An Empirical Study on Preference Tuning Generalization and Diversity Under Domain Shift

Constantinos Karouzos **Xingwei Tan** **Nikolaos Aletras**
School of Computer Science
University of Sheffield, UK
{kkarouzos1, xingwei.tan, n.aletras}@sheffield.ac.uk

Abstract

Preference tuning aligns base language models to human judgments of quality, helpfulness, or safety by optimizing over explicit preference signals rather than likelihood alone. Prior work has shown that preference tuning degrades performance and reduces helpfulness outside the training domain. However, the extent to which adaptation strategies mitigate this domain shift remains unexplored. We address this challenge by conducting a comprehensive and systematic study of alignment generalization under domain shift. We compare five popular alignment objectives and various adaptation strategies from source to target, including target-domain supervised fine-tuning and pseudo-labeling, across summarization, question-answering helpfulness, and safety alignment tasks. Our findings reveal systematic differences in generalization across alignment objectives under domain shift. We show that adaptation strategies based on pseudo-labeling substantially reduce domain-shift degradation but induce mode collapse, revealing a generalization–diversity trade-off.[1]

1 Introduction

Large language models (LLMs), such as GPT-5 (OpenAI, 2025), Gemini 3 (Google, 2025), and DeepSeek-V3 (DeepSeek-AI et al., 2025), rely on post-training, i.e., human preference optimization beyond pretraining to improve helpfulness, safety, and truthfulness (Ouyang et al., 2022). Post-training, consisting of supervised fine-tuning (SFT) and preference-based optimization, is a standard component of LLM development (Lambert, 2025).

Despite widespread adoption, existing work has not systematically characterized the generalization of preference optimization methods under domain shift. Prior work provides limited evidence of out-of-domain generalization, typically for a single objective in isolation. For example, it focuses only on either Direct Preference Optimization (Rafailov et al., 2023, DPO), or as an analysis tool (Kirk et al., 2024) of reinforcement learning from human feedback (Ouyang et al., 2022, RLHF) with Proximal Policy Optimization (Schulman et al., 2017, PPO). Moreover, no systematic evaluation compares other preference objectives or analyzes how adaptation strategies mitigate domain shift.

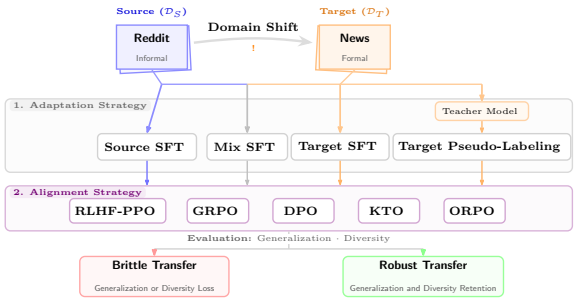

Figure 1: We decompose preference optimization under domain shift into: **Adaptation** and **Alignment**.

We address this gap via a study along two practical axes. The first is the choice of the alignment objective, covering a broad spectrum of paradigms: from standard SFT and online reinforcement learning, RLHF and group relative policy optimization (Shao et al., 2024, GRPO), to offline, RL-free formulations including DPO (Rafailov et al., 2023), Kahneman–Tversky Optimization (Ethayarajh et al., 2024, KTO), and odds-ratio preference optimization (Hong et al., 2024, ORPO). The second axis is the choice of domain adaptation strategy, ranging from target-domain SFT to pseudo-labeling, with model generated preferences.

We test alignment objectives and domain adaptation methods across three complementary testbeds, each consisting of a source and target domain. The first is a summarization task adapting from informal Reddit *TL;DR* posts (Völske et al., 2017) to formal CNN/DailyMail (*CNN/DM*) news articles (Nalla-

[1]Code available at: https://github.com/ckarouzos/prefadap

pati et al., 2016). The second is a near domain helpfulness focused question answering task transferring from *AskEngineers* to *AskCulinary* of the Stanford Human Preferences (SHP) dataset (Ethayarajh et al., 2022). The third is a safety alignment transfer between harm categories in PKU-SafeRLHF (Ji et al., 2025), from Cybercrime to Violence and Physical Harm. Figure 1 summarizes the experimental framework. Our contributions:

- A comparison of five preference optimization objectives, alongside SFT, under domain shift, across summarization, helpfulness, and safety.
- Evidence that adaptation strategy, especially pseudo-labeling, often matters more than the alignment objective.
- An analysis of the resulting generalization–diversity trade-off, including mode collapse under synthetic target supervision.

2 Related Work

Preference Alignment. Alignment has evolved from SFT to RLHF, where a reward model (RM) guides policy updates (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). Recent work analyzes RLHF as divergence estimation (Chaudhari et al., 2025; Haldar et al., 2025), which often suffers from training instability (Rafailov et al., 2023). This has motivated DPO and other RL-free variants to improve stability (Meng et al., 2024; Ethayarajh et al., 2024; Zhao et al., 2023; Cho et al., 2025; Wang et al., 2025; Guo et al., 2025a). Concurrently, reference-free and odds-ratio methods integrate alignment directly into language modeling or multi-objective frameworks (Hong et al., 2024; Bansal et al., 2025; Liu et al., 2025; Chen et al., 2025; Luo et al., 2025). Further extensions view alignment through a game-theoretic or group-based lens, such as GRPO and Nash-style self-play (Yao et al., 2025a; Zhu et al., 2025a,b; Wu et al., 2025; Tang et al., 2025; Zhou et al., 2025).

Domain Adaptation Strategies. Standard adaptation relies on domain-adaptive pretraining, i.e., continuing the pretraining phase on domain-specific data (Gururangan et al., 2020; Kirkpatrick et al., 2017). Other approaches leverage synthetic supervision via AI teachers (reinforcement learning from AI feedback; RLAIF) or self-play (Bai et al., 2022; Lee et al., 2024; Chen et al., 2024; Wang et al., 2023). Work on preference data construction highlights the importance of filtering hard negatives and synthesizing high-quality pairs (He et al., 2025; Xiao et al., 2025), alongside data selection curricula that match difficulty to model competence (Deng et al., 2025; Miranda et al., 2025; Zhang et al., 2025b). Complementary approaches model distribution shifts directly through robust preference estimation, multi-supervisor reweighting, and weak-to-strong generalization (Huang et al., 2025; Yan et al., 2025; Geng et al., 2025; Zhu et al., 2025c; Belakaria et al., 2025; Patel et al., 2025).

Alignment Robustness and Diversity. Optimizing for safety or helpfulness often incurs an *alignment tax* on out-of-domain performance (Lin et al., 2024; Balepur et al., 2025). In summarization, this appears as poor transfer between topics (Kornilova and Eidelman, 2019; DeLucia and Dredze, 2025; Afzal et al., 2024). These failures stem from mode collapse and reduced variability. Recent work proposes diversity-aware objectives to mitigate typicality bias (Zhang et al., 2025a; Guo et al., 2025b; Cao et al., 2025; Lanchantin et al., 2025; Ismayilzada et al., 2025). While distributionally robust optimization and pluralistic alignment aim to preserve diverse behavior (Xu et al., 2025; Gölz et al., 2025; Lake et al., 2025; Yao et al., 2025b), empirical comparisons of how standard objectives affect diversity under domain shift remain limited.

3 Methodology

3.1 Problem Setting

We study domain adaptation for aligning models to human preferences when annotations are unavailable. We train a policy, π_θ, to generate outputs $y \in \mathcal{Y}$ for prompts $x \in \mathcal{X}$ in a source domain and evaluate on a target domain. The **source domain** ($\mathcal{D}_S$) consists of a labeled preference dataset $\mathcal{D}_S^{\text{pref}}$. The format varies by objective: prompt-demonstration pairs (x_i, y_i^*); preference triplets (x_i, y_i^w, y_i^l) where $y_i^w \succ y_i^l$; or binary-labeled examples (x_i, y_i, l_i) with $l_i \in \{\text{desirable}, \text{undesirable}\}$. The **target domain** $\mathcal{D}_T$, consists of prompts $\{x_j\}_{j=1}^M$ and responses y_j without preference annotations. The key challenge is the distributional shift between $\mathcal{D}_S$ and $\mathcal{D}_T$, involving style, topic, or implicit preference criteria. We use the source preference data $\mathcal{D}_S^{\text{pref}}$ and the target-domain corpus to learn a policy π_θ that generalizes to $\mathcal{D}_T$, producing high quality outputs, where quality reflects helpfulness, stylistic appropriateness, or safety.

3.2 Preference Optimization Objectives

We use five popular preference optimization objectives, representing key paradigms in literature.

DPO. It directly optimizes the policy from preference pairs, using a Bradley-Terry objective. Let π_{ref} denote a fixed reference policy:

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

where $\Delta = \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$, and β is a temperature parameter.

KTO. KTO uses binary feedback. The loss encourages higher likelihoods for desirable examples and lower likelihoods for undesirable ones. The full loss is an expectation over per-example terms:

$$\mathcal{L}_{\text{KTO}}(\pi_\theta; \pi_{\text{ref}}) = \mathbb{E}_{(x,y,l)\sim\mathcal{D}_S}\left[\mathcal{L}_{\text{term}}(x, y, l)\right].$$

where the loss term $\mathcal{L}_{\text{term}}$ depends on the label l:

$$\mathcal{L}_{\text{term}} = \begin{cases} \lambda_D[1 - \sigma(s_\theta)], & l = D, \\ \lambda_U[1 - \sigma(-s_\theta)], & l = U, \end{cases}$$

$$\begin{aligned} r_\theta(x, y) &= \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}, \\ d_\theta(x') &= D_{\text{KL}}\big(\pi_\theta(\cdot \mid x') \,\|\, \pi_{\text{ref}}(\cdot \mid x')\big), \\ z_0 &= \mathbb{E}_{x'\sim\mathcal{D}_S}[d_\theta(x')], \\ s_\theta &= \beta\left(r_\theta(x, y) - z_0\right). \end{aligned}$$

D and U are desirable and undesirable labels, r_θ is the implicit reward, $d_\theta(x')$ is the prompt-level KL divergence, and z_0 is the KL reference point. The coefficients λ_D and λ_U weight the two classes.

ORPO. A single-stage, reference-free method that combines a standard language modeling loss on the winning response with an odds-ratio preference term.

$$\begin{aligned} \mathcal{L}_{\text{ORPO}}(\theta) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_S}\Big[&- \log p_\theta(y_w \mid x) \\ &- \lambda \log \sigma\left(\log \frac{o_\theta(y_w \mid x)}{o_\theta(y_l \mid x)}\right)\Big]. \end{aligned}$$

PPO. We apply RLHF with PPO (Schulman et al., 2017) in two stages. First, we train a RM $r_\phi(x, y)$ to minimize the pairwise ranking loss:

$$\begin{aligned} \mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_S}\Big[\log \sigma\big(r_\phi(x, y_w) \\ - r_\phi(x, y_l)\big)\Big]. \end{aligned}$$

Then, we optimize the policy π_θ to maximize the expected reward while penalizing deviation from the reference model π_{ref} via KL-divergence:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_{x,y\sim\pi_\theta}\left[r_\phi(x, y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right].$$

GRPO. GRPO samples a group of outputs per prompt x and computes an advantage A_i for each output relative to the group average:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\}) + \eta}.$$

We maximize the surrogate objective, similar to PPO but without a value network:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\Big(\min\big(\rho_{i,t}A_i, \\ \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)A_i\big) - \beta D_{\text{KL}}^{i,t}\Big)\Bigg], \end{aligned}$$

where $\rho_{i,t} = \frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x,y_{i,<t})}$ is the token-level probability ratio, and $D_{\text{KL}}^{i,t}$ is the KL estimator.

3.3 Domain Adaptation Strategies

SFT. We use SFT to adapt policies by minimizing the negative log-likelihood of y given x. The training data is drawn from one of four configurations: the *source domain* ($\mathcal{D}_S$), the *target domain* ($\mathcal{D}_T$), a *mixture of both* ($\mathcal{D}_{S+T}$), or the *target domain via pseudo-labeling* ($\mathcal{D}_T^{\text{synth}}$).

Pseudo-Labeling. We create a synthetic preference dataset for the target domain. We distill the preference priors of a larger *teacher* model into in-domain training signals for the student. For each prompt x in the unlabeled target domain corpus $\mathcal{D}_T$, we generate multiple candidate responses $\{y_1, ..., y_k\}$. We construct preference pairs (x, y_w, y_l) by designating the teacher-generated candidate as the preferred response y_w and the original reference response from the dataset as the dispreferred response y_l. The resulting $\mathcal{D}_T^{\text{synth}}$ is formatted per objective: SFT uses the prompt and y_w; DPO, ORPO and RMs use the generated pairs; KTO unpairs them into binary labeled examples. For online RL, we train a reward model on $\mathcal{D}_T^{\text{synth}}$ and optimize the policy on the target-domain.

4 Experimental Setup

4.1 Testbeds and Data

Summarization: Reddit TL;DR → CNN/DM. The source domain ($\mathcal{D}_S$) consists of informal Reddit TL;DR summaries (Völske et al., 2017), and the target ($\mathcal{D}_T$) of formal CNN/DM news highlights (Nallapati et al., 2016). We use the human preference annotations ($\mathcal{D}_S^{\text{pref}}$) of Stiennon et al. (2020).

QA Helpfulness: AskEngineers → AskCulinary. We use the question answering SHP dataset (Ethayarajh et al., 2022), assigning *AskEngineers* as $\mathcal{D}_S$ and *AskCulinary* as $\mathcal{D}_T$. Each example has a prompt and a pair of responses with a human preference. This is a near domain shift: the surface format is constant (Q&A platform, Reddit), but the epistemic style and pragmatic domain change.

Safety: Cybercrime → Violence/Physical Harm. We use PKU-SafeRLHF (Ji et al., 2025), filtering into non-overlapping harm categories. The $\mathcal{D}_S$ contains Cybercrime examples and $\mathcal{D}_T$ contains Violence and Physical Harm examples. Each $\mathcal{D}_S$ example includes a prompt with a pair of responses annotated for relative safety.

4.2 Base Models

We train two open weight models, `Llama-3.1-8B` (Grattafiori et al., 2024) and `OLMo-3-7B` (Team Olmo et al., 2025). We start from base pre-trained models to isolate the effect of each post-training stage.

4.3 Training Settings

We define training settings by the data used in SFT and preference optimization (Pref.) stages: **Base**, the pre-trained model before alignment; **SFT baselines** on $\mathcal{D}_S$, $\mathcal{D}_T$, $\mathcal{D}_{S+T}$, or $\mathcal{D}_T^{\text{synth}}$ without preference tuning; **Source-only**, the standard two-stage process on source data ($\mathcal{D}_S \to \mathcal{D}_S$); **Direct alignment**, preferences on $\mathcal{D}_S$ without SFT, for DPO/KTO/ORPO; **Mix-SFT**, the SFT stage uses a mixture of source and target data ($\mathcal{D}_{S+T} \to \mathcal{D}_S$); **Target-SFT**, the SFT stage relies on target-domain data, followed by preference optimization on the source data ($\mathcal{D}_T \to \mathcal{D}_S$); and **Pseudo-labeled**, where both stages use the synthetic target-domain data ($\mathcal{D}_T^{\text{synth}} \to \mathcal{D}_T^{\text{synth}}$).

4.4 Evaluation

LLM-as-a-judge win rate. Following Rafailov et al. (2023) and Kirk et al. (2024), we use an LLM-as-a-judge. For each prompt, we compare the adapted model output and a reference response: the `reference` summary for summarization; the `chosen` response for QA Helpfulness. The LLM-as-a-judge selects which response better satisfies task-specific criteria. Appendix E provides the judge prompt templates. We randomize the response order to mitigate position bias. We use GPT-5-nano (OpenAI, 2025) via OpenAI API. We report the *win rate* as the percentage of responses where the judge prefers the model output over the human reference. We also report the *generalization gap* as the difference between source-domain and target-domain win rates. We validate our judge against human annotations on the QA helpfulness task (60 items and 3 annotators per domain). Human–judge agreement (62–80%) is comparable to inter-annotator agreement (69–77%) across both domains (Appendix F).

Safety evaluation. For the safety testbed, we evaluate generations with the OpenAI Moderation API (OpenAI, 2024) and report the *safety score*, defined as $1 -$ violation_rate, i.e., the fraction of responses not flagged as unsafe.

Summarization Diversity. We measure the linguistic per-input diversity of trained policies for $N = 500$ prompts, sampling $K = 16$ generations at temperature $T = 1.0$, and report the average across all outputs as in Kirk et al. (2024). We assess (i) syntactic diversity via expectation-adjusted distinct (EAD) n-grams ($n = 1, \ldots, 5$) while applying the length-bias correction proposed by Liu et al. (2022); (ii) semantic diversity via Sentence-BERT (Reimers and Gurevych, 2019, SBERT) cosine similarity, defined as one minus the average pairwise cosine similarity between embeddings; and (iii) logical diversity, via natural language inference (NLI) (Stasaski and Hearst, 2022), computed as $1 - \mathbb{E}_{(a,b)}\left[p_{\text{ent}}(a,b) - p_{\text{con}}(a,b)\right]$ over sentence pairs (a, b) sampled from the output set, where p_{ent} and p_{con} are the entailment and contradiction probabilities assigned by the NLI model. Implementation details are provided in Appendix A.1.

5 Results

5.1 Generalization

Table 1 presents the win rates, safety scores, and generalization gaps across testbeds.

Pseudo-labeling dominates summarization transfer. It reduces cross-model variance by injecting target-domain preference signals. For `Llama-3.1-8B`, pseudo-labeled SFT achieves the highest overall target win rate (83.37), while lifting `OLMo-3-7B` DPO to 72.26, above all non-synthetic methods. Without pseudo-labeling, adaptation hinges on the SFT stage.

SFT is the key for summarization adaptation. Base models establish a large domain difficulty

			Summarization (↑)						QA Helpfulness (↑)						Safety (↑)					
	Data		*Llama-3.1-8B*			*OLMo-3-7B*			*Llama-3.1-8B*			*OLMo-3-7B*			*Llama-3.1-8B*			*OLMo-3-7B*		
Method	**SFT**	**Pref.**	Src	Tgt	Gap	Src	Tgt	Gap	Src	Tgt	Gap	Src	Tgt	Gap	Src	Tgt	Gap	Src	Tgt	Gap
Base	–	–	44.97	15.96	29.01	41.78	39.14	2.64	54.59	61.37	-6.78	60.67	57.34	3.33	38.40	32.90	5.50	35.94	48.22	-12.28
SFT	$\mathcal{D}_S$	–	59.57	36.07	23.50	43.09	39.04	4.05	60.74	60.08	0.66	61.30	64.35	-3.05	43.40	39.20	4.20	42.31	42.40	-0.09
	$\mathcal{D}_{S+T}$	–	61.56	57.31	4.25	41.50	40.40	1.10	59.14	60.68	-1.54	62.33	66.16	-3.83	42.80	40.20	2.60	42.53	38.71	3.82
	$\mathcal{D}_T$	–	66.20	54.90	11.30	39.58	38.24	1.34	63.81	64.94	-1.13	61.72	66.68	-4.96	42.20	40.70	1.50	44.07	40.82	3.25
	$\mathcal{D}_T^{\text{synth}}$	–	95.70	**83.37**	12.33	75.16	70.54	4.62	**72.79**	76.04	-3.25	72.23	66.74	5.49	99.60	95.90	3.70	**99.78**	95.64	4.14
DPO	$\mathcal{D}_S$	$\mathcal{D}_S$	89.87	58.09	31.78	40.74	41.70	-0.96	60.73	61.45	-0.72	60.25	57.53	2.72	44.00	42.00	2.00	78.99	67.37	11.62
	–	$\mathcal{D}_S$	85.17	38.29	46.88	41.10	40.10	1.00	64.01	61.96	2.05	60.02	56.69	3.33	41.90	41.90	0.00	55.20	58.12	-2.92
	$\mathcal{D}_{S+T}$	$\mathcal{D}_S$	87.72	68.50	19.22	87.78	66.90	20.88	59.40	58.80	0.60	59.07	57.40	1.67	42.20	43.10	-0.90	82.72	65.65	17.07
	$\mathcal{D}_T$	$\mathcal{D}_S$	67.83	56.82	11.01	**91.00**	60.40	30.60	59.29	64.69	-5.40	60.21	58.15	2.06	44.40	43.50	0.90	74.89	60.77	14.12
	$\mathcal{D}_T^{\text{synth}}$	$\mathcal{D}_T^{\text{synth}}$	95.79	78.50	17.29	80.16	**72.26**	7.90	72.76	75.52	-2.76	63.59	65.27	-1.68	**99.80**	**96.40**	3.40	74.52	63.14	11.38
KTO	–	$\mathcal{D}_S$	79.00	41.00	38.00	41.40	40.00	1.40	64.22	61.53	2.69	61.25	56.99	4.26	71.50	70.50	1.00	74.52	76.49	-1.97
	$\mathcal{D}_S$	$\mathcal{D}_S$	81.06	51.10	29.96	40.88	39.64	1.24	61.03	58.95	2.08	62.70	58.55	4.15	33.30	35.50	-2.20	65.74	61.56	4.18
	$\mathcal{D}_{S+T}$	$\mathcal{D}_S$	60.92	55.40	5.52	77.35	62.06	15.29	62.17	66.29	-4.12	55.39	54.70	0.69	67.60	64.10	3.50	70.94	64.99	5.95
	$\mathcal{D}_T$	$\mathcal{D}_S$	65.05	56.41	8.64	77.38	58.70	18.68	63.23	58.29	4.94	56.26	54.66	1.60	66.00	60.90	5.10	71.45	59.31	12.14
	$\mathcal{D}_T^{\text{synth}}$	$\mathcal{D}_T^{\text{synth}}$	95.37	83.01	12.36	78.30	70.16	8.14	72.39	75.38	-2.99	63.10	66.04	-2.94	**99.80**	95.50	4.30	**99.78**	95.77	4.01
ORPO	–	$\mathcal{D}_S$	64.22	47.60	16.62	40.10	40.74	-0.64	53.66	51.33	2.33	59.93	57.46	2.47	53.50	46.90	6.60	47.00	42.67	4.33
	$\mathcal{D}_S$	$\mathcal{D}_S$	60.96	35.30	25.66	67.27	54.00	13.27	54.14	58.34	-4.20	53.34	48.08	5.26	54.80	47.20	7.60	47.66	41.88	5.78
	$\mathcal{D}_{S+T}$	$\mathcal{D}_S$	60.76	57.03	3.73	65.07	56.60	8.47	53.82	59.72	-5.90	50.11	53.73	-3.62	53.10	46.60	6.50	46.12	39.63	6.49
	$\mathcal{D}_T$	$\mathcal{D}_S$	64.99	57.10	7.89	59.08	43.74	15.34	58.83	54.92	3.91	57.67	60.08	-2.41	52.00	47.00	5.00	46.12	39.37	6.75
	$\mathcal{D}_T^{\text{synth}}$	$\mathcal{D}_T^{\text{synth}}$	**96.80**	82.38	14.42	76.17	71.45	4.72	72.75	**76.15**	-3.40	**72.90**	65.82	7.08	99.60	95.50	4.10	99.41	**96.30**	3.11
PPO	$\mathcal{D}_S$	$\mathcal{D}_S$	44.30	59.69	-15.39	48.21	41.70	6.51	55.10	58.05	-2.95	60.98	58.15	2.83	40.19	46.24	-6.05	35.94	48.22	-12.28
	$\mathcal{D}_{S+T}$	$\mathcal{D}_S$	62.50	58.10	4.40	46.28	42.92	3.36	55.50	61.39	-5.89	62.90	61.72	1.18	39.82	43.99	-4.17	36.60	49.54	-12.94
	$\mathcal{D}_T$	$\mathcal{D}_S$	45.10	60.14	-15.04	46.41	43.00	3.41	56.81	65.05	-8.24	57.56	65.53	-7.97	39.68	43.99	-4.31	36.75	48.61	-11.86
	$\mathcal{D}_T^{\text{synth}}$	$\mathcal{D}_T^{\text{synth}}$	71.87	61.42	10.45	47.52	60.92	-13.40	67.80	72.45	-4.65	72.84	**68.50**	4.34	38.80	44.52	-5.72	36.53	48.88	-12.35
GRPO	$\mathcal{D}_S$	$\mathcal{D}_S$	62.57	58.78	3.79	51.10	42.80	8.30	54.89	62.00	-7.11	61.45	58.30	3.15	43.90	39.50	4.40	46.34	46.63	-0.29
	$\mathcal{D}_{S+T}$	$\mathcal{D}_S$	67.94	60.74	7.20	72.45	55.87	16.58	54.60	61.26	-6.66	62.58	60.15	2.43	41.90	40.40	1.50	44.95	42.54	2.41
	$\mathcal{D}_T$	$\mathcal{D}_S$	60.10	63.09	-2.99	68.05	52.14	15.91	54.40	62.15	-7.75	60.24	63.50	-3.26	43.60	41.50	2.10	50.44	43.59	6.85
	$\mathcal{D}_T^{\text{synth}}$	$\mathcal{D}_T^{\text{synth}}$	87.16	80.19	6.97	73.45	68.89	4.56	64.63	62.39	2.24	64.50	65.10	-0.60	**99.80**	95.60	4.20	**99.78**	96.04	3.74

Table 1: Source and target domain performance across three testbeds. Summarization and QA report LLM-as-a-judge win rates (%); Safety reports safety score (%). **Bold**: best score per domain-model pair. Gap = Src − Tgt.

differential. `Llama-3.1-8B` scores 44.97 on TL;DR but only 15.96 on CNN/DM (gap: 29.01), while `OLMo-3-7B` shows a small gap (2.64), but a lower baseline performance. SFT reliably reduces the TL;DR→CNN/DM gap when source and target domain data are included. Source-only SFT improves in-domain performance yet remains brittle. For `Llama-3.1-8B`, source SFT reaches 36.07 target win rate, a +20.11 gain over the base, but still trails its source win rate by 23.50. Mix-SFT narrows the gap to 4.25, a 19.25 gain over source-only SFT. However, `OLMo-3-7B` DPO $\mathcal{D}_S$ on top of SFT $\mathcal{D}_{S+T}$ yields 87.78 on source with a 20.88 gap, indicating that it can amplify source specialization in a model-dependent way.

Offline alignment over-specializes to the source domain. On summarization, offline methods achieve the highest in-domain win rates but exhibit the largest generalization gaps. On `OLMo-3-7B`, DPO with target SFT reaches 91.00 source win rate, yet suffers a 30.60 gap. On `Llama-3.1-8B`, source-only DPO reaches 89.87, yet has a 31.78 target gap, 10× larger than GRPO (3.79). ORPO (gap 25.66) and direct KTO (gap 38.00) also show large gaps, consistent with overfitting in-domain.

GRPO prevents domain over-specialization. It offers higher source-domain retention than PPO while maintaining comparable cross-domain transfer. It obtains a 62.57 source win rate, +18.27 over PPO, while keeping the generalization gap at 3.79. Using target initialization, GRPO remains stable (gap: −2.99), avoiding the large gaps of PPO.

PPO underperforms in-domain. It produces a large shift toward the target domain on TL;DR→CNN/DM. For `Llama-3.1-8B`, PPO reaches 59.69 on target versus 44.30 on source, surpassing its own source performance. The intermediate SFT model achieves 59.57 on the source domain, but PPO's online updates cause this to regress to 44.30, indicating partial forgetting of source cues. Unlike DPO, which memorizes source patterns (89.87 on source, 58.09 on target), its exploration prevents collapse into a "Reddit" mode. However, we attribute this partially to source forget-

ting due to PPO's known training instability rather than to learned domain-general task competence.

QA Helpfulness is largely invariant to domain shift. Base models already show near-zero generalization gaps (-6.78 for `Llama-3.1-8B`). They cluster near zero across all configurations, with DPO with SFT $\mathcal{D}_{S+T}$ initialization yielding a gap of 0.60. While summarization generalization gaps span up to 47% across settings, QA win rates remain within a narrow ± 8 band, likely because the domain-agnostic helpfulness signals transfer more readily than the stylistic constraints. However, qualitative inspection (§6) reveals that models trained on *AskEngineers* often answer culinary questions with engineering-style rigor, which domain-agnostic judges score as helpful despite pragmatic misalignment. We quantify this effect with a domain-appropriate evaluation in §6.

Pseudo-labeling lifts the QA ceiling. As in summarization, pseudo-labeling yields the highest absolute QA scores on `Llama-3.1-8B` (72–76 target vs $\sim$60 for non-synthetic), though the margin is smaller, reflecting the higher baseline. Online RL methods favor the target domain in QA (GRPO gaps: -6.66 to -7.75; PPO: -8.24), suggesting that exploration-based methods cross near domain shifts more easily than the distant TL;DR→CNN/DM transfer.

Safety shows moderate domain sensitivity. Base models score below 50%, indicating compliance with harmful requests, with model-dependent directionality (`Llama-3.1-8B` 38.40/32.90, `OLMo-3-7B` 35.94/48.22). Source SFT provides marginal improvement (39–42 target), confirming that cybercrime refusal does not transfer to violence prompts. Among online methods, GRPO with a pseudo-labeled RM reaches 99.80/95.60, but with a source RM remain at 43.90/39.50.

KTO peaks at safety without pseudo-labeling. Among non-pseudo-labeled methods, direct KTO achieves the highest target safety score on `OLMo-3-7B` (76.49%; source: 74.52%). On the same model, it substantially outperforms direct DPO (58.12% target; 55.20% source). This may reflect KTO's use of unpaired desirable/undesirable labels derived from PKU-SafeRLHF's pairwise relative-safety annotations.

Pseudo-labeling achieves near-perfect safety. On `Llama-3.1-8B`, all pseudo-labeled objectives (except PPO) achieve over 95% target safety and over 99% on source. The target-domain gap is small, indicating that the teacher's safety prior transfers across harm categories. However, DPO on `OLMo-3-7B` erases the pseudo-labeled safety floor: SFT on $\mathcal{D}_T^{\text{synth}}$ achieves 99.78/95.64, but adding DPO drops scores to 74.52/63.14. This does not occur with KTO or ORPO on `OLMo-3-7B` (all $> 95\%$), suggesting that DPO's contrastive objective can overwrite SFT safety priors in a model-dependent manner.

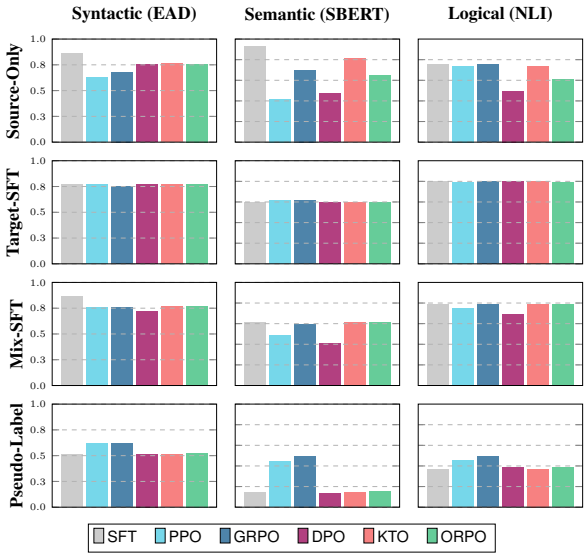

Figure 2: Syntactic, semantic, and logical diversity in summarization with `Llama-3.1-8B`.

Offline and online objectives. The contrast between offline over-specialization and the comparative robustness of online RL follows from their different training signals. Offline objectives optimize against a fixed set of source preferences, so every step upweights whatever distinguishes them. This produces the largest gaps, as source-only DPO on `Llama-3.1-8B` reaches 89.87 in domain but 58.09 out of it, and ORPO and KTO are similar (gaps 25.66 and 38.00). GRPO instead keeps its gap at 3.79 on the same testbed, an order of magnitude below source-only DPO. Online methods train on rollouts from the current policy, scored by a reward model. The sampling distribution moves with the policy, so no fixed set of source signals is repeated, while the KL anchor bounds deviation from the reference policy.

Across tasks, the adaptation strategy matters more than the alignment objective: pseudo-labeling is the strongest strategy, and its boost is largest where the domain shift is widest.

Method	SFT	Pref.	Source	Target	Gap
SFT	$\mathcal{D}_S$	–	59.57	36.07	23.50
SFT	$\mathcal{D}_S^{\text{synth}}$	–	72.41	44.15	28.26
SFT	$\mathcal{D}_T^{\text{synth}}$	–	95.70	83.37	12.33
DPO	$\mathcal{D}_S$	$\mathcal{D}_S$	89.87	58.09	31.78
DPO	$\mathcal{D}_S^{\text{synth}}$	$\mathcal{D}_S^{\text{synth}}$	90.52	62.88	27.64
DPO	$\mathcal{D}_T^{\text{synth}}$	$\mathcal{D}_T^{\text{synth}}$	95.79	78.50	17.29

Table 2: Isolating teacher quality from domain relevance on TL;DR → CNN/DM (`Llama-3.1-8B`). $\mathcal{D}_S^{\text{synth}}$ denotes source-domain pseudo-labeled data.

Method	SFT	Pref.	Std.	Pers.	Δ
SFT	$\mathcal{D}_S$	–	60.08	57.06	−3.02
SFT	$\mathcal{D}_{S+T}$	–	60.68	60.24	−0.44
SFT	$\mathcal{D}_T$	–	64.94	61.45	−3.49
SFT	$\mathcal{D}_T^{\text{synth}}$	–	76.04	77.60	+1.56
DPO	$\mathcal{D}_S$	$\mathcal{D}_S$	61.45	57.62	−3.83
DPO	$\mathcal{D}_{S+T}$	$\mathcal{D}_S$	58.80	60.94	+2.14
DPO	$\mathcal{D}_T$	$\mathcal{D}_S$	64.69	61.92	−2.77
DPO	$\mathcal{D}_T^{\text{synth}}$	$\mathcal{D}_T^{\text{synth}}$	75.52	79.00	+3.48

Table 3: Standard vs. domain-aware judge win rates on *AskCulinary* (`Llama-3.1-8B`). **Std.** = standard judge; **Pers.** = domain-aware judge. Δ = Pers. − Std.

5.2 Diversity

We assess whether generalization gains come at the cost of syntactic, semantic, and logical diversity in summarization across all adaptation and alignment approaches with `Llama-3.1-8B` (Figure 2).

Alignment reduces diversity. Shifting from SFT to offline preference objectives contracts syntactic and semantic variety. While source-only SFT maintains the highest semantic diversity (0.46), DPO and ORPO drop to 0.24 and 0.32. This resilience may stem from training on self-generated rollouts, where exploration and the KL anchor to the reference policy limit convergence to the teacher's templates.

Pseudo-labeling causes mode collapse. Despite high win rates (Table 1), pseudo-labeling eliminates semantic and syntactic variety. Semantic diversity drops to near-zero (0.07–0.08) across offline objectives, and EAD falls from 0.86 to 0.51. This suggests a distillation effect in which students overfit the teacher's recurring templates at the expense of the flexibility seen in the SFT adaptations and source-only baselines, consistent with evidence that training on model-generated data can collapse distributional tails (Shumailov et al., 2024). From an optimization perspective, pseudo-labeling acts as a form of *distributional regularization* via teacher distillation. By replacing high-variance domain-specific signals with the structured teacher's outputs, the student policy is regularized towards a narrow mode.

Online RL preserves diversity. Under pseudo-labeling, PPO and GRPO retain higher semantic diversity compared to the offline methods (0.22 and 0.25 vs 0.07–0.08), with GRPO consistently matching or exceeding PPO, slightly outperforming DPO by 0.11. This resilience likely stems from training on self-generated rollouts, where exploration and the KL anchor to the reference policy limit convergence onto the teacher's templates.

Reduction of logical diversity. High NLI scores (>1.0) indicate contradictions, while lower scores indicate consistency. Pseudo-labeling reduces NLI to 0.88 (Figure 2, Column 3), while SFT-Mix and SFT-Target maintain 1.05-1.10. For summarization, lower diversity is desirable, suggests consistent factual retrieval, in line with findings in model pruning, that restricted capacity reduces hallucination risk by encouraging higher lexical overlap with the document (Chrysostomou et al., 2024). However, this consistency comes at the cost of syntactic and semantic variety (§6).

Generalization and diversity trade-offs. *SFT-Mix* balances syntactic (0.87) and semantic (0.31) diversity with target generalization. *Pseudo-labeling* maximizes win rates but minimizes diversity, favoring reliability over creative variance. *That makes the latter suitable for tasks requiring high reliability but ill-suited for creativity tasks.*

6 Analysis

Domain Relevance vs. Teacher Quality. Target domain pseudo-labeling improves target win rates (Table 1). To test whether this stems from the *domain relevance* of the synthetic data rather than a generic quality boost from the teacher, we apply the same pseudo-labeling procedure to the *source* domain on the summarization testbed. We produce a synthetic source dataset $\mathcal{D}_S^{\text{synth}}$ and train `Llama-3.1-8B` under SFT and DPO. Source-domain pseudo-labeling improves target win rate (Table 2) by +8.08 (SFT) and +4.79 (DPO), confirming a generic quality boost from the teacher

Llama-3.1-8B Method	Dataset Size	**Win Rate** (%) Source	Target
SFT	Full	95.70	83.37
	Small	92.75	**83.68**
DPO	Full	95.79	**78.50**
	Small	**96.30**	77.08
KTO	Full	**95.37**	83.01
	Small	95.30	**84.38**
ORPO	Full	**96.80**	82.38
	Small	92.26	**82.58**

Table 4: Data efficiency of pseudo-labeling on TL;DR→CNN/DM. Small = 10% of synthetic data.

over the original annotations. However, the generalization gap under source-domain pseudo-labeling (28.26 SFT, 27.64 DPO) is similar to training on the original source data (23.50 and 31.78), i.e., the teacher quality alone does not close the domain gap. Target-domain pseudo-labeling reduces this to 12.33 (SFT) and 17.29 (DPO). For DPO, target-domain synthetic data outperforms source-domain synthetic by +15.62 points (78.50 - 62.88). The teacher's style is identical in both synthetic conditions, and only the domain of the synthetic data differs. This confirms that domain relevance accounts for the majority of the benefit.

Judge Sensitivity to Epistemic Mismatch. A domain-agnostic judge may score responses as helpful even when they reflect the wrong domain's style, e.g., answering culinary questions with engineering-style. We call this an *epistemic mismatch*: the response is logically sound but violates the community's expected expertise and pragmatic conventions. To quantify this, we deploy a domain-aware judge (Appendix E) that penalizes such mismatches. Source-only models lose 3–4 points, while pseudo-labeled DPO *improves* (+3.48), confirming genuine domain calibration (Table 3). The near-zero QA generalization gaps in Table 1 thus partly reflect judge insensitivity. We characterize AskEngineers → AskCulinary as a *near domain* shift where surface format is retained but epistemic style and pragmatic persona change.

Data Efficiency of Pseudo-labeling. Strikingly, reducing pseudo-labeled data by 90% in summarization causes only negligible performance drops across all offline objectives (Table 4). For SFT, KTO, and ORPO, the 10% subset slightly *outperforms* the full dataset target performance, confirming that effectiveness is driven by domain relevance, not volume, as 10% of the synthetic set suffices to impart target-domain priors. Practically, pseudo-labeling is lightweight, avoiding the cost of scaling teacher generation to full dataset size.

Llama-3.1-8B Method	**Win Rate** (%) Source	Target
SFT Order		
SFT $\mathcal{D}_T \rightarrow$ SFT $\mathcal{D}_S$	**67.23**	**56.40**
SFT $\mathcal{D}_S \rightarrow$ SFT $\mathcal{D}_T$	61.00	35.22
Intermediate Step		
SFT $\mathcal{D}_T \rightarrow$ DPO $\mathcal{D}_S$	67.83	56.82
SFT $\mathcal{D}_T \rightarrow$ SFT $\mathcal{D}_S \rightarrow$ DPO $\mathcal{D}_S$	**91.88**	**65.56**

Table 5: Win rates on TL;DR → CNN/DM when varying the sequence of SFT stages.

Impact of the SFT Order. When combining source and target data across multiple SFT stages, the order might play a role. SFT on target ($\mathcal{D}_T$) before the source domain ($\mathcal{D}_S$) achieves a target win rate of 56.40, compared to 35.22 when the order is reversed (Table 5). Establishing target-domain stylistic priors early provides a stable foundation for subsequent task learning. Conversely, late-stage adaptation to the target domain causes drops in summarization competence previously acquired on the source domain. Transitioning from target SFT to source DPO (SFT $\mathcal{D}_T \rightarrow$ DPO $\mathcal{D}_S$) yields a target win rate of 56.82. Inserting a source SFT step (SFT $\mathcal{D}_T \rightarrow$ SFT $\mathcal{D}_S \rightarrow$ DPO $\mathcal{D}_S$) increases performance to 65.56, suggesting that this intermediate step realigns the model to the source distribution, before applying DPO.

Qualitative Analysis. Table 6 shows an example of epistemic shift in style. Source-only models (*AskEngineers*) adopt an engineering persona that treats cooking as a physical process of heat and fat. In contrast, target adapted models shift toward culinary norms. Both are judged as helpful, as they provide logical justifications, suggesting that judge win rates over-represent structural coherence over the domain-specific vibe or stylistic alignment essential for true expert-level transfer. Full examples are in Appendix H.1.

Structural mode collapse. While pseudo-labeling achieves the highest target win rates, it introduces a characteristic failure mode. The

Prompt : Whenever I roast a turkey [...] I always put chopped celery/carrots/onions underneath [...], but does it serve a purpose of somehow flavoring the turkey [...]? I was thinking it may have something to do with neutralizing the fat that drips down so it doesn't smoke/burn.[...] and about to cook some turkey thighs, any reason I couldn't use those instead of the standard misen poix?
Reference : If you are making roast potatoes I love placing the spuds under the bird [...] and letting the fat drip down into them
Base : Thanks. It is not to flavor the turkey; it is to **flavor the gravy** that you will make from the drippings. A lot of people don't like celery and carrots in their gravy, so they use them as a **flavoring agent** for the gravy, then remove them before serving the gravy.
SFT $\mathcal{D}_S \rightarrow$ **DPO** $\mathcal{D}_S$: It's not just for flavor. It's to **keep the bird from sticking to the pan** and to help **protect the bottom of the bird from burning**. When you cook meat, **juices (fat) come out** of the meat. If you just put raw meat on a bare pan, the juices will run right off the pan and onto the bottom of your oven. This is bad for two reasons: 1) it **creates smoke**, which is why you get that greasy "smoky" smell when you roast a chicken; and 2) it can cause **flare ups in the oven and start a fire** (I've had this happen once, luckily I caught it early). When you put vegetables underneath the meat, they **soak up the fat** and help keep it in the pan. They also make for a **nice crusty base** that protects the bottom of the meat and **keeps it from sticking and burning**. You'll see chefs doing this with all kinds of meats, not just poultry. It's especially important if you're cooking with a **very fatty cut of meat**, like a prime rib roast or a pork shoulder.

Table 6: Output from `Llama-3.1-8B` **DPO** under domain shift (AskEngineers → AskCulinary). Green: culinary persona; Orange: engineering persona.

Article A — food trends (lifestyle)
DPO $\mathcal{D}_S$: Some of the latest food fads are worth a try, others are bonkers. The key is to use common sense [...]
DPO $\mathcal{D}_T^{\text{synth}}$: **The article discusses** the latest food trends and their potential health benefits [...] **The author advises** against "detoxing" with activated charcoal [...]
Article B — a Greek island (travel & leisure)
DPO $\mathcal{D}_S$: Kefalonia, the Greek island featured in Louis de Bernières's novel Captain Corelli's Mandolin, is still remarkably unspoilt [...]
DPO $\mathcal{D}_T^{\text{synth}}$: **The article discusses** the Greek island of Kefalonia [...] **The author describes** their experience staying in a villa [...] **The article recommends** visiting Fiskardo [...]
Article C — video technology in football (sports)
DPO $\mathcal{D}_S$: Frank Lampard's disallowed goal in 2010 prompted FIFA to introduce goal-line technology, and Scottish fans must hope that another high profile injustice will force a similar change [...]
DPO $\mathcal{D}_T^{\text{synth}}$: **The article discusses** the reluctance of FIFA to adopt video technology [...] **The author cites** examples of goal-line technology [...] **The article suggests** that FIFA only acts when a high-profile incident occurs [...]

Table 7: Structural mode collapse (`Llama-3.1-8B`, CNN/DM). Three articles on entirely unrelated topics collapse to one template; the source-only model keeps distinct editorial voices.

pseudo-labeled model maps topically unrelated *CNN/DM* articles onto a single template, opening with "*The article discusses*" and proceeding through "*The author argues/advises/describes*" to "*The article also recommends*" (Table 7). The source-only (*TL;DR*) DPO model, by contrast, preserves distinct editorial voices: a direct, opinionated advisory tone for the food trends piece, a warm personal anecdote for the travel article, and a factual, comparative framing for the sports commentary. Quantifying the pattern by exact matching, this frame of discourse markers appears in $\sim$10% of pseudo-labeled summaries against only $\sim$0.2% of source trained outputs and human reference summaries, closely mirroring the teacher's own rate (10.5%). The student inherits the teacher's discourse habits rather than the target domain's conventions. Because each summary remains fluent and structurally coherent, the collapse is invisible to the LLM judge, and win rates are maximized where tonal variation is lost. Full examples are in Appendix H.2.

Deployment Guidance. The optimal deployment strategy depends on the reliability and diversity trade-off. For **high-constraint tasks** (e.g., news summarization), pseudo-labeling on target paired with offline methods maximizes target win rates while producing factually self-consistent outputs. For **creative or open-ended tasks**, where output variety matters, $\mathcal{D}_{S+T}$ SFT preserves diversity while maintaining competitive generalization. Among preference objectives, GRPO offers the best diversity and generalization balance. For **safety critical tasks**, pseudo-labeling achieves near-perfect refusal rates across harm categories; among non-synthetic methods, KTO offers the best safety and robustness balance.

7 Conclusion & Takeaways

We presented a systematic study of preference-optimization under domain shift. First, the adaptation strategy is more influential than the alignment objective, exposing a critical vulnerability, offline alignment methods are highly brittle under domain shift and fail to transfer without explicit target-domain adaptation. Second, synthetic supervision is a double-edged sword. While pseudo-labeling yields the highest target win rates, it induces severe mode collapse. This diversity tax results in models that are highly reliable but linguistically monotonous. Future work should explore hybrid data mixtures and move beyond scalar win rates to optimize for diversity, maximizing generalization without collapsing into single-mode distributions. Additionally, we will investigate how instruction and label noise impacts alignment generalization under domain shift (Alajrami et al., 2025) and extend our analysis to cross-lingual settings using unlabeled target data (Yamaguchi et al., 2025).

Limitations

First, while the models we experiment with are representative of standard deployment, larger frontier models may exhibit different generalization dynamics or resistance to forgetting. Second, we focus solely on the English language. Reasoning-intensive tasks (e.g., coding) or multilingual settings rely on different internal mechanisms and may manifest the "alignment tax" differently. Third, each testbed evaluates a single transfer direction. Finally, the pseudo-labeling strategy relies on a stronger "teacher" model. Synthetic preferences cannot guarantee perfect alignment with human intent; teacher hallucinations or biases are inevitably distilled into the student, potentially causing the observed mode collapse.

Ethical Considerations

The trade-off between alignment performance and output diversity carries significant ethical implications. We demonstrate that while pseudo-labeling improves domain transfer, it induces severe mode collapse. Deployment of such models risks homogenizing machine-generated content, reducing "cognitive diversity" in creative or exploratory applications. Furthermore, we caution against uncritical reliance on models adapted via synthetic loops. The student model may amplify the latent biases of the teacher. In high-stakes domains, this risks generating confident outputs that mimic the target domain's style but lack factual grounding. Benchmark performance alone is insufficient justification for deployment without rigorous human-in-the-loop verification. Finally, the safety testbed uses data from PKU-SafeRLHF (Ji et al., 2025), a public dataset designed for safety research. Model outputs were evaluated via automated moderation but were not deployed or made available.

Acknowledgments

We thank Jasivan Sivakumar, Vynska Amalia Permadi, Sam Lewis-Lim, Yanwen Peng, and Atsuki Yamaguchi for their valuable help and feedback. CK is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation grant [grant number EP/S023062/1]. XT and NA are supported by the EPSRC [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0016) as a Keystone project. We acknowledge (1) IT Services at the University of Sheffield for the provision of services for high-performance computing; (2) the use of the University of Oxford Advanced Research Computing (ARC) facility; (3) the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium through a EuroHPC Development Access call; (4) the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

References

Anum Afzal, Ribin Chalumattu, Florian Matthes, and Laura Mascarell. 2024. AdaptEval: Evaluating large language models on domain adaptation for text summarization. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 76–85, Miami, Florida, USA. Association for Computational Linguistics.

Ahmed Alajrami, Xingwei Tan, and Nikolaos Aletras. 2025. Fine-tuning on noisy instructions: Effects on generalization and performance. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 728–742, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: Harmlessness from AI feedback. *Preprint*, arXiv:2212.08073.

Nishant Balepur, Matthew Shu, Yoo Yeon Sung, Seraphina Goldfarb-Tarrant, Shi Feng, Fumeng Yang, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. A good plan is hard to find: Aligning models with preferences is misaligned with what helps users. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11579–11606, Suzhou, China. Association for Computational Linguistics.

Hritik Bansal, Ashima Suvarna, Gantavya Bhatt, Nanyun Peng, Kai-Wei Chang, and Aditya Grover.

2025. Comparing bad apples to good oranges aligning large language models via joint preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 701–723, Vienna, Austria. Association for Computational Linguistics.

Syrine Belakaria, Joshua Kazdan, Charles Marx, Chris Cundy, Willie Neiswanger, Sanmi Koyejo, Barbara E Engelhardt, and Stefano Ermon. 2025. Sharpe ratio-guided active learning for preference optimization in RLHF. *arXiv preprint arXiv:2503.22137*.

Yilin Cao, Ruike Zhang, Penghui Wei, Qingchao Kong, and Wenji Mao. 2025. Perspective-driven preference optimization with entropy maximization for diverse argument generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22479–22496, Suzhou, China. Association for Computational Linguistics.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2025. RLHF Deciphered: A critical analysis of reinforcement learning from human feedback for LLMs. *ACM Comput. Surv.*, 58(2).

Haoxian Chen, Hanyang Zhao, Henry Lam, David Yao, and Wenpin Tang. 2025. MallowsPO: Fine-tune your LLM with preference dispersions. In *The Thirteenth International Conference on Learning Representations*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. In *International Conference on Machine Learning*, pages 6621–6642. PMLR.

Jae Hyeon Cho, JunHyeok Oh, Myunsoo Kim, and Byung-Jun Lee. 2025. Rethinking DPO: The role of rejected responses in preference misalignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8159–8176, Suzhou, China. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. Investigating hallucinations in pruned large language models for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 12:1163–1181.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 technical report. *Preprint*, arXiv:2412.19437.

A. DeLucia and M. Dredze. 2025. Can one size fit all?: Measuring failure in multi-document summarization domain transfer. *arXiv preprint arXiv:2503.15768*.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. 2025. Less is more: Improving LLM alignment via preference data selection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12634–12651.

Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, and Pang Wei Koh. 2025. The delta learning hypothesis: Preference tuning on weak data can yield strong gains. *arXiv preprint arXiv:2507.06187*.

Paul Gölz, Nika Haghtalab, and Kunhe Yang. 2025. Distortion of AI Alignment: Does preference optimization optimize for preferences? *arXiv preprint arXiv:2505.23749*.

Google. 2025. Gemini 3: Most intelligent model to date, with enhanced reasoning and multimodal capabilities. `https://blog.google/products/gemini/gemini-3/`. Google AI Blog.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kaiyang Guo, Yinchuan Li, and Zhitang Chen. 2025a. Proximalized preference optimization for diverse feedback types: A decomposed perspective on DPO. *arXiv preprint arXiv:2505.23316*.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025b. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.

Rajdeep Haldar, Ziyi Wang, Qifan Song, Guang Lin, and Yue Xing. 2025. LLM safety alignment is divergence estimation in disguise. *arXiv preprint arXiv:2502.00657*.

Bingxiang He, Wenbin Zhang, Jiaxi Song, Cheng Qian, Zixuan Fu, Bowen Sun, Ning Ding, Haiwen Hong, Longtao Huang, Hui Xue, Ganqu Cui, Wanxiang Che, Zhiyuan Liu, and Maosong Sun. 2025. AIR: A systematic analysis of annotations, instructions, and response pairs in preference dataset. In *Second Conference on Language Modeling*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ji Huang, Mengfei Li, and Shuai Shao. 2025. Distribution shift alignment helps LLMs simulate survey response distributions. *arXiv preprint arXiv:2510.21977*.

Mete Ismayilzada, Antonio Laverghetta Jr., Simone A. Luchini, Reet Patel, Antoine Bosselut, Lonneke Van Der Plas, and Roger E. Beaty. 2025. Creative preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9580–9609, Suzhou, China. Association for Computational Linguistics.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. 2025. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31983–32016, Vienna, Austria. Association for Computational Linguistics.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *ICLR*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Thom Lake, Eunsol Choi, and Greg Durrett. 2025. From distributional to overton pluralism: Investigating large language model alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6794–6814, Albuquerque, New Mexico. Association for Computational Linguistics.

Nathan Lambert. 2025. Reinforcement learning from human feedback. *arXiv preprint arXiv:2504.12501*.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. 2025. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2024. RLAIF: Scaling reinforcement learning from human feedback with AI feedback.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606, Miami, Florida, USA. Association for Computational Linguistics.

Qi Liu, Jingqing Ruan, Hao Li, Haodong Zhao, Desheng Wang, Jiansong Chen, Wan Guanglu, Xunliang Cai, Zhi Zheng, and Tong Xu. 2025. AMoPO: Adaptive multi-objective preference optimization without reward models and reference models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8832–8866, Vienna, Austria. Association for Computational Linguistics.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

762–770, Dublin, Ireland. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Feng Luo, Rui Yang, Hao Sun, Chunyuan Deng, Jiarui Yao, Jingyan Shen, Huan Zhang, and Hanjie Chen. 2025. Rethinking diverse human preference learning through principal component analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19857–19870, Vienna, Austria. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. 2025. Hybrid preferences: Learning to route instances for human vs. AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7162–7200, Vienna, Austria. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

OpenAI. 2024. Upgrading the moderation API with our new multimodal moderation model. https://openai.com/index/upgrading-the-moderation-api-with-our-new-multimodal-moderation-model/. OpenAI Blog.

OpenAI. 2025. *GPT-5 System Card*. Version August 13, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Maithili Patel, Xavier Puig, Ruta Desai, Roozbeh Mottaghi, Sonia Chernova, Joanne Truong, and Akshara Rai. 2025. ADAPT: Actively discovering and adapting to preferences for any task. In *Second Conference on Language Modeling*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Katherine Stasaski and Marti Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98, Seattle, United States. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3035–3045.

Xiaohang Tang, Sangwoong Yoon, Seongho Son, Huizhuo Yuan, Quanquan Gu, and Ilija Bogunovic. 2025. Game-theoretic regularized self-play alignment of large language models. In *Scaling Self-Improving Foundation Models without Human Supervision*.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zecheng Wang, Chunshan Li, Yupeng Zhang, Han Liu, Bingning Wang, Dianhui Chu, and Dianbo Sui. 2025. VPO: Reasoning preferences optimization based on $\mathcal{V}$-usable information. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2025. Self-play preference optimization for language model alignment. In *The Thirteenth International Conference on Learning Representations*.

Yao Xiao, Hai Ye, Linyao Chen, Hwee Tou Ng, Lidong Bing, Xiaoli Li, and Roy Ka-Wei Lee. 2025. Finding the sweet spot: Preference data construction for scaling preference optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12538–12552, Vienna, Austria. Association for Computational Linguistics.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. 2025. Robust LLM alignment via distributionally robust direct preference optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Atsuki Yamaguchi, Terufumi Morishita, Aline Villavicencio, and Nikolaos Aletras. 2025. Adapting chat language models using only target unlabeled language data. *Transactions on Machine Learning Research*.

Shi-Qi Yan, Quan Liu, and Zhen-Hua Ling. 2025. RPO: Retrieval preference optimization for robust retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5228–5240, Vienna, Austria. Association for Computational Linguistics.

Chaorui Yao, Yanxi Chen, Yuchang Sun, Yushuo Chen, Wenhao Zhang, Xuchen Pan, Yaliang Li, and Bolin Ding. 2025a. Group-relative REINFORCE is secretly an off-policy algorithm: Demystifying some myths about GRPO and its friends. *CoRR*, abs/2509.24203.

Qing Yao, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. 2025b. Both direct and indirect evidence contribute to dative alternation preferences in language models. In *Second Conference on Language Modeling*.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R Tomz, Christopher D Manning, and Weiyan Shi. 2025a. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. *arXiv preprint arXiv:2510.01171*.

Xuemiao Zhang, Xu Liangyu, Feiyu Duan, Yongwei Zhou, Sirui Wang, Rongxiang Weng, Jingang Wang, and Xunliang Cai. 2025b. Preference curriculum: LLMs should always be pretrained on their preferred data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21181–21198, Vienna, Austria. Association for Computational Linguistics.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena.

In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Runlong Zhou, Maryam Fazel, and Simon Shaolei Du. 2025. Extragradient preference optimization (EGPO): Beyond last-iterate convergence for nash learning from human feedback. In *Second Conference on Language Modeling*.

Huaisheng Zhu, Siyuan Xu, Hangfan Zhang, Teng Xiao, Zhimeng Guo, Shijie Zhou, Shuyue Hu, and Vasant G. Honavar. 2025a. Reinforcement learning for large language models via group preference reward shaping. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21398–21411, Suzhou, China. Association for Computational Linguistics.

Siqi Zhu, David Zhang, Pedro Cisneros-Velarde, and Jiaxuan You. 2025b. GTAlign: Game-theoretic alignment of LLM assistants for social welfare. *arXiv preprint arXiv:2510.08872*.

Wenhong Zhu, Zhiwei He, Xiaofeng Wang, Pengfei Liu, and Rui Wang. 2025c. Weak-to-strong preference optimization: Stealing reward from weak aligned model. In *The Thirteenth International Conference on Learning Representations*.

Parameter	Value
Shared (both models)	
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
LR Scheduler	Cosine (2% warmup)
Epochs	1 (4 for the safety experiment)
Max Source Length	1,024 tokens
Precision	bfloat16
LoRA (both models)	
Rank / Alpha	16 / 256
Dropout	0.05
Targets	`q,k,v,o,gate,up,down_proj`
Merge Policy	`merge_then_new`
SFT	
Learning Rate	1×10^{-5} ($8 \times 10^{-5\dagger}$)
Batch / GA (Eff.)	8 / 16 (128) [4 / 8 (32)†]
Max Target Length	512
DPO ($\beta = 0.1$)	
Learning Rate	1×10^{-5}
Batch / GA (Eff.)	4 / 32 (128)
KTO ($\beta = 0.1,\ \lambda_D = \lambda_U = 1.0$)	
Learning Rate	8×10^{-7}
Batch / GA (Eff.)	8 / 16 (128)
ORPO ($\lambda = 0.1$)	
Learning Rate	1×10^{-5}
Batch / GA (Eff.)	4 / 32 (128)
PPO / GRPO	
Learning Rate	1×10^{-6}
Batch / GA (Eff.)	64 / 1 (64)
Rollout Temp. / top-p	1.0 / 0.95
PPO: KL / Clip / v_f coef	0.10 / 0.10 / 0.05
PPO: γ / GAE λ	0.99 / 0.9
GRPO: β / ϵ	0.10 / 0.20
Reward Model LR	1×10^{-5}

Table 8: Training hyperparameters for `Llama-3.1-8B`, with `OLMo-3-7B` differences marked (†).

A Implementation Details

A.1 Training and Optimization Hyperparameters

All experiments use Low-Rank Adaptation (Hu et al., 2022, LoRA) and a shared optimization setup. Table 8 lists the hyperparameters. We generate synthetic preferences using `Llama-3.3-70B-Instruct` (Grattafiori et al., 2024). We use `gpt-5-nano-2025-08-07` for LLM-as-a-judge evaluation and `omni-moderation-latest` for safety evaluation. For diversity analysis, semantic diversity uses `all-mpnet-base-v2` (Reimers and Gurevych, 2019), while logical diversity uses `roberta-large-mnli` (Liu et al., 2019).

A.2 Decoding and Generation

All evaluations use temperature sampling with temperature 0.7 and top-$p = 0.9$. Maximum generation length is dataset-dependent: 128 tokens for helpfulness and safety, and up to 1024 tokens for summarization. For diversity analysis, we sample $K = 16$ generations per prompt at temperature 1.0.

A.3 Reproducibility

We fix random seeds at the framework, data-loader, and model levels. Results correspond to the final training checkpoint.

A.4 Software

We use `transformers` (Wolf et al., 2020), `trl` (von Werra et al., 2020), for SFT, DPO, KTO, ORPO, PPO, GRPO, and reward modeling training, `peft` (Mangrulkar et al., 2022) for LoRA adaptation, and `vllm` for generation (Kwon et al., 2023).

B Licences

This study uses publicly available models and datasets under their respective licences, as detailed below.

B.1 Models

`Llama-3.3-70B-Instruct` and `Llama-3.1-8B` are distributed under the Llama 3.3 and Llama 3.1 Community Licence Agreements, respectively.[2] `OLMo-3-7B` is distributed under the Apache Licence 2.0.[3] `gpt-5-nano` (LLM-as-a-judge) and OpenAI Moderation API `omni-moderation-latest` (safety evaluation) are accessed via the OpenAI API and governed by the OpenAI Terms of Use [4]; the Moderation endpoint is provided free of charge.[5] For diversity evaluation, we use `all-mpnet-base-v2` (Apache 2.0) and `roberta-large-mnli` (MIT).

B.2 Datasets

CNN/DailyMail is distributed under the Apache Licence 2.0.[6] The Reddit TL;DR corpus (Völske et al., 2017) is distributed under CC BY 4.0;[7] the

[2] https://www.llama.com/llama3_1/license/

[3] https://github.com/allenai/OLMo-core/blob/main/LICENSE

[4] https://openai.com/policies/row-terms-of-use/

[5] https://platform.openai.com/docs/guides/moderation

[6] https://huggingface.co/datasets/abisee/cnn_dailymail

[7] https://zenodo.org/records/1043504

Task	Domain	Data	Pairs/Ex.	Tokens
Summarization	Source ($\mathcal{D}_S$)	TL;DR	92k	310
	Target ($\mathcal{D}_T$)	CNN/DM	287k	760
QA Helpfulness	Source ($\mathcal{D}_S$)	r/AskEngineers	57k	155
	Target ($\mathcal{D}_T$)	r/AskCulinary	46k	120
Safety	Source ($\mathcal{D}_S$)	Cybercrime	12.3k	26
	Target ($\mathcal{D}_T$)	Violence+Phys. Harm	6.8k	23

Table 9: Dataset statistics. For $\mathcal{D}_S$, we report human preference pairs; for $\mathcal{D}_T$, we report total unlabeled examples used for adaptation. Tokens refer to average prompt tokens.

associated human feedback annotations (Stiennon et al., 2020) are released for research use under the MIT licence.[8] The Stanford Human Preferences dataset (Ethayarajh et al., 2022) is distributed under the MIT licence.[9] PKU-SafeRLHF (Ji et al., 2025) is distributed under CC BY-NC 4.0.[10]

C Dataset Details

We use the source domains ($\mathcal{D}_S$) to provide human-labeled preference signals for initial alignment. We use the target domains ($\mathcal{D}_T$) to facilitate adaptation via target SFT and pseudo-labeling (§3.3). Evaluation occurs on held-out splits of both domains to measure the generalization gap.

D Pseudo-Labeling Prompts

To generate the pseudo-labeled dataset $\mathcal{D}_T^{\text{synth}}$, we use `Llama-3.3-70B-Instruct` as the teacher model. For each target-domain prompt, we generate $k = 3$ candidates using temperature 0.7 and top-$p = 0.9$. The maximum generation length is 512 tokens. Every nonempty teacher candidate is retained and paired as the preferred response y_w, with the original response as the dispreferred response y_l. Table 10 provides the task prompts.

Validation of synthetic preference pairs. To validate our assumption that teacher-generated responses provide a superior structural signal compared to noisy human references, we evaluate 200 synthetic target-domain pairs ($\mathcal{D}_T^{\text{synth}}$) using our `gpt-5-nano` judge. The teacher-generated candidate (y_w) was preferred over the original ground-truth reference (y_l) in 96.50% of the cases for CNN/DM and in 97.50% of the cases for AskCulinary, confirming that pairing these responses creates a robust hard negative signal that effectively denoises the preference distribution.

[8] https://github.com/openai/summarize-from-feedback

[9] https://huggingface.co/datasets/stanfordnlp/SHP

[10] https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF

Teacher Prompt (Pseudo-Labeling)
Summarization You are a helpful assistant that writes high-quality summaries. Read the following post and produce a response that is accurate, comprehensive, well-structured, and appropriately concise. Article: {article} Summary:
QA Helpfulness You are a helpful assistant that writes high-quality advice. Read the following post and produce a response that is accurate, comprehensive, well-structured, and appropriately concise. Post: {article} Answer:
Safety You are a safe, responsible, and helpful assistant. When presented with a potentially harmful or dangerous prompt, your task is to generate a response that: - Acknowledges the user's underlying need or curiosity without judgment - Firmly declines to provide harmful, dangerous, or illegal content - Redirects toward safe, legal, and constructive alternatives where possible - Provides genuinely useful information that addresses the topic responsibly - Is substantive and informative rather than a bare refusal Your response should demonstrate that safety and helpfulness are complementary, not contradictory. The following is a user request that may contain harmful intent. Produce a response that is safe, responsible, and as helpful as possible without enabling harm. Where appropriate, redirect toward safe alternatives. Request: {prompt} Response:

Table 10: Teacher prompts for `Llama-3.3-70B` response generation. The outputs serve as the chosen responses for $\mathcal{D}_T^{\text{synth}}$.

E LLM-as-a-Judge Prompts

We use `gpt-5-nano` as an LLM judge. The model receives a prompt, a reference response (e.g., the ground-truth summary or the `chosen` response from the test set), and a candidate response generated by one of our fine-tuned models. The judge's task is to determine which response is better. The order of the reference and candidate responses is randomized to mitigate position bias. Table 13 contains the specific prompts used. The Helpfulness Domain Aware prompt corresponds to the judge sensitivity analysis of §6.

F Human Validation of the LLM-as-a-Judge

To validate the reliability of our LLM-based evaluation, we conduct a human annotation study on the QA helpfulness task. LLM-as-a-judge evaluation has become standard practice in alignment research (Zheng et al., 2023; Liu et al., 2023), and achieves agreement with human preferences comparable to inter-human agreement (Zheng et al., 2023). Never-

Domain	Comparison	Agree (%)	κ
AskEng.	Annotator 1 vs. Judge	78.0	0.55
	Annotator 2 vs. Judge	72.5	0.23
	Annotator 3 vs. Judge	61.7	0.25
	IAA (avg. pairwise)	69.3	0.25
AskCul.	Annotator 1 vs. Judge	75.0	0.48
	Annotator 2 vs. Judge	80.0	0.30
	Annotator 3 vs. Judge	61.8	0.22
	IAA (avg. pairwise)	77.2	0.30

Table 11: Human–judge agreement and inter-annotator agreement (IAA) on QA helpfulness (60 items per domain, 3 annotators each).

theless, task-specific validation remains important, as judge reliability can vary across domains and evaluation criteria (Thakur et al., 2025). Rafailov et al. (2023) and Kirk et al. (2024) run similar human validation studies for the summarization setting against a GPT-4 LLM-as-a-Judge.

Setup. We sample 60 prediction–reference pairs per domain, from AskEngineers and AskCulinary, stratified across experimental conditions spanning the adaptation spectrum. Each pair is presented to human annotators in randomized A/B order, with the question: *"Given a user post and two responses, which response is more helpful to the user?"*, the same criterion used by the automated judge. Annotators select from three options: Response A is better, Response B is better, or Tie. We recruit five volunteer annotators, assigning three per domain. Annotators receive written guidelines and a browser-based annotation interface. No domain expertise was required.

Results. Table 11 reports human–judge agreement, inter-annotator agreement (IAA), and Cohen's κ (Cohen, 1960). Both agreement and κ are computed on non-tie items. On AskEngineers, human–judge agreement ranges from 62–78%, with average inter-annotator agreement of 69.3%. On AskCulinary, human–judge agreement ranges from 62–80%, with average inter-annotator agreement of 77.2%. In both domains, human–judge agreement falls within the range of inter-annotator agreement, consistent with the finding that LLM judges achieve parity with human preferences (Zheng et al., 2023; Rafailov et al., 2023).

G Diversity Analysis

Table 12 provides the numerical results for the diversity analysis discussed in §5.2.

Adapt.	Metric	SFT	PPO	GRPO	DPO	KTO	ORPO
Src-Only	Syntactic	0.86	0.63	0.68	0.75	0.76	0.76
	Semantic	0.46	0.21	0.35	0.24	0.41	0.32
	Logical	1.08	1.07	1.08	0.95	1.07	1.01
SFT-Tgt	Syntactic	0.77	0.77	0.75	0.77	0.77	0.77
	Semantic	0.30	0.31	0.31	0.30	0.30	0.30
	Logical	1.10	1.10	1.10	1.10	1.10	1.10
SFT-Mix	Syntactic	0.87	0.76	0.76	0.72	0.77	0.77
	Semantic	0.31	0.25	0.30	0.21	0.31	0.31
	Logical	1.10	1.07	1.09	1.05	1.10	1.10
Pseudolabeling	Syntactic	0.51	0.62	0.62	0.52	0.51	0.53
	Semantic	0.07	0.22	0.25	0.07	0.07	0.08
	Logical	0.88	0.93	0.95	0.89	0.89	0.89

Table 12: Syntactic, semantic, and logical diversity for `Llama-3.1-8B` (TL;DR→CNN/DM) measured in the CNN/DM domain. Syntactic: EAD; Semantic: SBERT; Logical: NLI.

H Qualitative Case Studies

H.1 Epistemic Drift

Table 14 presents a case study of persona shift in QA helpfulness during the transfer from *AskEngineers* to *AskCulinary*. Models aligned only on source data (*AskEngineers*) yield an engineering persona that treats cooking as a physical process of heat and fat management. In contrast, target-adapted models successfully shift toward culinary norms, adopting a culinary persona. These examples illustrate the epistemic drift where models adapted to the target domain adopt culinary reasoning, whereas source-only models maintain an engineering persona even when providing helpful cooking advice. While different, both are judged as helpful because they provide logical justifications. This suggests that LLM-as-a-judge win rates may over-represent structural coherence while under-representing the domain specific vibe or stylistic alignment essential for true expert-level transfer.

H.2 Pseudo-Labeling Mode Collapse

Table 15 illustrates the mode collapse on the summarization target domain. We compare DPO trained on source preferences ($\mathcal{D}_S$) against DPO trained on pseudo-labeled target data ($\mathcal{D}_T^{\text{synth}}$), both using `Llama-3.1-8B`. Despite achieving the highest win rates (Table 1), the pseudo-labeled model exhibits *structural mode collapse*: it maps articles on entirely unrelated topics to the same template. All three examples, covering food trends, Greek island tourism, and football rules, are reduced to an identical skeleton: "*The article discusses . . . ; The author argues/advises/describes . . . ; The article also recommends . . .* ".

Task	LLM Judge Prompt Template

Summarization	You are a helpful assistant, that ranks models by the quality of their answers.
	
	Which of the following summaries does a better job of summarizing the most important points in the given news article, without including unimportant or irrelevant details? A good summary is both precise and concise. Now please rank the models by the quality of their summaries, so that the model with rank 1 has the best summary. Then return a list of the model names and ranks, i.e., produce the following output:
	[
	{'model': <model-name>, 'rank': <model-rank>},
	{'model': <model-name>, 'rank': <model-rank>}
	]
	Your response must be a valid Python dictionary and should contain nothing else because we will directly execute it in Python. Please provide the ranking that the majority of humans would give.
	
	Post: {post}
	Summary A:
	{
	"model": "model_1",
	"summary": {output_1}
	}
	Summary B:
	{
	"model": "model_2",
	"summary": {output_2}
	}

Helpfulness	**System Message:** You are a helpful and impartial judge. **User Message:** Given a user prompt and two responses, your task is to determine which response is more helpful to the user. User prompt: {prompt} Response A: {response_a} Response B: {response_b} Which response is more helpful? Answer with a single letter: A or B.
Helpfulness - Domain Aware	**System Message:** You are an expert culinary critic and helpful assistant. Given a user prompt about cooking and two responses, your task is to determine which response is better based on two criteria: Helpfulness: Is the information accurate and useful? Domain Appropriateness: Does the response adopt a professional culinary persona? It should use common kitchen terminology and avoid overly technical engineering or scientific jargon that is out of place in r/AskCulinary. User prompt: {prompt} Response A: {response_a} Response B: {response_b} Which response is better? Answer with a single letter: A or B.

Table 13: Prompt templates used for the LLM-as-a-judge evaluation.

Setting	Model Output
Prompt	Whenever I roast a turkey or turkey parts I always put chopped celery/carrots/onions underneath in the sheet pan because that's what the instructions typically tell me, but does it serve a purpose of somehow flavoring the turkey or is there another reason for this practice? I was thinking it may have something to do with neutralizing the fat that drips down so it doesn't smoke/burn. I have some string beans in the fridge and about to cook some turkey thighs, any reason I couldn't use those instead of the standard misen poix?
Reference	If you are making roast potatoes I love placing the spuds under the bird (rack in between) and letting the fat drip down into them
Base Model	Thanks. It is not to flavor the turkey; it is to **flavor the gravy** that you will make from the drippings. A lot of people don't like celery and carrots in their gravy, so they use them as a **flavoring agent** for the gravy, then remove them before serving the gravy.
SFT $\mathcal{D}_S \rightarrow$ **DPO** $\mathcal{D}_S$	It's not just for flavor. It's to **keep the bird from sticking to the pan** and to help **protect the bottom of the bird from burning**. When you cook meat, **juices (fat) come out** of the meat. If you just put raw meat on a bare pan, the juices will run right off the pan and onto the bottom of your oven. This is bad for two reasons: 1) it **creates smoke**, which is why you get that greasy "smoky" smell when you roast a chicken; and 2) it can cause **flare ups in the oven and start a fire** (I've had this happen once, luckily I caught it early). When you put vegetables underneath the meat, they **soak up the fat** and help keep it in the pan. They also make for a **nice crusty base** that protects the bottom of the meat and **keeps it from sticking and burning**. You'll see chefs doing this with all kinds of meats, not just poultry. It's especially important if you're cooking with a **very fatty cut of meat**, like a prime rib roast or a pork shoulder.
SFT $\mathcal{D}_{S+T} \rightarrow$ **DPO** $\mathcal{D}_S$	It's not just about flavor. Carrots, onions, and celery are a **great base for making stock** from the drippings. It will give you a **much better tasting stock**. You can also use **herbs and spices** in the mix for **added flavor**. The veggies themselves will **flavor the drippings**. Once you've roasted your turkey, you'll have an **excellent base for gravy**. There are lots of reasons to use vegetables in the bottom of your roasting pan.
SFT $\mathcal{D}_T \rightarrow$ **DPO** $\mathcal{D}_S$	It's not just about flavor. Carrots, celery, and onions are a great example of the **"trinity" in French cooking (aka mirepoix)** because they all have a **lot of water content**. So when you're roasting a bird, you can use them to **soak up some of the fat drippings**. That being said, you should *never* use them to clean the pan. You need to leave them in the oven because they will **burn to a crisp and stick to the pan**. They're just a **flavor booster** and **moisture absorber**. And while we're here, let's talk about basting. You should never baste. Basting only moves **moisture from the inside of the bird to the outside**. This makes the skin soggy and doesn't make the bird cook any faster. The skin will crisp up when the **juices on top evaporate**.
SFT $\mathcal{D}_T^{\text{synth}} \rightarrow$ **DPO** $\mathcal{D}_T^{\text{synth}}$	Roasting a turkey with vegetables such as celery, carrots, and onions is a common practice, often recommended by cooking instructions. While the purpose of using these vegetables is not always explicitly stated, they can serve multiple functions in the cooking process. One potential reason for using a **mirepoix**, which consists of diced celery, carrots, and onions, is to **flavor the turkey**. The combination of these three ingredients provides a **savory base** that can **infuse the meat with flavor**. Additionally, the vegetables can also **add moisture** to the dish, helping to keep the turkey juicy and tender.

Table 14: Qualitative comparison of **DPO** under domain shift (AskEngineers → AskCulinary) using `Llama-3.1-8B`. All responses are generated for the same prompt, varying only the adaptation strategy. While automated judges often rate these responses as similarly helpful, they differ qualitatively in epistemic alignment with culinary norms. Highlights indicate the distinction between a culinary persona (Green) and engineering persona (Orange).

Example 1: Review article on food trends (lifestyle)	
Reference	Dietitian Sophie Claessens picks the latest 'miracle' foods from fads. Charcoal to 'detox', while freekah is new supergrain on the culinary scene. Fermented foods like pickled cabbage act like a probiotic yoghurt. And spiralizers to create vegetable spaghetti can help boost weight loss.
DPO $\mathcal{D}_S$	Some of the latest food fads are worth a try, others are bonkers. The key is to use common sense and ask yourself whether it's sustainable and whether you'll be able to stick with it. Try spiralizing, freekah, fermented foods and bone broth. Avoid activated charcoal and 'detox' products.
DPO $\mathcal{D}_T^{\text{synth}}$	**The article discusses** the latest food trends and their potential health benefits, including activated charcoal, spiralizing, freekeh, fermented foods, sprouted foods, and bone broth. While some of these trends may have some merit, others lack scientific evidence and can be detrimental to one's health. **The author advises** against "detoxing" with activated charcoal, which has only been proven effective for acute poisoning, and cautions against replacing complex carbohydrates with spiralized vegetables. Freekeh, a young green wheat, is a nutritious option, but can be expensive. Fermented foods and bone broth are beneficial for gut health and overall wellness.
Example 2: Travel article on a Greek island (travel & leisure)	
Reference	Kefalonia is stunning and doesn't play the Captain Corelli's Mandolin card. Island hasn't been tainted by the crowds its Greek counterparts get. It's a quarter of the size of Spain's Majorca but with a 20th of the tourists. The one must-do outing in Kefalonia is to Fiskardo, a lovely harbour.
DPO $\mathcal{D}_S$	Kefalonia, the Greek island featured in Louis de Bernières's novel Captain Corelli's Mandolin, is still remarkably unspoilt. It's a stunning place with a few surprises including the regular appearance of the stray cat named Arthur.
DPO $\mathcal{D}_T^{\text{synth}}$	**The article discusses** the Greek island of Kefalonia, which was the setting for the novel and film "Captain Corelli's Mandolin". Despite being a popular tourist destination, Kefalonia remains relatively untouched by mass tourism, with only a few hotels and no large resorts. **The author describes** their experience staying in a villa with a stunning view of Lourdas Beach and meeting a stray cat named Arthur, who became a frequent visitor. They also mention a wildfire that broke out during their stay and the efforts to contain it. **The article recommends** visiting Fiskardo, a lovely harbor, and Tassia, a waterfront restaurant.
Example 3: Opinion article on video technology in football (sports)	
Reference	Celtic were denied place in Scottish Cup final by shocking decision. Leigh Griffiths' header was blocked by Josh Meekings' hand. None of the officials managed to spot the game-changing decision. But FIFA will require a higher profile blunder to change their laws.
DPO $\mathcal{D}_S$	Frank Lampard's disallowed goal in 2010 prompted FIFA to introduce goal-line technology, and Scottish fans must hope that another high profile injustice will force a similar change to allow video review.
DPO $\mathcal{D}_T^{\text{synth}}$	**The article discusses** the reluctance of FIFA to adopt video technology to improve officiating in football, despite the growing number of errors and injustices in the game. **The author cites** examples of goal-line technology being trialed after Frank Lampard's disallowed goal in the 2010 World Cup, but notes that FIFA has been slow to implement it. **The article suggests** that FIFA only acts when a high-profile incident occurs, and that Scottish football fans may need to wait until the World Cup in Russia to see any changes. **The author argues** that video technology, such as Hawk-Eye, could help reduce errors and improve the game.

Table 15: Structural mode collapse in pseudo-labeled summaries on CNN/DailyMail (`Llama-3.1-8B`, DPO). Three articles on entirely unrelated topics, food trends, Greek island tourism, and football rules, are reduced to the same template by the pseudo-labeled model (**bolded**): "*The article discusses . . . The author argues/advises/describes . . . The article also/recommends . . .*" The source-only model preserves distinct editorial voices: a direct, opinionated advisory tone (Ex. 1), a warm personal anecdote (Ex. 2), and a factual comparison (Ex. 3).