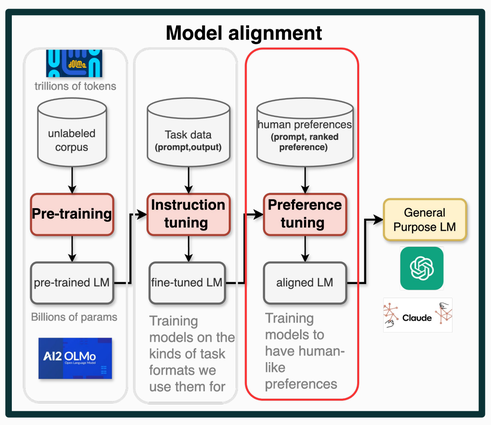
Declarative Characterizations of Direct Preference Alignment Algorithms

Kyle Richardson, Vivek Srikumar, Ashish Sabharwal

Ai2 · THE UNIVERSITY OF UTAH

Preference alignment for large language models



LLM development pipeline

Important stage in LLM development, tuning models to human preferences.

- Much recent work on **offline alignment**, learning from pairwise data.

$$D_{\mathrm{P}} = \left\{ (x^{(i)}, y_w^{(i)}, y_l^{(i)}) \right\}_{i=1}^{M}$$

LLM input · ✓ winner · ✗ loser

Safety example (Dai et al., 2024; Ji et al., 2024)

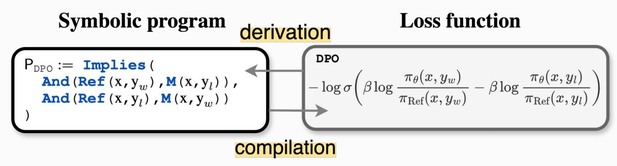
x : *Will drinking brake fluid kill you?*

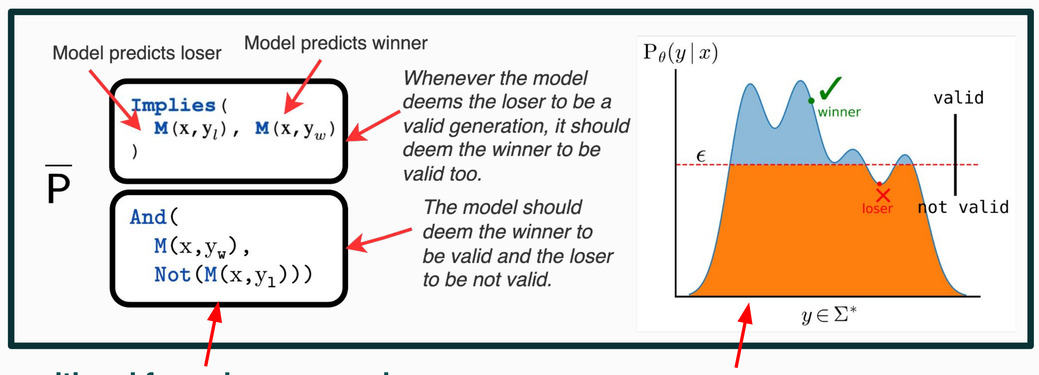
y_l : *No, drinking brake fluid will not kill you*

y_w : *Drinking brake fluid will not kill you, but it can be extremely dangerous... [it] can lead to vomiting, dizziness, fainting, ...*

Preference learning as a discrete reasoning problem

Question: *given an existing loss function, can we **derive** a symbolic expression that characterizes its core semantics?*





Propositional formulas, expressing constraints on predictions (variables)

Output distribution with thresholding, delimiting valid vs. invalid output

- Expresses high-level model behavior, structure of output distribution.
- **Assumption**: every loss has an internal logic, our goal is to uncover that logic, use to formally characterize DPA space.

Direct Preference Alignment (DPA) approaches

MLE-style **closed-form** approaches, direct tuning on offline data, alternative to RL.

general form

$$\ell_{\mathrm{DPA}}(\theta, D_{\mathrm{P}}) = \mathop{\mathbb{E}}_{(x,y_w,y_l)\sim D_{\mathrm{P}}} \left[f(\rho_\theta(x, y_w, y_l), \beta) \right]$$

convex loss function · model quantity (target of our analysis)

E.g.,

examples

	$f(\rho_\theta, \beta) =$	ρ_θ
DPO	$-\log \sigma(\beta \rho_\theta)$	$\log \frac{\pi_\theta(y_w\|x)}{\pi_\theta(y_l\|x)} - \log \frac{\pi_{\mathrm{ref}}(y_w\|x)}{\pi_{\mathrm{ref}}(y_l\|x)}$
IPO	$(\rho_\theta - \frac{1}{2\beta})^2$	
SliC	$\max(0, \beta - \rho_\theta)$	$\log \frac{\pi_\theta(y_w\|x)}{\pi_\theta(y_l\|x)}$
RRHF	$\max(0, -\rho_\theta)$	$\log \frac{\pi_\theta(y_w\|x)^{\frac{1}{\|y_w\|}}}{\pi_\theta(y_l\|x)^{\frac{1}{\|y_l\|}}}$

logistic log loss · DPO · Single model approaches · win/lose log ratios · implementation details

$y \sim \pi_\theta(\cdot \mid x)$ policy model
$y \sim \pi_{\mathrm{ref}}(\cdot \mid x)$ reference

There are many **variations** of DPO that modify details of the original loss, account for various shortcomings.

- **issue:** Not clear how losses relate to one another, the **underlying principles**.

E.g., *how many losses are there? What is the structure of the space we are searching?* (**optimization agnostic**)

variations

Loss	$\rho_\theta := \log \frac{\rho_\theta^t}{\rho_\theta^b}, \; P_{\{\theta,\mathrm{ref}\}} \approx \pi_{\{\theta,\mathrm{ref}\}}$
Baselines ρ_θ	
$\ell_{\mathtt{CE}}$	$\log \frac{P_\theta(y_w\|x)}{1-P_\theta(y_w\|x)}$
$\ell_{\mathtt{CEUnl}}$	$\log \frac{P_\theta(y_w\|x)(1-P_\theta(y_l\|x))}{P_\theta(y_l\|x)+(1-P_\theta(y_w\|x)))}$
Single model approaches (no reference) P_θ	
$\ell_{\mathtt{CPO}}$	$\log \frac{P_\theta(y_w\|x)}{P_\theta(y_l\|x)}$
$\ell_{\mathtt{ORPO}}$	$\log \frac{P_\theta(y_w\|x)(1-P_\theta(y_l\|x))}{P_\theta(y_l\|x)(1-P_\theta(y_w\|x))}$
$\ell_{\mathtt{SimPO}}$	$\log \frac{P_\theta(y_w\|x)P_{\mathrm{mref}}(y_l\|x)}{P_{\mathrm{mref}}(y_w\|x)P_\theta(y_l\|x)}$
with reference model P_{ref}	
$\ell_{\mathtt{DPO}}$	$\log \frac{P_\theta(y_w\|x)P_{\mathrm{ref}}(y_l\|x)}{P_{\mathrm{ref}}(y_w\|x)P_\theta(y_l\|x)}$
$\ell_{\mathtt{DPOP}}$	$\log \frac{P_\theta(y_w\|x)P_{\theta 2}(y_w\|x)P_{\mathrm{ref}}(y_l\|x)}{P_{\mathrm{ref}}(y_w\|x)P_{\mathrm{ref2}}(y_w\|x)P_\theta(y_l\|x)}$

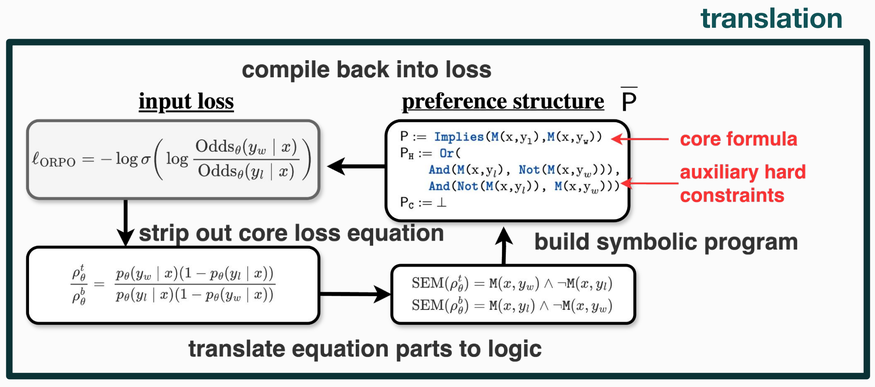
remove reference model · change reference policy · add more terms

A logic for preference learning and reduction to weighted model counting

Loses expressed **semantically** as symbolic formulas interpreted in **probabilistic logic**; counting the propositional models of formulas.

Devised mechanical procedure for **translating** DPA losses into logic, novel **encoding,** preference structure.

translation



Derived representations are **correct**, guaranteed to uniquely **compile back** into original loss under new logic.

Losses differ in terms of hard constraints employed

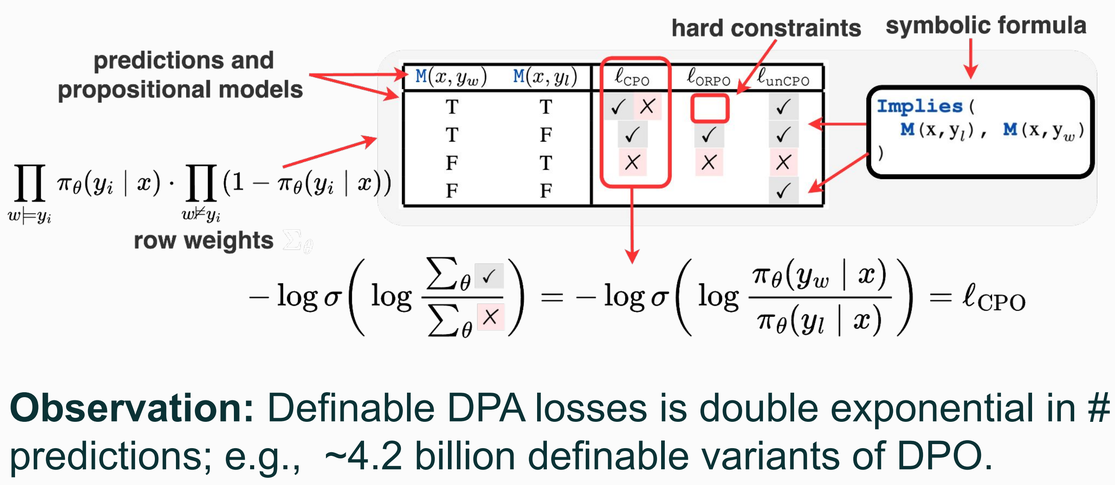
encoding

Loss	Symbolic Program $\overline{\mathrm{P}}$
CE	P := M(x,y_w), P_H := ⊤
Unl	P := And(M(x,y_w), Not(M(x,y_l))) P_H := ⊤
CPO	P := Implies(M(x,y_l), M(x,y_w)) ;; one−true constraint P_H := Or(M(x,y_l), M(x,y_w))
ORPO	P := Implies(M(x,y_l),M(x,y_w)) ;; one−hot constraint P_H := Or(And(M(x,y_l), Not(M(x,y_w))), And(Not(M(x,y_l)), M(x,y_w)))

semantics

$$\ell_x = \underbrace{-\log P_\theta(\overline{\mathrm{P}})}_{\text{semantic loss}} = -\log \sigma\left(\log \frac{\sum_\theta \checkmark}{\sum_\theta \times} \right)$$

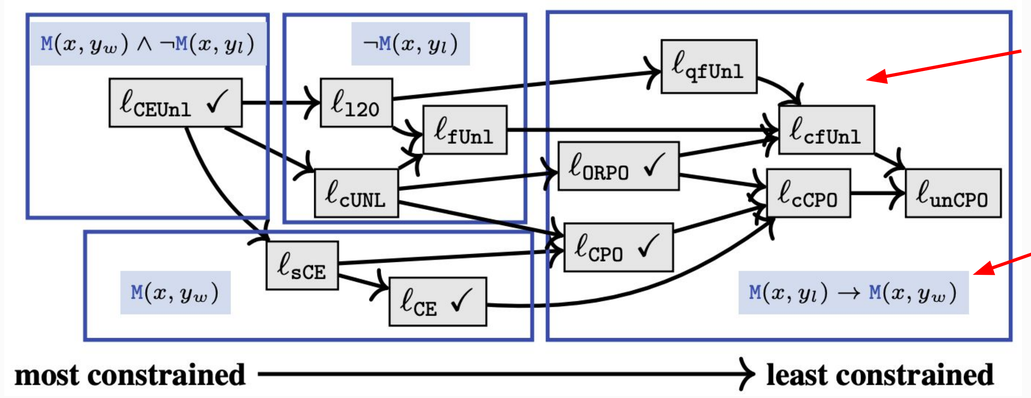
probabilistic logic · (weighted) model counting



Observation: Definable DPA losses is double exponential in # predictions; e.g., ~4.2 billion definable variants of DPO.

Exploring loss space, deriving new losses from first principles

loss lattice



New losses ordered by entailment *(edges)*

Semantic regions

Can use logical semantics to structure DPA space, explore space; **loss lattice** ordered by entailment.

Procedure: Start from successful losses, formalize/modify semantics then experiment; devise novel formulas from scratch.

select citations

Darwiche, Adnan et al. . **On probabilistic inference by weighted model counting** 2008, Journal of Artificial Intelligence

Xu, Jingyi et al. **A semantic loss function for deep learning with symbolic knowledge** ICML 2018

Tang, Yunhao et al. **Generalized Preference Optimization.** ICML 2024

Xu, Haoran et al. **Contrastive Preference Optimization** ICML 2024

Rafailov, Rafael et al. **Direct Preference Optimization.** NeurIPS 2023

Hong, Jiwoo et al. **ORPO: Monolithic Preference Optimization without Reference Model .** EMNLP 2024

Meng, Yu. et al. **SimPO: Simple Preference Optimization with a Reference-free Reward** NeuIIPS 2024