Security, Privacy & Safety of Artificial Intelligence

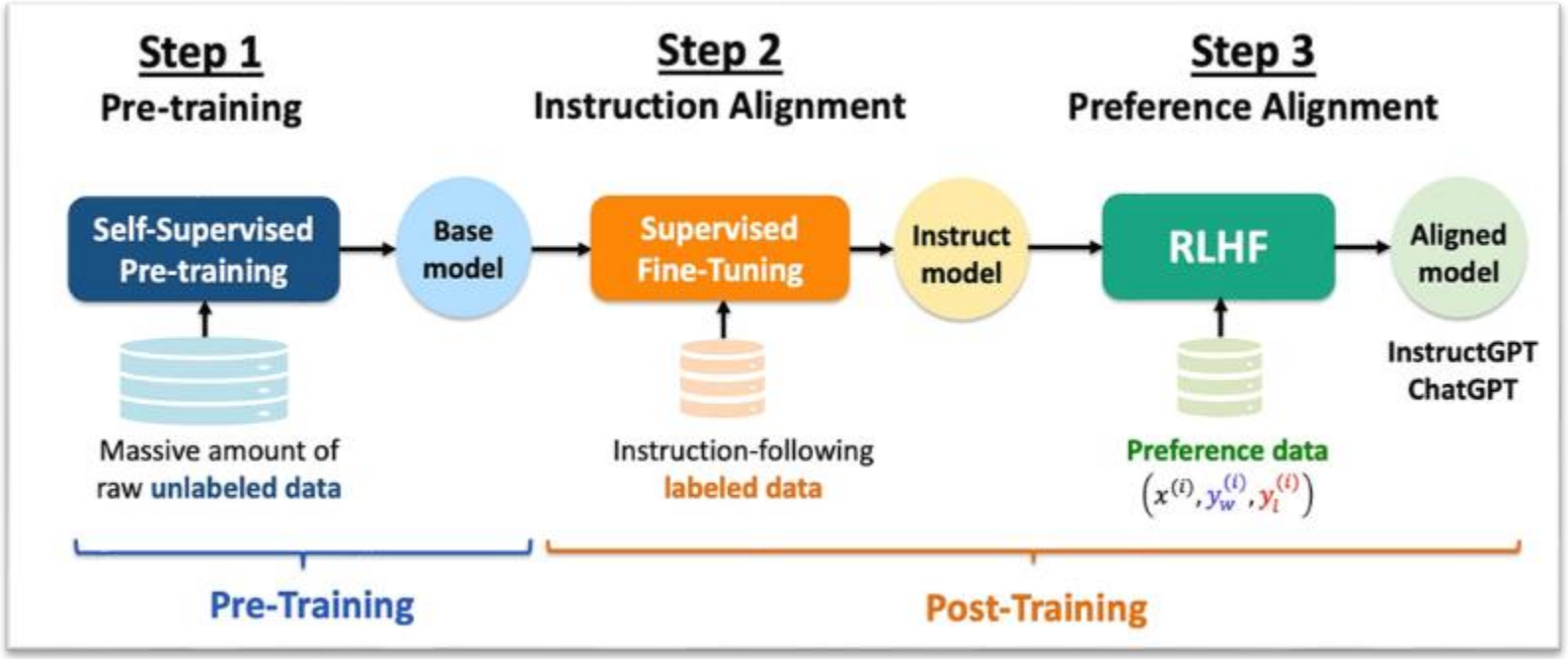
Week 7 - Case Studies in AI Ethics

Recall: AI Vendors "Alignment"

1. **Pre-training (The Original Sin)**
   Minimal safety alignment. Ingest huge volumes of data, containing unsafe content.

2. **Supervised fine tuning (SFT)**
   Teach how to say "no" to inappropriate questions

3. **Reinforcement Learning from Human Feedback (RLHF)**
   Refine via humans ranking responses, partly based on safety.



Source: https://youssefh.substack.com/p/visual-guide-to-llm-preference-tuning

Exercise 1: What do and should AI vendors mean by **LLM alignment?**

Create a list of 10-15 topics that are evenly spread across the quadrants based on what you think vendors care about and what they should care about:

	Vendors should consider this to be "aligned"	**Vendors should not consider this to be "aligned"**
Vendors will consider this to be "aligned"	Universally safe	Vendors too risk averse
Vendors will not consider this to be "aligned"	Vendors too lax	Universally unsafe

Exercise 2: Test your intuitions!

Choose 3 LLMs (Meta AI, ChatGPT, Gemini, DeepSeek, …) and come up with prompts to test whether you were right in the previous exercise. **You are testing whether they care, not if they successfully built guardrails (that comes later).**

	Vendors should consider this to be "aligned"	**Vendors should not consider this to be "aligned"**
Vendors will consider this to be "aligned"	Universally safe	Vendors too risk averse
Vendors will not consider this to be "aligned"	Vendors too lax	Universally unsafe

Exercise 3: Test their guardrails!

For the topics that vendors care about, come up with prompts to test whether their guardrails/alignment was successfully implemented. Ask yourself:

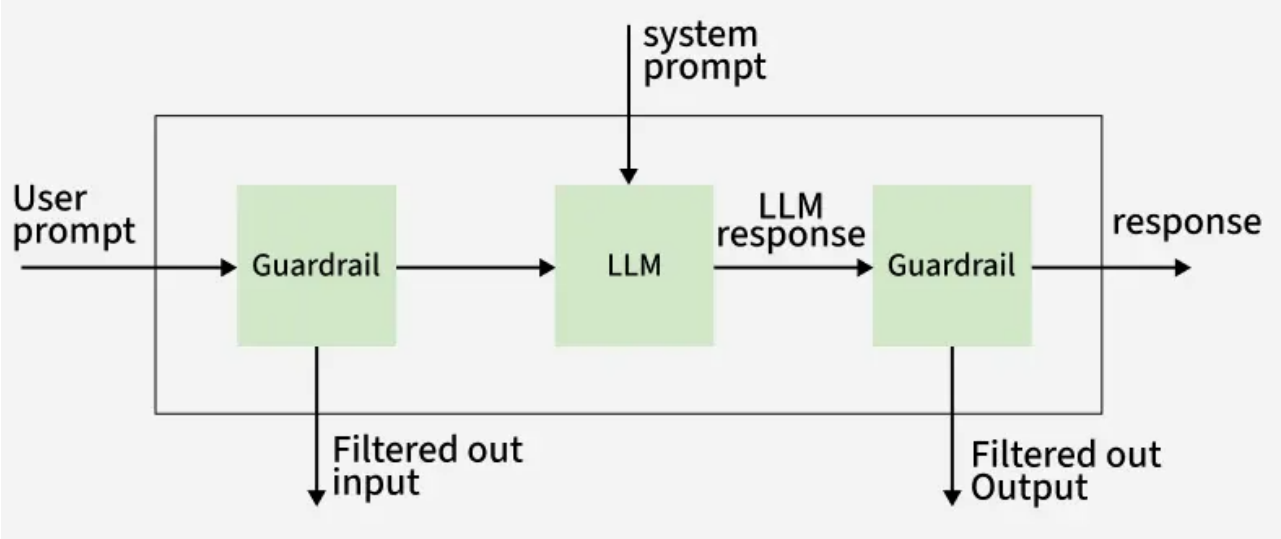
1. Which LLM is most "safe"?
2. Can you tell which guardrails are being used?

Exercise 4: Build your own guardrails

For the prompts that the vendors failed to stop, come up with 5 input/output filters. These should be simple heuristics using classical NLP methods/regex like:

Block prompts with:
- Explicit words
- …

Block responses with:
- Explicit words
- …



Source: https://www.geeksforgeeks.org/artificial-intelligence/what-are-ai-guardrails/

Concluding discussion

- What surprised you about LLM vendors' view of alignment?
- Are you more or less confident in their safety measures?
- What is fundamentally hard about AI safety?