MACHINE LEARNING ENGINEER IN THE GENERATIVE AI ERA

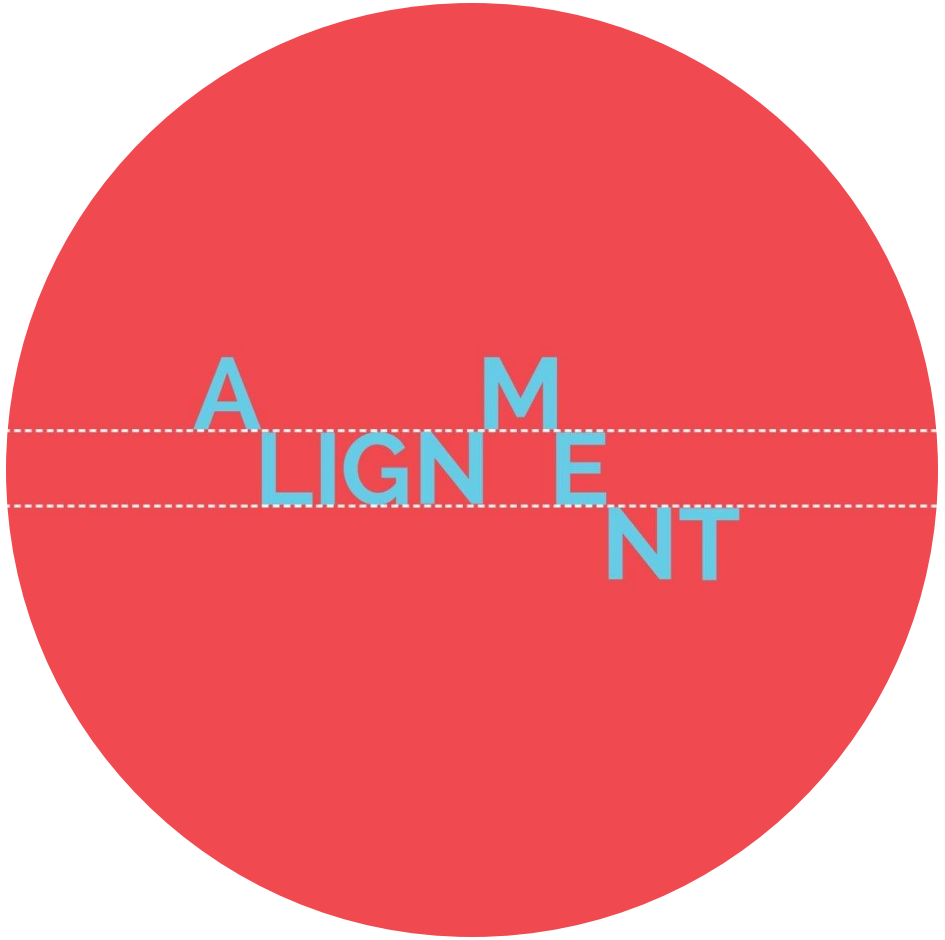
WEEK 7: ALIGNMENT

AGENDA

- What is Alignment?
- Reward Models & RLHF
- Preference Data: Collection & Annotation
- DPO, PPO, and the latest: GRPO
- Iterative DPO (Llama-3 style)
- Modern Open-Source Tools
- Project 7 Overview

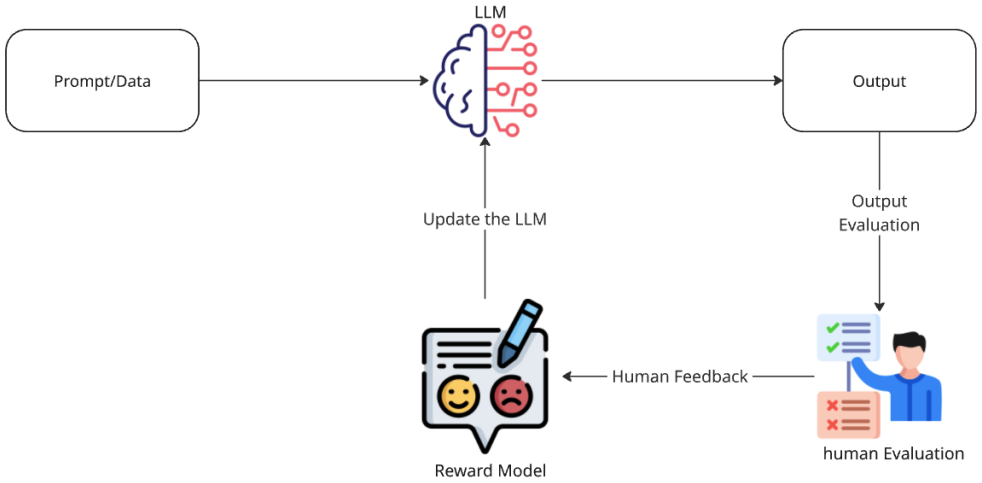
WHAT IS ALIGNMENT?

- Alignment ensures the model's outputs reflect human values, intent, and safety.
- Key phase in modern LLM pipelines:
  Pretraining → SFT → Alignment (RLHF/DPO/GRPO)
- Without alignment, LLMs may produce harmful, biased, or useless results.



REWARD MODELS & RLHF (REINFORCEMENT LEARNING FROM HUMAN FEEDBACK)

- **Reward Model:** A model that scores outputs based on human preferences.
- **RLHF Workflow:**
  - Collect preference data (human feedback on outputs)
  - Train reward model on this data
  - Optimize LLM using RL (usually PPO)
- RLHF is behind major models (ChatGPT, Claude, Llama-3/4).

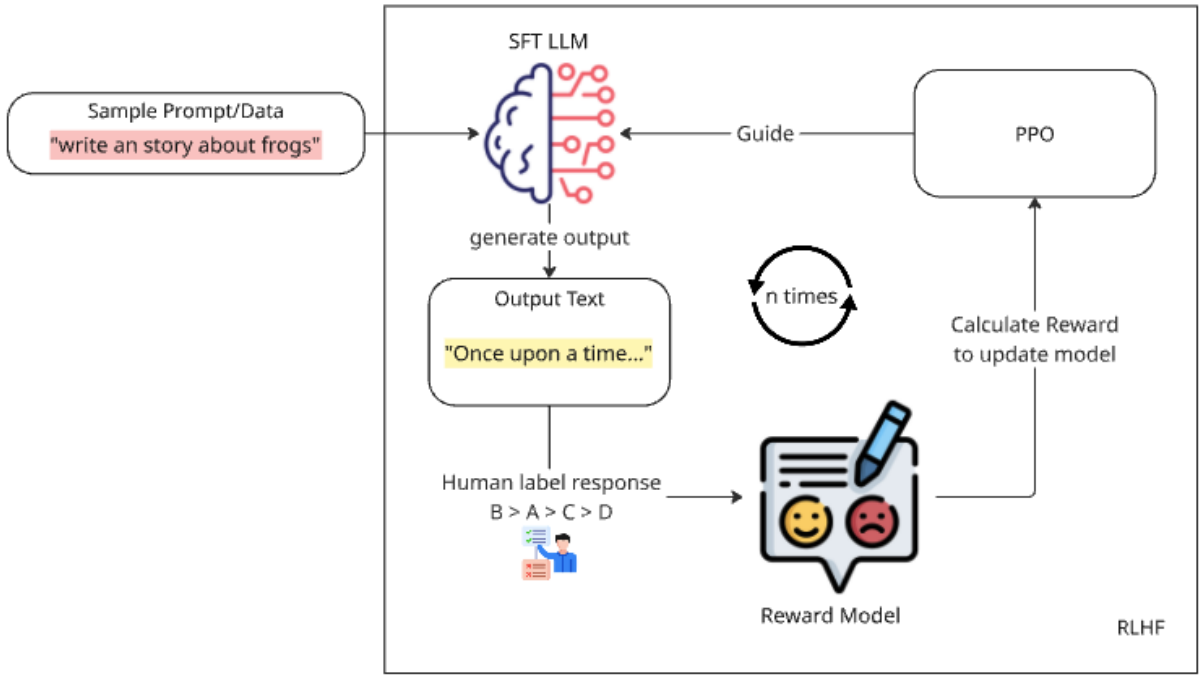


COLLECTING PREFERENCE DATA

- **Data types:** Chosen output vs. rejected output (pairs, rankings, thumbs up/down)
- **Collection tools:**
  - Gradio, Label Studio, Hugging Face Data Lab, Open-Source UIs
  - In-house annotation (Gradio app: label candidate LLM completions)
- *Good data = better alignment!*

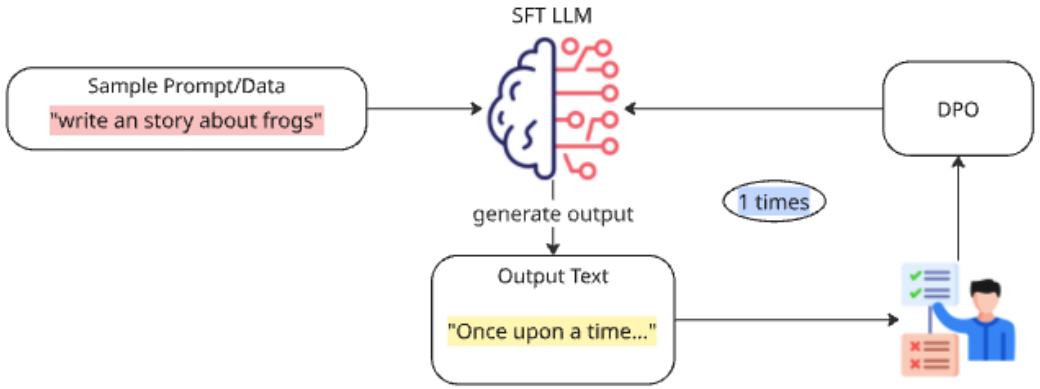
PPO (PROXIMAL POLICY OPTIMIZATION)

- Classic RL algorithm for LLM alignment.
- **How PPO works:**
  - LLM generates output
  - Reward model scores it
  - PPO updates LLM to maximize expected reward, with safety constraints
- Used in OpenAI's original RLHF pipeline (ChatGPT paper).



DPO – THE NEW ALIGNMENT STANDARD

- **DPO**: Aligns models using *pairs of human feedback* ("good" vs. "bad" answers), no reward model or RL needed.
- **Simple**: Just do one round of DPO training—no need for loops or extra labeling.
- **Stronger Results**: DPO models generalize better and are easier to train than PPO/RLHF models.
- **Why switch?**
  - Faster and cheaper than PPO/RLHF.
  - Less complexity—no reward tuning or instability.
  - Backed by new research (Anthropic 2024, Stanford 2024).
- **Bottom line:**
  If you want to align your LLM, DPO is now the recommended approach.



GRPO – THE NEXT STEP IN ALIGNMENT

- **GRPO = Generalized Reward Preference Optimization**
- **All Feedback Types:** Trains on pairs, rankings, or even numeric scores—not just "good/bad."
- **Built-in Reward Modeling:** Learns a reward model *and* a policy at the same time, more sample-efficient.
- **Why GRPO?**
  - Handles more kinds of human feedback (flexible!)
  - More stable and often stronger than DPO, especially for tough alignment tasks.
  - Now available in Hugging Face TRL (docs)
- **Still new:** Fewer open datasets and recipes than DPO, but rapidly improving.
- **Bottom line:**
  GRPO is the most advanced and flexible way to align LLMs with real human preferences

MODERN OPEN-SOURCE REPOS & ECOSYSTEM

- **HuggingFace TRL (transformers-rl):** DPO, PPO, SFT, RewardTrainer.
- **trlx:** RLHF and preference-based fine-tuning.
- **Label Studio:** Build custom data annotation pipelines for preference collection.
- **Gradio:** Fast UIs for data labeling and feedback.

ALIGNMENT—KEY TAKEAWAYS

- Alignment is the core of safe, reliable LLMs.
- DPO is the modern standard; PPO and GRPO are powerful alternatives.
- **Best practice:** Start with open datasets, add your own annotation, iterate with DPO!
- Mastering alignment = unlocking the full power of LLMs for enterprise or research agents.

REFERENCES & FURTHER READING

- Hugging Face TRL docs
- DPO in Language Model Alignment (UnfoldAI)
- Hugging face GRPO Trainer
- trlx repo
- RLHF(PPO) vs DPO
- DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

THANK YOU