Aligning Language Models with Human Feedback without Reinforcement Learning

2nd Cycle Integrated Project in Computer Science and Engineering 2023/2024

Presentation by :

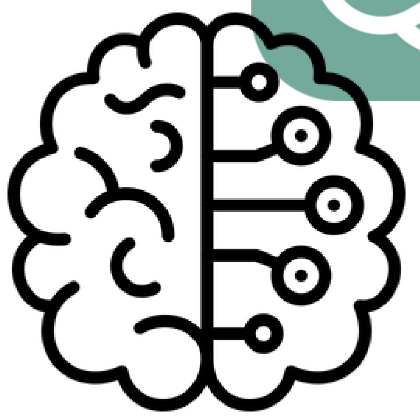
MARTIM SANTOS

TÉCNICO LISBOA

Roadmap

Motivation

- LLMs can often contain misleading and toxic content
- Well known models such as GPT-3.5, GPT-4, ... → **RLHF**
- Impressive outcomes <u>but</u> with downsides: complexity, instability and sensitivity to hyperparameters.
- Empirical success and usability in real-life scenarios



Roadmap

Objectives

- **Compare thoroughly** the different **novel RL-free approaches**, mainly in the **dialogue (single-turn and multi-turn), machine translation, and summarization** tasks.

- Examine how **benchmarks for assessing model performance** in these tasks align with human preferences for usefulness and safety.

- Propose and implement a **novel approach** that uses similar backbone models and aims to outperform the existing ones in the referred tasks, by **combining the strengths** of each model and aligning better with human preferences.

TÉCNICO LISBOA

Roadmap

Background

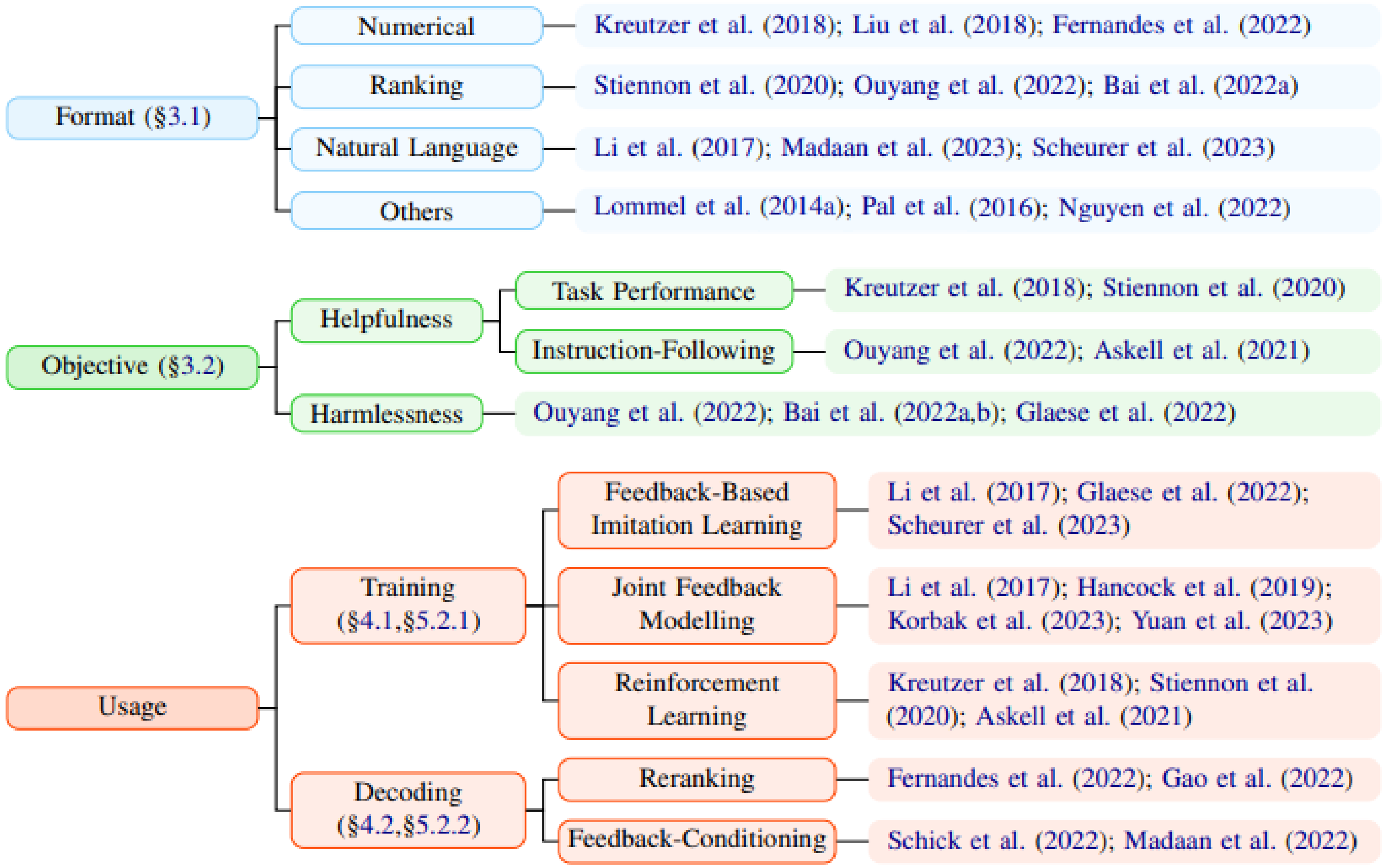
- LLMs (e.g. OpenAI's GPT series, Google's PaLM, and Meta's LLaMA)
- Training procedure: pre-training and then SFT (commonly Instruction-tuning)
- Challenges remain: misalignments objective-expectations, critical error detection and subtle language nuances
- Beyond fine-tuning

Leveraging human feedback is key.

Background

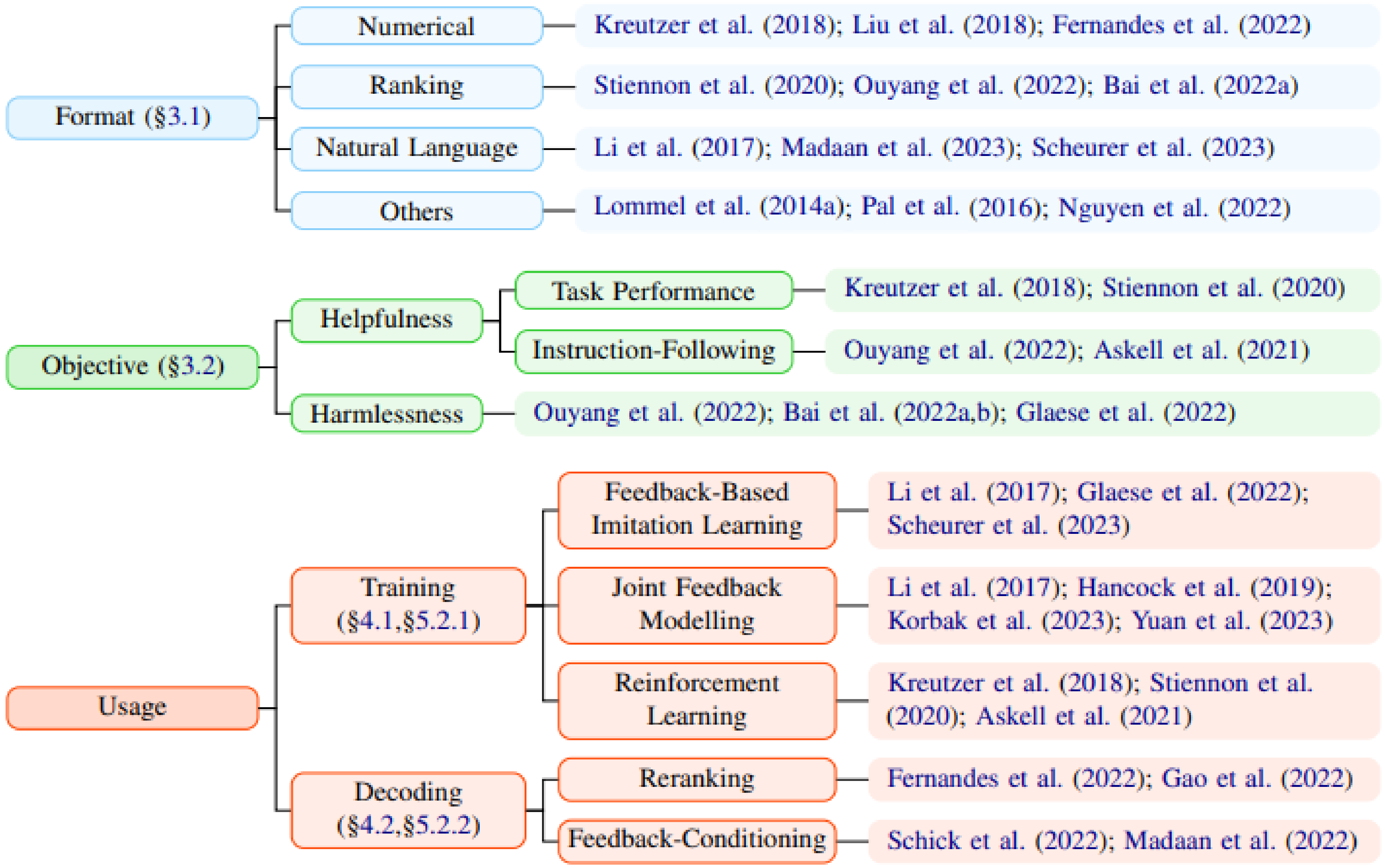
- Feedback **format**:
  - **Numerical** - simple scalar
  - **Ranking-based** - ranking outputs
  - **Natural Language** - details
  - Domain-specific
  - Others

The choice of feedback format has implications on: expressivity, the ease of its collection, usage to improve models.



Background

- Alignment **objective**:
  - **Harmlessness** - avoiding harmful outputs.
  - **Helpfulness** - is assessed through task-related feedback, enhancing overall system usefulness.

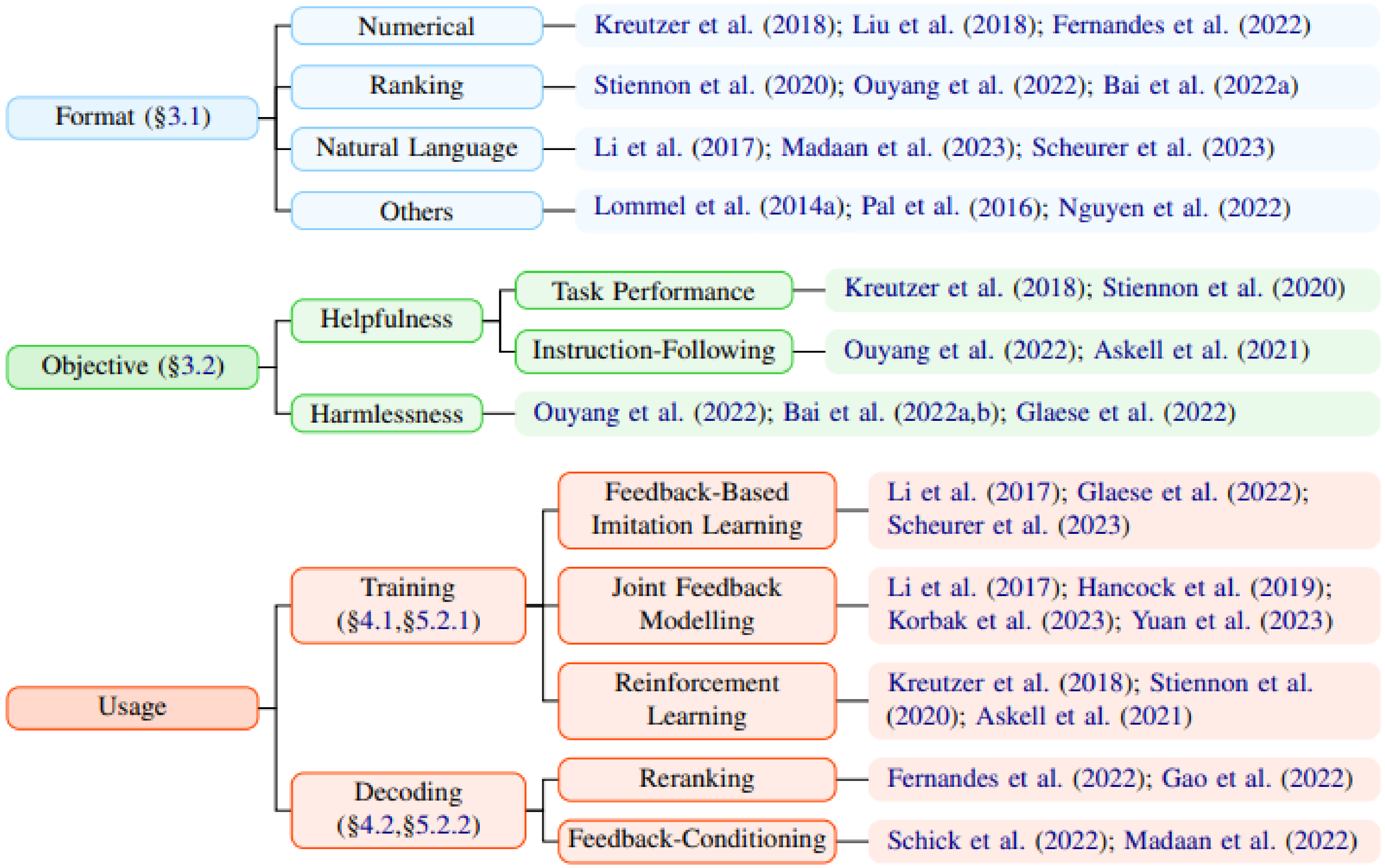


Background

- **Training approaches:**
  - **Feedback-based imitation learning** - only positively labeled generations
  - **Joint feedback modeling** - all human feedback, diverse formats.

$$\mathcal{L}_i(\theta) = -\log p_\theta(f_i \mid y_i, x_i) + \log p_\theta(y_i \mid x_i)$$

AND...



Reinforcement Learning from Human Feedback

- addresses misalignment in FT objectives
- human feedback
- toxicity, user preferences and ethical outputs

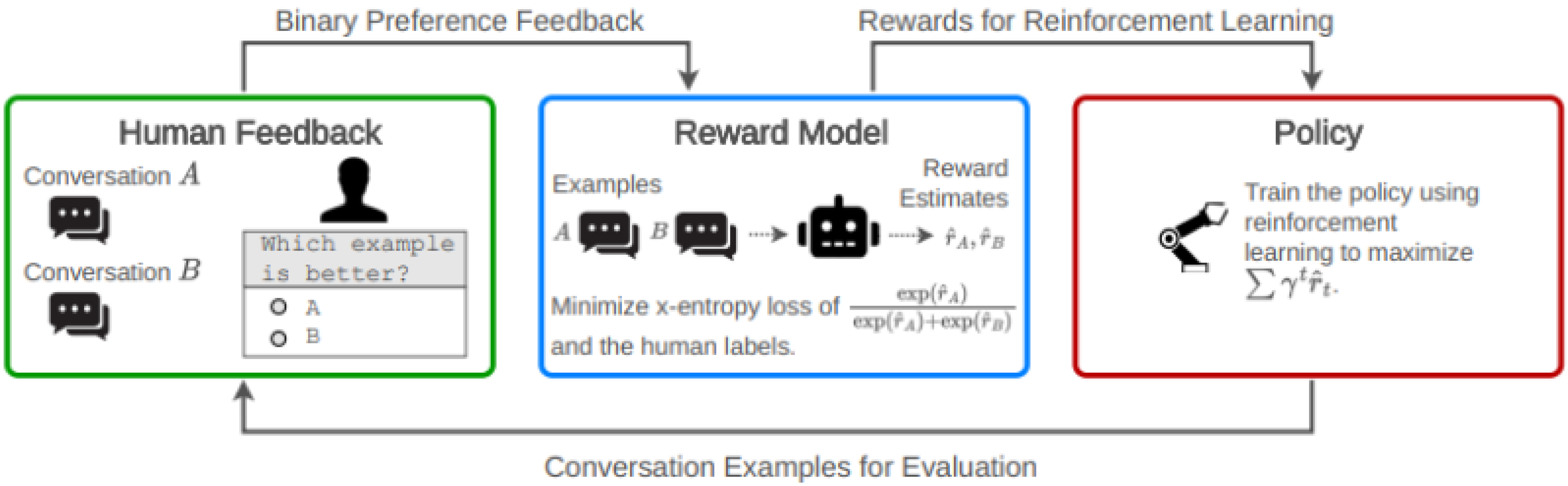
- Extends across various NLP domains
- Scalability - hundreds of billions of parameters

Reinforcement Learning from Human Feedback

- First stage: **Supervised Fine-tuning (SFT)**
- Second phase: **prompting and human labelers**
- **Bradley-Terry model (BT)**: higher score if preferable
- **Reward model**: estimated using maximum likelihood on a dataset of human judgments.

$$\mathcal{L}(r_\theta, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\Big[\log \sigma(r_\theta(x, y_w) - r_\theta(x, y_l))\Big],$$

Reinforcement Learning from Human Feedback



Reinforcement Learning from Human Feedback

- Reward function to guide model training, optimizing for higher-quality human-judged outputs:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(y \mid x)}\left[r_\theta(x,y) - \beta D_{KL}(\pi_\theta(y \mid x) \parallel \pi_{\text{ref}}(y \mid x))\right]$$

- KL divergence term dual purpose
- Objective is non-differentiable and is typically optimized using reinforcement learning methods such as Proximal Policy Optimization (PPO) or REINFORCE.

- <u>Standard</u>: reward function + PPO

Reinforcement Learning from Human Feedback

LIMITATIONS:

- **PPO trial-and-error, instability and computational demands**
  - Reward rAnked FineTuning (**RAFT**)

- **Additional training** dedicated to **refining reward models**
  - loading multiple LLMs for PPO training
  - heavy memory burden, limits scalability and practical applicability

Roadmap

Related Work

Alternative Approaches (offline/RL-free)

Direct Preference Optimization (DPO):

- uses preferences *directly*
- new parameterization for the reward model

$$r(x, y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x).$$

$$p^*(y_1 \succ y_2 \mid x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

- loss function over rewards → loss funcion over policies

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right].$$

Related Work

Rank Responses to Human Feedback (RRHF):

- responses from various sources
- responses sampled using different policies
- reward function assigns scores

$$
\begin{aligned}
\mathcal{L} = \mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{ft}} &= \sum_{r_i < r_j} \max(0, p_i - p_j) - ||y_{i'}||\, p_{i'} \\
&= \sum_{r_i < r_j} \max\left(0, \pi_\theta(y_i \mid x) - \pi_\theta(y_j \mid x)\right) - ||y_{i'}||\, \pi_\theta(y_{i'} \mid x)
\end{aligned}
$$

- Simpler and requires only 1-2 models, less hyperparameters and is easy to implement

Related Work

Sequence Likelihood Calibration (SLiC):

- leverages feedback from another model
- simple loss function
  - rank calibration + cross-entropy regularization

$$\mathcal{L}(\theta) = \max\left(0, \delta - \log \pi_\theta(y_w \mid x) + \log \pi_\theta(y_l \mid x)\right) - \lambda \log \pi_\theta(y_{\text{ref}} \mid x)$$

Related Work

Statistical Rejection Sampling Optimization (RSO):

- addresses:
  - DPO's containment in sampling
  - SLiC restriction to sampling only from SFT policy
- higher-reward regions → enhance sampling from optimal policy

Step 1: Generate $y \sim \pi_{\text{SFT}}(y \mid x)$ and $u \sim U[0, 1]$.

Step 2: Let $M = \min\{m \mid m\pi_{\text{SFT}}(y \mid x) \geq \pi_{r_\theta}(y \mid x) \text{ for all } y\}$. If $u < \frac{\pi_{r_\theta}(y \mid x)}{M\pi_{\text{SFT}}(y \mid x)}$, then accept y. Otherwise, reject y and redo the sampling.

Related Work

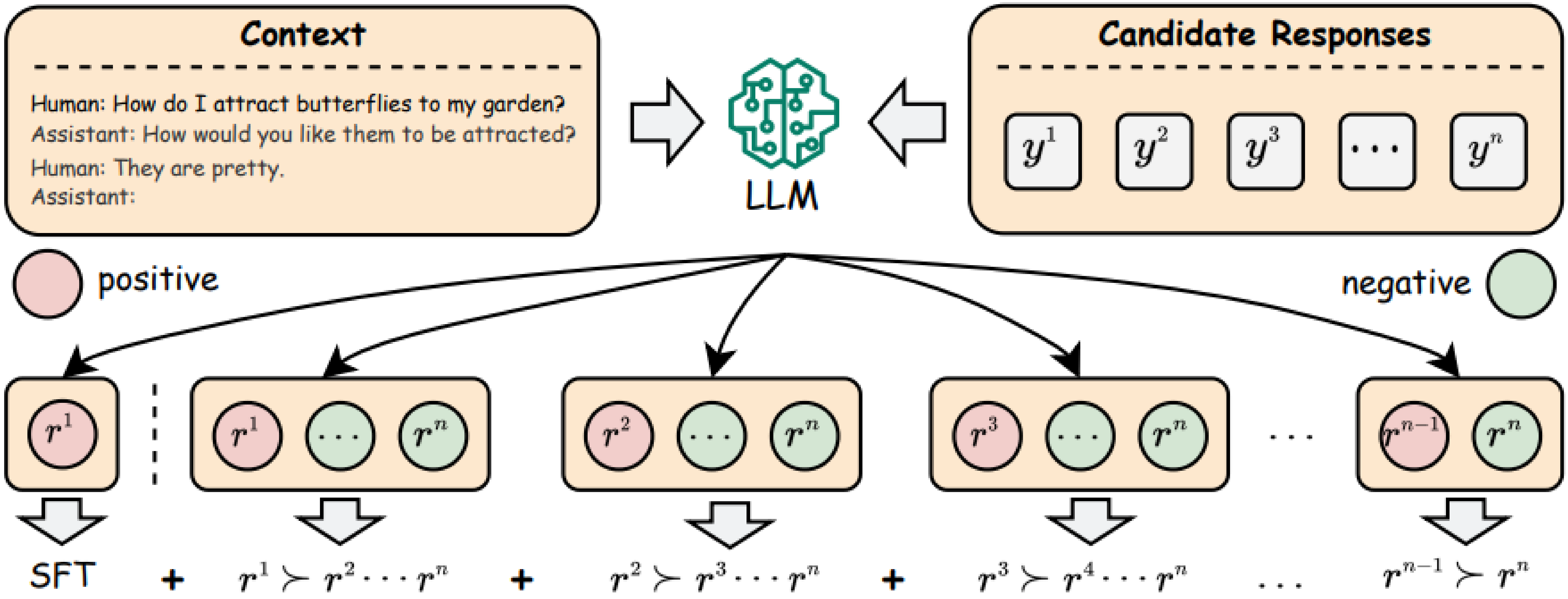
Preference Ranking Optimization (PRO)

- extends BT pairwise comparison

$$y_1 \succ y_2 \succ \ldots \succ y_n \qquad\qquad y_{1,2:n} = y_1 \succ \{y_2, \ldots, y_n\}$$

- fully leverage the preference rankings

$$P(y_{1,\ldots,n} \mid x) = \prod_{k=1}^{n-1} P(y_{k,k+1:n} \mid x) = \prod_{k=1}^{n-1} \frac{\exp(r(x, y_k))}{\sum_{i=k}^{n} \exp(r(x, y_i))}$$



- first response, others responses as negatives. Then removes the current, and moves to the next

Related Work

Preference Ranking Optimization (PRO)

- overall optimization objective:

$$\mathcal{L}(y_{1,\dots,n} \mid x) = \mathcal{L}_{\text{PRO}} + \beta \mathcal{L}_{\text{SFT}}$$

- where LSFT is the negative log-likelihood loss of the top 1 candidate and LPRO is defined as:

$$\mathcal{L}_{\text{PRO}} = -\sum_{k=1}^{n-1} \log \frac{\exp(r_{\pi_{\text{PRO}}}(x, y_k))}{\sum_{i=k}^{n} \exp(r_{\pi_{\text{PRO}}}(x, y_i))}$$

Roadmap

Proposed Solution

- **Focus:** dialogue (single and multi-turn), machine translation, and summarization tasks.

- Fine-tune backbone model via **SFT**, implement **DPO**, **RSO** and **PRO**

- **Novel approach** that uses similar backbone models and combines the strengths of these different approaches

Proposed Solution

- **<u>Objective:</u>** Incorporate **PRO's BT model extension** with **reward reparameterization from DPO**

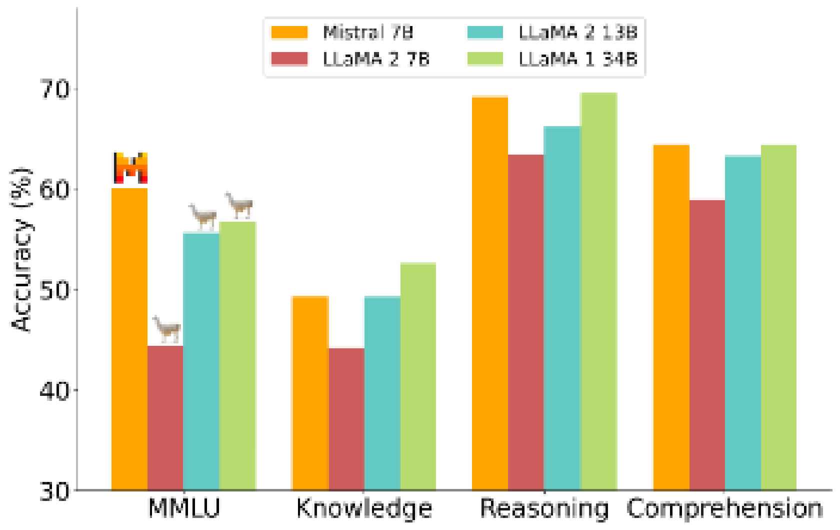
$$P^*(y_1 \succ \{y_2, \cdots, y_n\} \mid x) = \frac{1}{1 + \sum_{i=2}^{n} \exp\left(\beta \log \frac{\pi^*(y_i \mid x)}{\pi_{\text{ref}}(y_i \mid x)} - \beta \log \frac{\pi^*(y_1 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)}\right)}$$

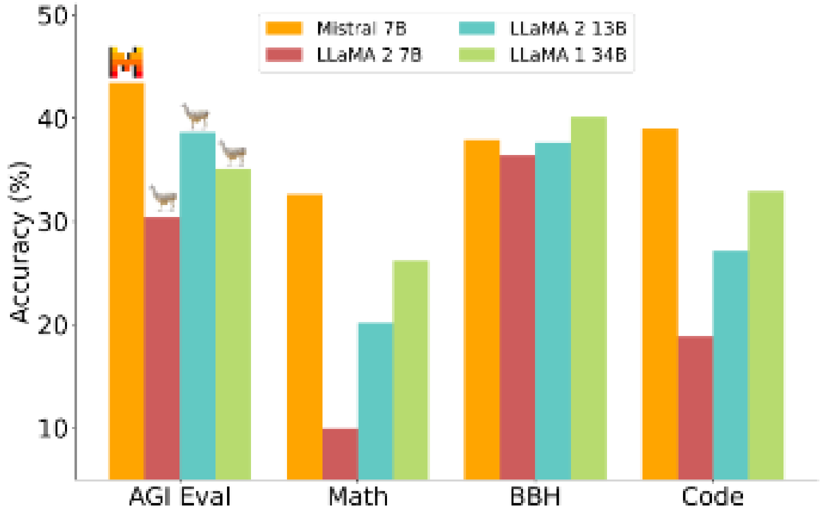
To fully leverage the ranking as it is done in PRO, we deduce

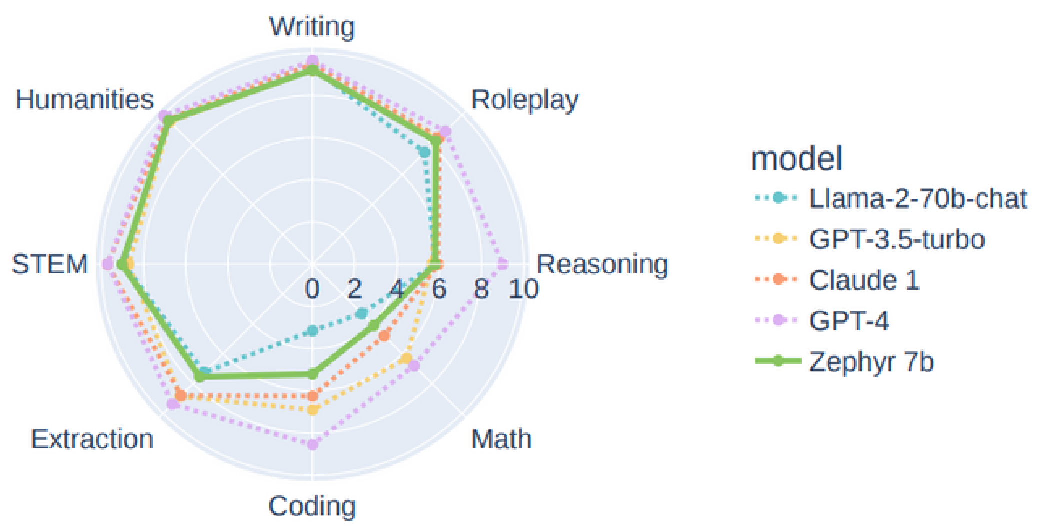
$$\begin{aligned} P^*(y_1 \succ \cdots \succ y_n \mid x) &= \prod_{k=1}^{n-1} P^*(y_k \succ \{y_{k+1}, \ldots, y_n\}) \\ &= \prod_{k=1}^{n-1} \frac{1}{1 + \sum_{i=k+1}^{n} \exp\left(\beta \log \frac{\pi^*(y_i \mid x)}{\pi_{\text{ref}}(y_i \mid x)} - \beta \log \frac{\pi^*(y_1 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)}\right)} \end{aligned}$$

Architecture

- We propose **LLaMa-7B**, used in multiple approaches
- Explore alternative models such as **Mistral 7B** given its great performance results and **Zephyr-7B**, an aligned version of Mistral, which surpasses LLama2-Chat-70B, the best open source RLHF-based model, on many benchmarks







Datasets

Focus on 3:

- **UltraChat** – which consists of 1.47M multi-turn dialogues generated by GPT-3.5-TURBO over 30 topics and 20 different types of text material.
- **UltraFeedback** – consists of 64k prompts, each of which has four LLM responses that are rated by GPT-4 according to different criteria.
- **Reddit TL;DR** – a popular summarization dataset that contains human preferences gathered by previous works.

Roadmap

Evaluation

- Compare results across various approaches, including RLHF and RL-free baselines
- Leverage powerful LMs (e.g. GPT-4 and Claude) as automatic evaluators
- Win rate against baseline policy, using GPT-4 as proxy for human evalutation

Evaluation

Single and multi-turn dialogue:

- **MT-Bench –** multi-turn benchmark comprising 160 questions spanning 8 different domains.
- **AlpacaEval –** single turn benchmark involving the generation of responses to 805 questions of various topics.

We can also evaluate our models on the **Open LLM Leaderboard,** from HuggingFace

Evaluation

Summarization:

- Discard some automatic metrics (e.g. **ROUGE**)
- **BLEU** to assess the text quality, to compare inference results with preferred responses in test sets.

Evaluation

Similarly to recent works, we plan to evaluate other existing approaches to training language models to adhere to human preferences, namely :

- zero-shot prompting with **GPT-J** in the summarization task
- 2-shot prompting with **Pythia-2.8B** in the dialogue task.

In addition, we plan to evaluate the **SFT** model as well as **Preferred-FT** fine-tuned for each of the tasks, and finally **PPO** using a reward function learned from the preference data.

TÉCNICO LISBOA

Roadmap

Thank You!

Presentation by :

MARTIM SANTOS