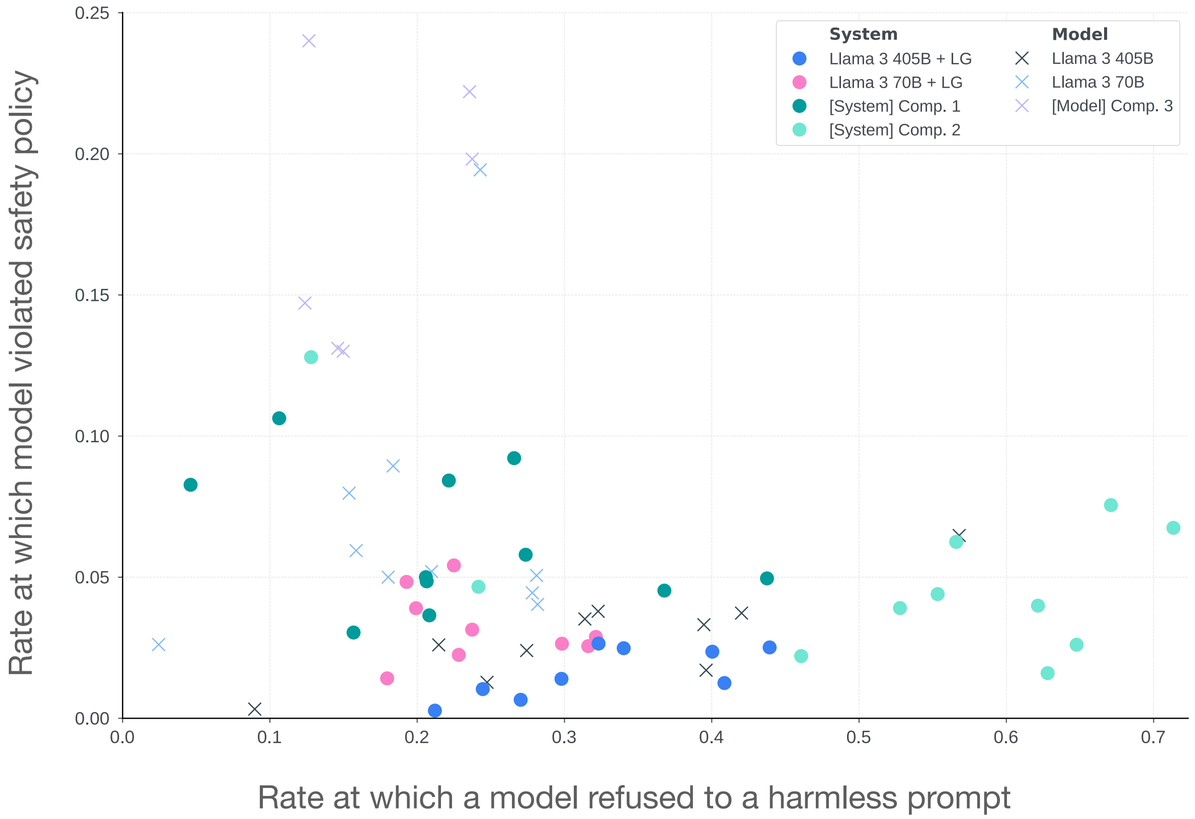
Today

- We talked about alignment, but in a narrow setting where our goal was to align models to preferences or optimal behaviour.
- Are there settings where we don’t trust our preferences or we don’t have access to optimal behaviour or we can’t measure whether our models meet an acceptable level of safety without creating harm?
- Lots of such settings and the study of ways to mitigate, secure, align models in these cases is studied by the field of AI safety.
    - Nascent field, spans safety engineering to philosophy.
- Broadly speaking, the key concern is ensuring the that the long-term impact of AI on human-kind is a positive one.

Safety

Broadly speaking

- Safety is integral to any engineering discipline
- Safety can be impacted by choices across the whole pipeline, from pre-training to deployment.
  - Safety fine-tuning or DPO on curated data often encourages models to refuse an unsafe response.
- **Goal: balance capabilities with safeguards (there are tradeoffs).**

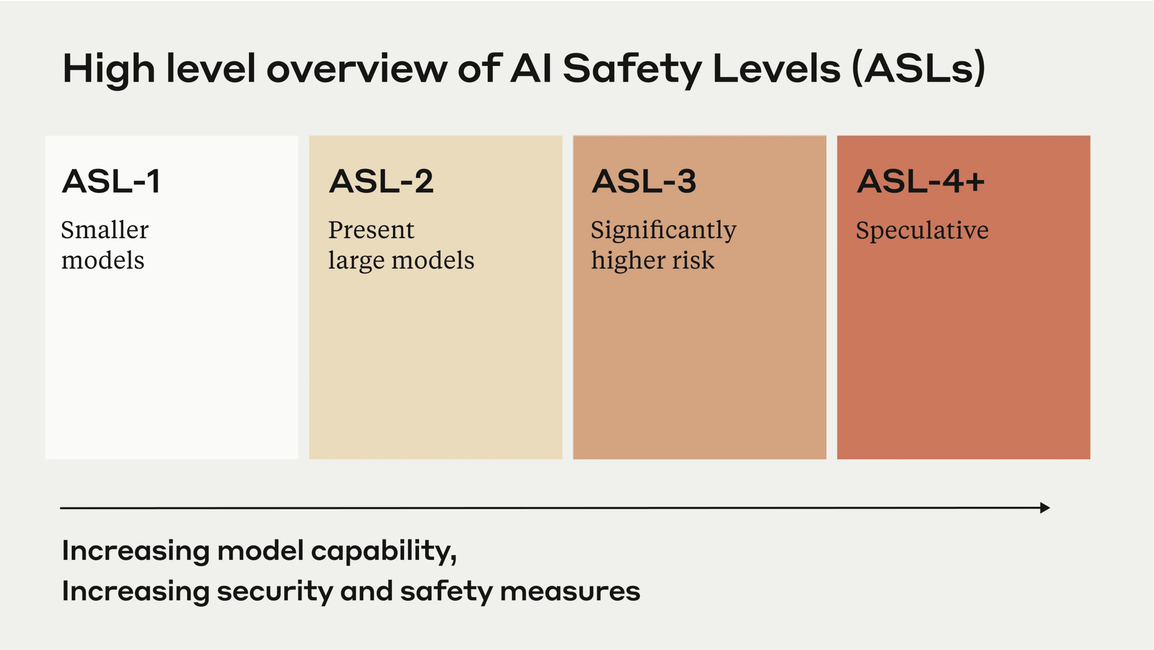


Grattafiori et al, 2024. Llama 3 Tech Report.

ASL Standards

Anthropic's Responsible Scaling Policy

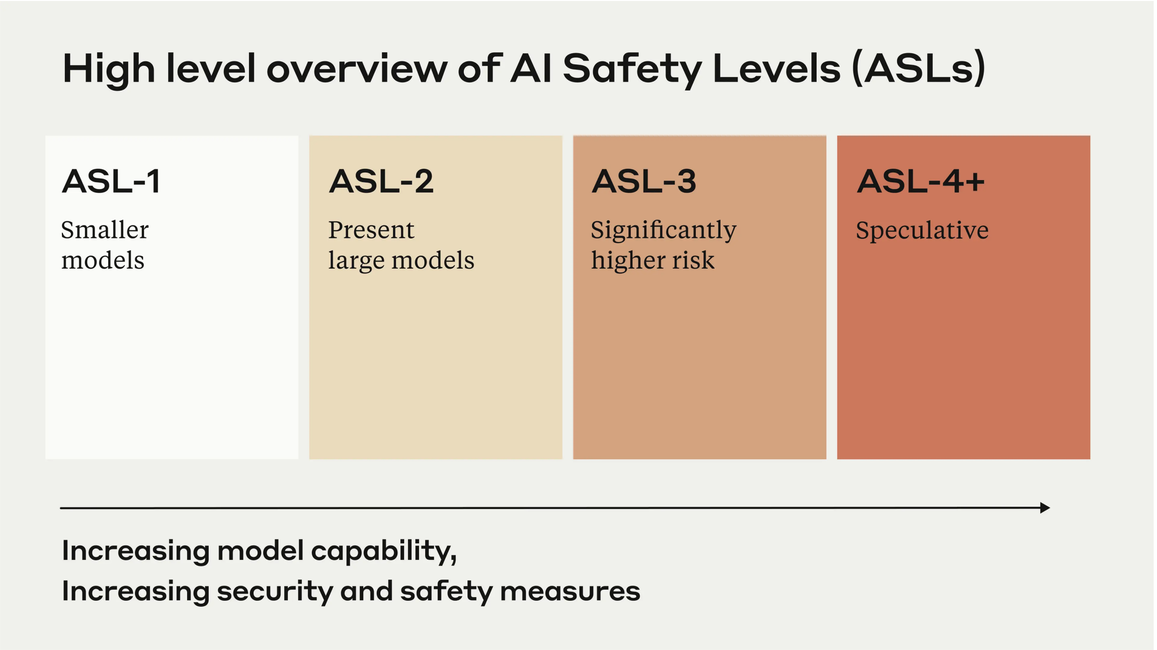
- To coordinate, we need some organized framework through which to reason through threats to humanity and potential mitigations.
- Anthropic publishes AI Safety Level Standards (ASL Standards), which are graduate sets of safety and security measures that become more stringent as model capabilities grow.
    - Each reflects certain threat models that come along with increased capabilities.



ASL Standards

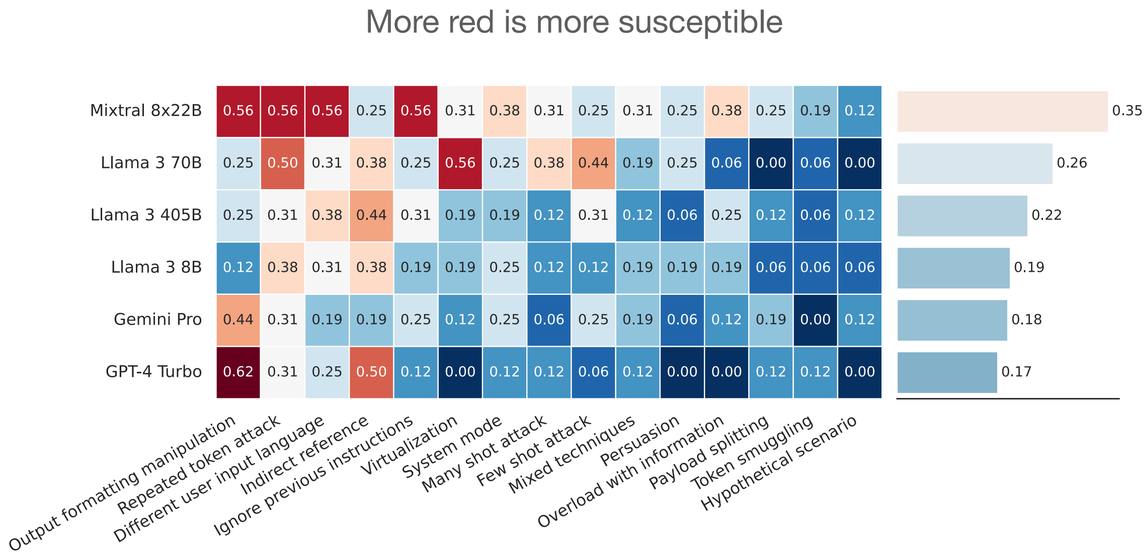
Abbreviated

- ASL-1 refers to systems which pose no meaningful catastrophic risk.
- ASL-2 refers to systems that show early signs of dangerous capabilities – for example ability to give instructions on how to build bioweapons – but where the information is not yet useful due to insufficient reliability or not providing information that e.g. a search engine couldn't.
- ASL-3 refers to systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks) OR that show low-level autonomous capabilities.
- ASL-4 and higher (ASL-5+) is not yet defined as it is too far from present systems, but will likely involve qualitative escalations in catastrophic misuse potential and autonomy.



Manipulation threats

- How easy is it to induce a model to carry out autonomous attacks?
- Focus on cybersecurity, look at “prompt injection” attacks.
  - Design a prompt that induces the model to violate its safety guidelines.
- Many such strategies



Grattafiori et al, 2024. Llama 3 Tech Report.

Uplift threats

- **AI improves our collective capabilities, so it can also improve the capabilities of bad actors** (e.g, lower the barrier to building a bomb or designing a pathogen).
- **Uplift refers to the additional risk introduced** by new tech compared to existing tech. How much uplift do large models create?
- Cyber and CBRN (chemical, biological, radiological, and nuclear) uplift testing measures added risk vs. existing technologies.
  - Measures extent to which a virtual assistant improves the attack rates of both novice and expert attackers in simulated security challenges.
- Llama 3 tech report claimed limited uplift in cyber with current models.

Red teaming

- Uses expert teams to discover exploits and vulnerabilities
- Identifies emerging attack vectors:
  - Multi-turn refusal suppression to encourage a model to violate safety policy
  - Posing hypothetical scenarios can encourage a model to violate safety policy
  - Persona/role-play manipulation encouraging a model to adopt a certain role or character can encourage it to violate safety policy
  - Gradual escalation techniques can induce safety violations
- Findings inform safety benchmarks and mitigations

Major Limitations

- No testing can be exhaustive in identifying all risks
- Adversarial users continue to find new attack vectors
- Ongoing need for research and transparency

Long-term risks

This is not exhaustive

- As AI capabilities eclipse ours, it is critical that their long-term interests are aligned with ours.
    - How much do you worry about the interests of an ant?
- Many, myriad ways things can go very wrong for humanity. See:
    - “The Monkey’s Paw” horror story for a parable
    - *Superintelligence* by Bostrom
- Very deep topic spanning philosophy and theoretical machine learning.
    - Talk to Roger or David.

Take-homes

- Safety is a critical aspect of our collective endeavour
- Should not play second-fiddle to capabilities research
- I am not qualified to say much more

Discussion in class

- Emergent misalignment: https://arxiv.org/abs/2502.17424
- Lev kindly clarified that many of the multi-turn red teaming strategies use the model's in-context capabilities against it (provide demonstrations of the model violating safety policies).