

Diverse Preference Learning for Capabilities and Alignment

Stewy Slocum, Asher Parker-Sartori, Dylan Hadfield-Menell

Motivation

Alignment algorithms like RLHF and DPO reduce LLM output diversity

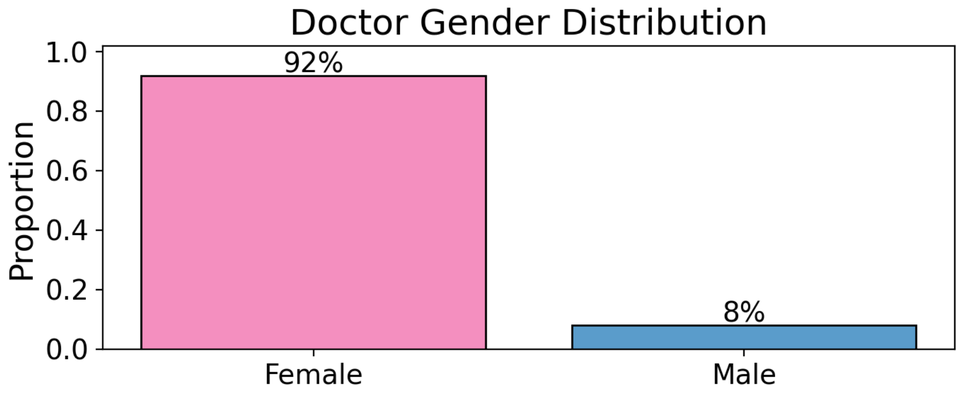
- Repetitive word choice, response structure
- Similar high-level ideas, societal perspectives
- This reduces diversity of joint human-LLM essays [Kobak 2024, Padkumar 2023]
- Does not occur for SFT models [Kirk 2023, Padkumar 2023]

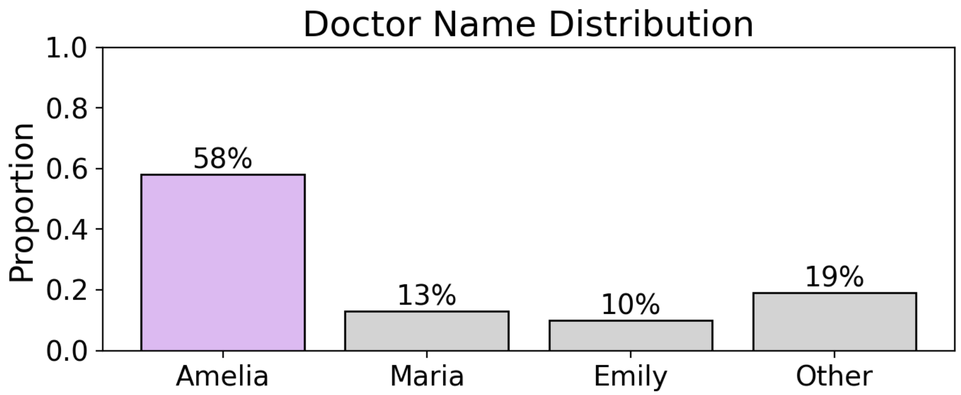
Prompt: Please write a two sentence story about a doctor and a close family member

Dr. Amelia watched with a heavy heart as **her** elderly father...

Dr. Amelia tended to **her** ailing father...

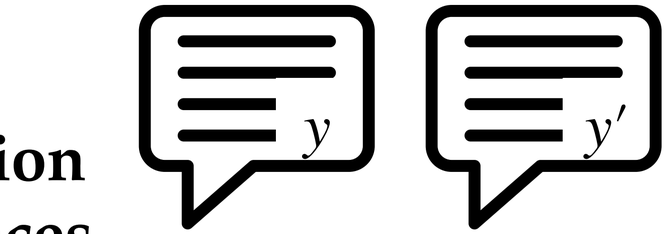
Dr. Amelia's heart sank as she held **her** ailing grandmother's hand...

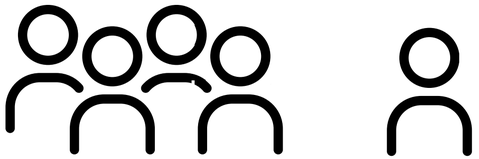




Algorithmic Alignment Group

Population Preferences





$p =$ 

Optimal SPL policy

$$\pi_{SPL}(y) \propto (\pi_{ref}(y)p^{1/\beta})^{\beta/\alpha}$$

RLHF

$$\pi_{RLHF}(y) \propto \pi_{ref}(y)p^{1/\beta}$$

SPL (Soft Preference Learning)

$$\max_{\pi} \mathbb{E}[r(y)] + \alpha H(\pi) - \beta H(\pi, \pi_{ref})$$

Optimal RLHF policy overweights majority preferences and sacrifices diversity

SPL entropy term α acts like temperature scaling at *sequence level* to control diversity

Token vs Sequence-Level Temperature Scaling

Prompt: Please write a two sentence story about a doctor and a close family member

DPO

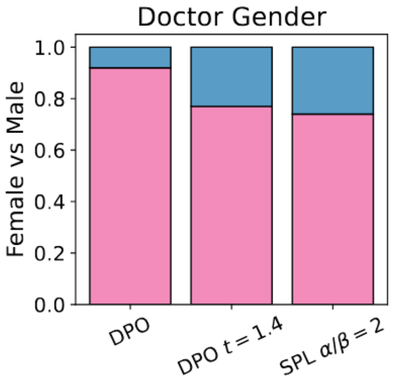
Dr. Amelia watched with a heavy heart as her elderly father, tears streaming down his face, admitted to no longer being able to care for himself at home. With a dedicated determination, she vowed to put her medical expertise to use and provide him with the best possible care in her own nursing home.

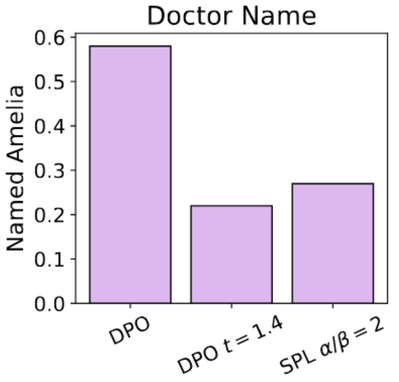
DPO $t = 1.4$

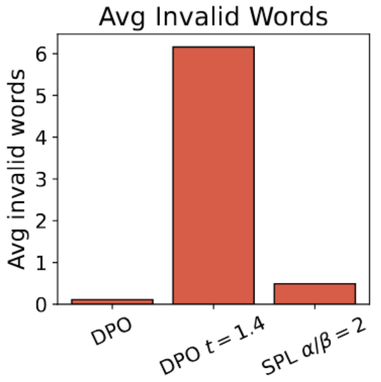
Dr. Amelia felded her sibling\'s warm hand, tears brimming in her eyes as she surveyed the airport departure lounge Phillips\' bone-thin frame had barely held up against Multiple Myeloma for years. Their embrace was an ignorant权dmermiment of infirm}[como handsome hero][COL shootColors(new Array manuscript Josh Goldman I L…

SPL $\alpha/\beta = 2$

Dr. Alanna sat quietly in her hospital office, the sun casting golden rays through the window over her quiet, motionless mother. Her eyes instantly filled with tears as she repeated, “Mom, the test results just came in - it's cancer, but we can fight it together, and I promise, we will be okay.”







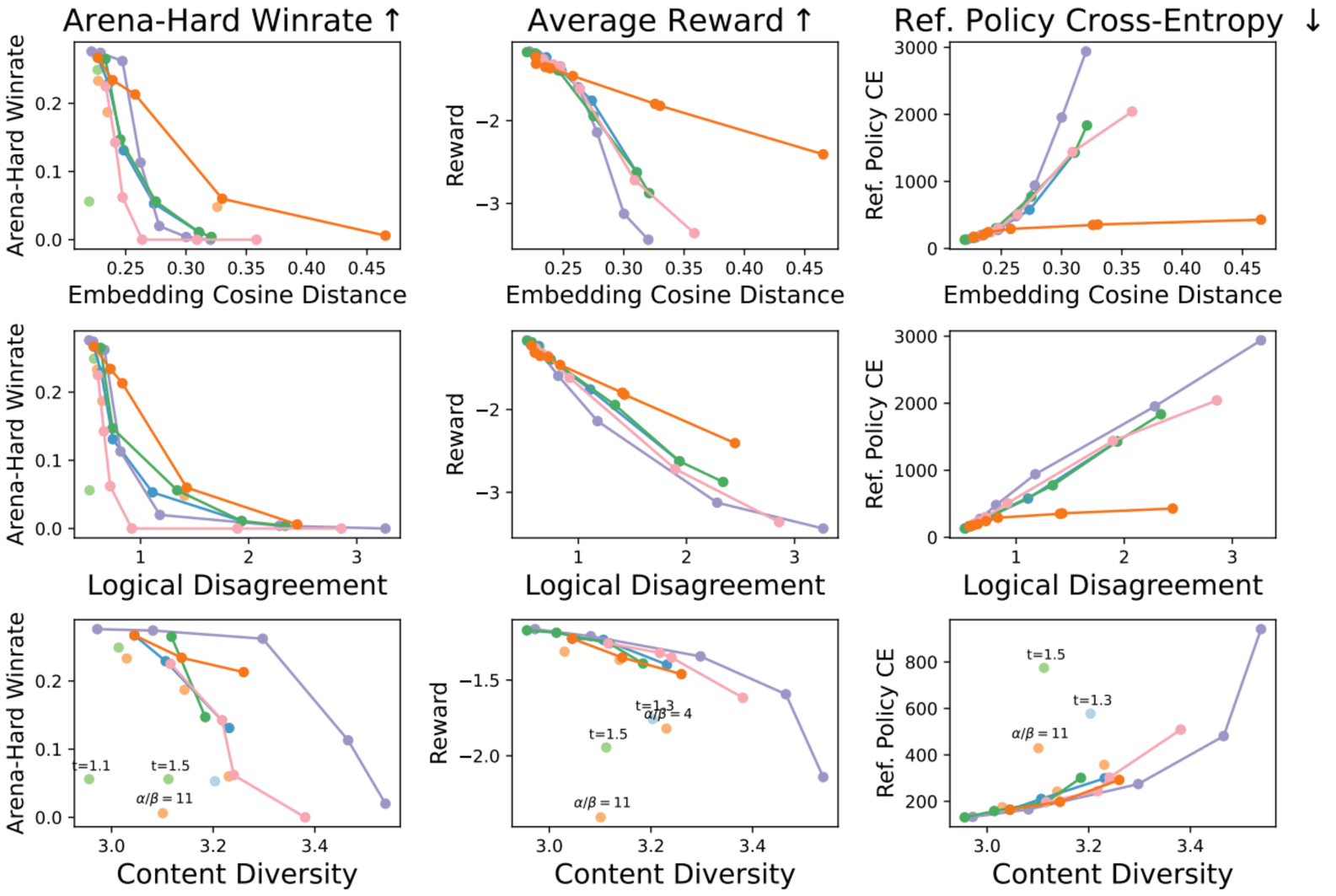
Algorithmic Alignment Group

Experiments — Diversity-Quality Tradeoff

Improved diversity-quality Pareto curve

Baselines: DPO with temperature scaling + top-p, top-k, and min-p sampling

DPO loses coherence, SPL doesn't



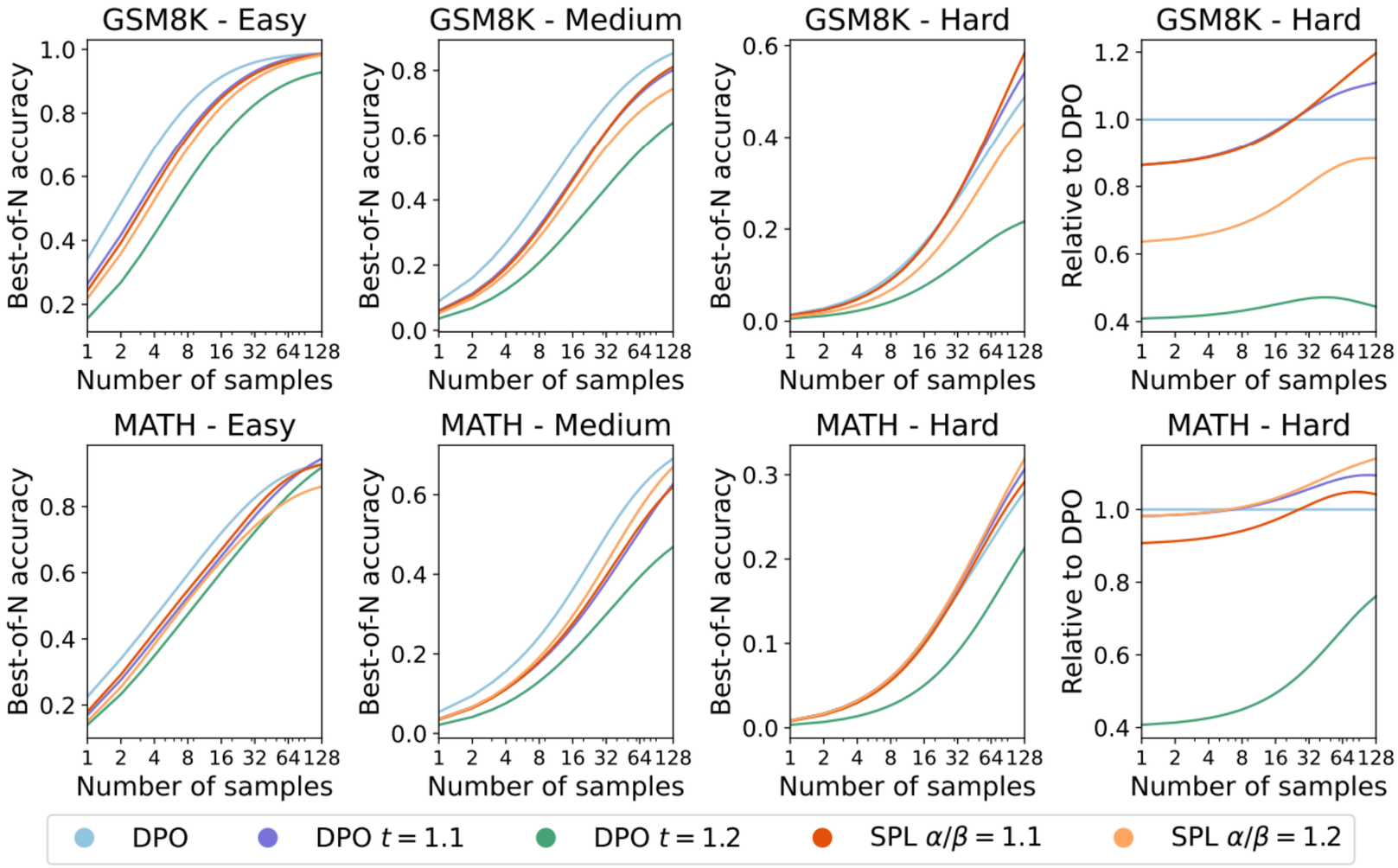
Algorithmic Alignment Group

Experiments — Inference-time Scaling, Calibration

Improved inference-time scaling (best-of-N) on challenging math datasets

- DPO → 30-40% duplicates
- SPL samples more diverse reasoning traces

Improved multiple-choice logit calibration while preserving accuracy



Algorithmic Alignment Group

Thank You

Stewart Slocum

stew@csail.mit.edu