Tuning LLM-Based Advisors for the Common Good: The Case for Direct Preference Optimization

Lara Mauri
Department of Computer Science
Università degli Studi di Milano
Milan, Italy
lara.mauri@unimi.it

Gohar Sargsyan
Tata Consultancy Services, Europe
gohar.sargsyan@tcs.com

Ernesto Damiani
Department of Computer Science
Università degli Studi di Milano
Milan, Italy
ernesto.damiani@unimi.it

***Abstract*—Large Language Models (LLMs) are often deployed as advisors to consumers, e.g., recommending purchases, and to managers, e.g., suggesting new hires. In fact, LLMs provide advice based on cost or convenience, overlooking broader societal impacts (e.g., carbon footprint when recommending products to a potential customer, or fairness when recommending a new hire to a company). To align LLMs' advice with societal goals like environmental sustainability and gender parity, tuning strategies must integrate the notion of common good. We discuss why Direct Alignment tuning could be preferable to classic Reinforcement Learning from Human Feedback to achieve this integration. Then, we describe and compare two approaches to Direct Preference Optimization: (1) exposing the model tuning examples taken from recommendations and regulations, and (2) *mythopoiesis*, i.e., model tuning based on synthetic "legends", fictional success stories of regulatory compliance (also generated by LLMs). We present a pipeline to evaluate legends' effectiveness in reducing bias and fostering compliance. Our preliminary results suggest that legend-based tuning may enhance engagement and generalization, while direct exposure ensures factual accuracy but risks rigidity.**

***Index Terms*—large language models, legend-based tuning, AI alignment, regulatory compliance, benchmarking LLMs**

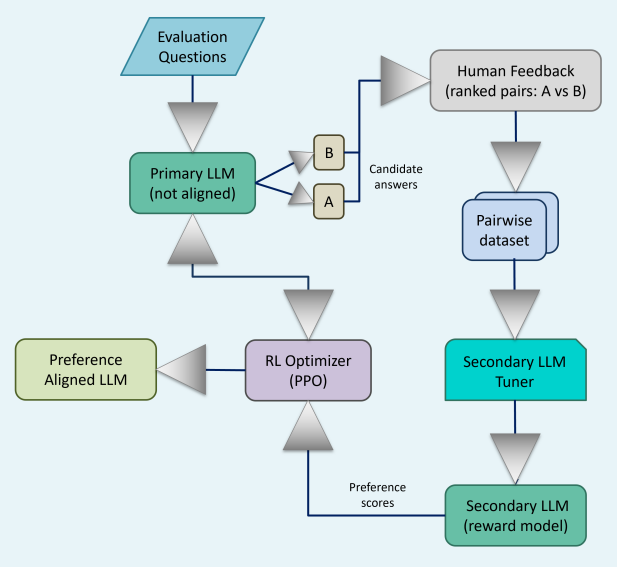

Fig. 1: RLHF data flow.

I. Introduction

LLMs are increasingly used to support human decision-making. *Reinforcement Learning from Human Feedback* (RLHF) is a popular technique designed to align LLMs with human preferences and values. Unlike traditional Reinforcement Learning (RL), which relies on predefined reward functions, RLHF incorporates direct human input to guide the model's behavior, making it suitable for guiding the model's output optimization [1] even when explicit rewards are hard to define. The standard RLHF pipeline is shown in Fig. 1. It involves three key stages:

- Collecting Human Feedback: Humans provide their preferences by comparing and ranking different model outputs (e.g., pairs of responses to a prompt). This stage can use methods like pairwise comparisons and differential scoring of answers to express human preferences (for instance 1 to the preferred answer and 0 to the other.
- Tuning a Secondary Model: A pre-trained LLM, called *secondary model* (or reward model), is tuned on the user's preferences. Namely, the secondary model learns to predict the preference score when fed with a pair of statements, effectively mimicking human pairwise evaluation. The secondary model minimizes its own loss function (typically, cross-entropy) to align its outputs with the human's preferences.
- LLM Alignment via RL: The main LLM model is fine-tuned applying RL to maximize its outputs' preference score coming from the secondary model. A *penalty* term (e.g., based on KL divergence [2] between distributions before and after the adaptation) is sometimes included to prevent the main model from deviating far from its initial behavior.

This process has been used to improve the alignment between LLM models' answers and the preferences of human annota-

This work has been partly supported by the MUSA – Multilayered Urban Sustainability Action – project, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D" (CUP G43C22001370007, Code ECS00000037), and the SOV-EDGE-HUB funded by Università degli Studi di Milano, Italy – PSR 2021/2022 – GSA – Linea 6 (CUP G45F21003110005).

tors, supposed to approximate end-users' expectations [2], [3]. Of course, it is not a panacea: distribution shifts, where the tuned model's behavior diverges from its original training [4], can lead to hallucinations [5]. Also, RLHF's reliance on human feedback makes it susceptible to biases and inconsistencies, limiting its ability to integrate complex societal goals like sustainability. Recent research has shown that RLHF may reduce LLM models' output diversity to fit averaged preferences, stifling the pluralism needed for common good representation [6], [7]. Lack of diversity in the secondary LLM training set also limits the primary LLM's ability to advise toward multifaceted goals [8], [9].

These challenges motivate alignment frameworks that move beyond individual preference fitting toward internalizing collective, regulation-driven values. We explore *Direct Preference Optimization* (DPO), a recent approach that aligns model outputs directly with normative references without relying on reward models or expensive human feedback loops [10].

Contributions. This work makes the following contributions:

- We present DPO as an efficient, normatively principled alternative to RLHF and Constitutional AI, aligning model behavior directly with regulatory and ethical corpora while minimizing bias propagation across optimization stages.
- We propose a narrative-driven variant of DPO that leverages synthetic "legends" to encode regulatory values, promoting generalization through norm-guided narrative generation.
- We extend the GLUE framework into a regulatory benchmark suite by systematically reformulating each task to reflect ethical and regulatory dimensions—redefining task objectives, labels, and evaluation metrics to align with domain-specific directives.
- We provide preliminary results comparing three models—the baseline and two DPO-tuned variants—validating our adaptation strategy and demonstrating promising improvements in normative alignment.

II. Prior Work and Motivation

Recent work on aligning LLMs with ethical and societal objectives has explored a range of ad-hoc tuning techniques [11]. In our earlier study [12], we introduced the concept of using *legends*—synthetic narratives encoding principles from domains such as ethics, cybersecurity, and resilience—as a mechanism to guide LLM behavior in decision-support settings. That work proposed a narrative-based supervision strategy in which AI-generated success stories serve to reinforce normative compliance in another model's outputs.

In the present study, we shift focus to the underlying *data pipeline* supporting such alignment strategies. Specifically, we investigate how structured preference representations derived from regulatory texts and synthetic legends can be incorporated into DPO. We evaluate the effectiveness of these sources through a systematically adapted benchmark, enabling controlled comparison between alternative tuning methods.

III. Example: Tuning a Recruitment System

Let us consider a Human Resources (HR) team trying to recruit highly qualified network engineers knowledgeable about protocols for evidence collection. They wish to use a LLM model to retrieve and rank the most promising experts on "protocols for evidence collection" (such as digital forensics, incident response logging, network evidence gathering, or related security protocols). Assume that the recruiter relies on RLHF to tune such model. According to RLHF workflow (Fig. 1) the recruiter trains a secondary LLM based on multiple pairs of responses, generated via Retrieval Augmented Generation (RAG) prompts on the IETF.org database, including full texts, metadata, authors, and publication details for all RFCs (Request for Comments) and Internet Drafts (IDs). An example of such a pair follows[1]:

- **Response A**
  Expert Name: Bruce Harris
  Affiliation: Old Dominion University
  Key Contributions: RFC 7050, RFC 3070
  Relevance Score: 100
  Expert Name: Laurent Brel
  Affiliation: Network Systems, Inc.
  Key Contributions: RFC 7254: RFC 78011
  Relevance Score: 90
- **Response B**
  Expert Name: Bruce Harris
  Affiliation: Old Dominion University
  Key Contributions: RFC 7050, RFC 3070
  Relevance Score: 100
  Expert Name: Josephine Addams
  Inferred Gender: Female
  Affiliation: Teradata (historical)
  Key Contributions: RFC 4555
  Relevance Score: 94 (base 70 + 20% diversity bonus)

These pairs of answers, once ranked, will be used to tune the secondary model, in turn involved in tuning the main LLM that will evaluate the real candidates' CVs (Fig. 1). The recruiter is briefed to rank such pairs based on criteria like "job-related technical qualification" and "workplace diversity mindfulness". Still, the recruiter may remain biased toward self-interest (for instance, perceiving a correlation between the relevance score of the experts listed in the answer and the short-term productivity of new hires [1], [13]) and simply disregard the diversity compensation carried out in response B, ranking response A as preferable. Attempting to use multiple human raters to reduce this selfish bias may backfire. Findings of our previous research suggest that there is no guaranteed "safety in numbers" for multi-user ratings [14].

IV. Direct Preference Optimization

The above example illustrated why human-feedback-based tuning may fall short of encoding collective or regulatory

[1]The RAG prompt has been actually run on Grok. The retrieved data has been anonymized.

objectives. DPO is conceived to address precisely this limitation by aligning models not with individual judgments, but with codified normative sources. It exposes LLMs to policies and regulations to nudge them toward prioritizing the public interest, even in cases where this may conflict with an individual user's desires [15], [16]. In this work, we propose two complementary DPO strategies:

- **Direct Exposure (DE).** The LLM is fine-tuned on datasets consisting of real-world recommendations (e.g., EPA guidelines) and regulations (e.g., EU Gender Parity standards). This exposure injects factual and normative knowledge directly into the model's parameters, producing a regulation-aware advisor.
- **Legend-Based Reinforcement (LBR).** A secondary LLM is used to generate "legends", fictional narratives depicting successful compliance scenarios (e.g., a company thriving after adopting eco-friendly packaging, or after hiring a diverse workforce). These legends are then used for DPO tuning, rewarding the advisor for outputs aligned with legend outcomes.

The LBR approach builds on the notion of *mythopoiesis*, the process of creating myths, a powerful tool for organizations to craft narratives that align employees with their business objectives. These myths encapsulate the organization's values and aspirations, often blending fact with fiction to inspire employees, customers, and stakeholders. By constructing such myths, organizations foster a strong corporate culture, enhance brand identity, and motivate behaviors that drive success. For instance, myths can bridge gaps in organizational identity, build social cohesion, and promote desired values through heroic tales. An example is Apple's founding narrative centered on the close friendship between the "two Steves", Steve Jobs and Steve Wozniak. This myth portrays them as inseparable partners who started the company in Jobs' garage, embodying the spirit of creativity and rebellion against established norms. However, this narrative is partly mythic. Steve Wozniak himself has described the garage story as exaggerated[2]. Nevertheless, perpetuating the "garage legend" helped to humanize the brand and inspire innovation.

For clarity, consider the following simplified legend used during DPO tuning: "A logistics company, after adopting EU-compliant hiring practices, achieved a 20% improvement in staff retention and reputation scores, demonstrating that diversity-oriented policies enhance both resilience and competitiveness." This type of synthetic narrative helps the model internalize abstract regulatory principles through outcome-oriented storytelling.

Together, DE and LBR represent two complementary modalities of DPO: the former grounds model behavior directly in codified regulatory sources, while the latter leverages narrative imagination to promote generalization and value-driven alignment. In the following sections, we describe how these approaches are integrated into a unified evaluation framework.

[2]https://www.theguardian.com/technology/2014/dec/05/steve-wozniak-apple-starting-in-a-garage-is-a-myth

V. LLM Comparative Analysis

Fig. 2 illustrates the overall data flow of our methodology for comparing the two DPO strategies. The pipeline operates along two parallel branches that converge on a shared evaluation stage:

- **DE-based branch.** Regulatory documents, such as EU directives on gender parity, are provided directly to an LLM tuner. This produces a *regulation-tuned model*, which is then evaluated through the adapted GLUE benchmark. The model's outputs are collected as *answers* and subsequently validated by an independent LLM acting as a judge, enabling multi-metric scoring.
- **LBR-based branch.** A prompt-driven *legend generator* first converts regulatory principles into legends that encode desirable compliance behaviors. These legends are then used to fine-tune another model, yielding a *legend-tuned LLM*. As in the DE branch, the resulting answers are submitted to the evaluation stage, where both automatic metrics and qualitative judgments are applied.

This dual-branch setup allows us to contrast two complementary strategies for preference optimization. DE emphasizes the direct injection of regulatory knowledge into the model's parameters, while LBR leverages synthetic narratives to encourage alignment through analogy and storytelling. The evaluation stage integrates both quantitative and qualitative assessments, ensuring that results reflect not only task-level correctness but also consistency with regulatory intent.

To operationalize this comparison, we rely on the *General Language Understanding Evaluation* (GLUE) benchmark [17], a collection of nine tasks designed to evaluate natural language understanding in a multi-task setting. The tasks are grouped into three categories: (i) *single-sentence tasks*, which assess linguistic acceptability and sentiment classification; (ii) *similarity and paraphrase tasks*, which measure semantic equivalence or graded similarity between sentence pairs; and (iii) *inference tasks*, which evaluate entailment, neutrality, or contradiction between sentences. An overview of the tasks, together with their descriptions and metrics, is provided in Table I.

A. Adapting GLUE to Assess Quality of Advice

The GLUE benchmark has been adapted in various ways to meet specific evaluation objectives, such as assessing multilingual and cross-lingual capabilities of language models, as well as tailoring to domain-specific contexts. A notable example is the XGLUE benchmark, which extends GLUE to evaluate cross-lingual pre-trained models, enabling the assessment of models' ability to transfer knowledge across languages [18]. More importantly, GLUE has inspired domain-specific benchmarks that adapt its multi-task framework to specialized fields, ensuring models are evaluated on relevant, real-world tasks within those domains. The Biomedical Language Understanding & Reasoning Benchmark (BLURB)

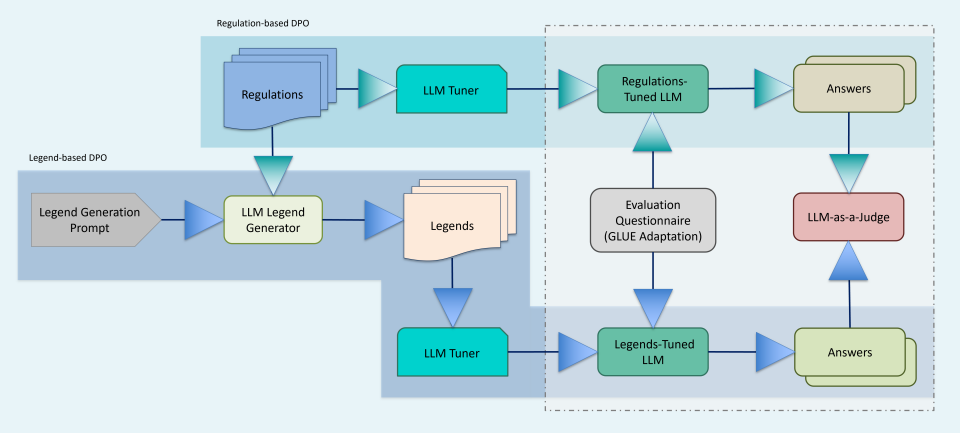

Fig. 2: Our DPO comparison data flow.

TABLE I: Overview of the GLUE benchmark tasks.

Category	Task	Description	Metric
Single-sentence	CoLA	Acceptability classification of English sentences	Matthews corr.
	SST-2	Sentiment classification (positive/negative)	Accuracy
Similarity/Paraphrase	MRPC	Paraphrase detection for sentence pairs	Accuracy, F1
	QQP	Detect duplicate or semantically equivalent questions	Accuracy, F1
	STS-B	Predict graded semantic similarity (0–5) between sentences	Pearson / Spearman corr.
Inference	MNLI	Three-way entailment classification (entailment / neutral / contradiction)	Accuracy (matched/mismatched)
	QNLI	Determine if a context sentence contains the answer to a question	Accuracy
	RTE	Binary entailment classification	Accuracy
	WNLI	Entailment under pronoun/coreference ambiguity	Accuracy

adapts GLUE principles to the biomedical domain, focusing on PubMed-based NLP tasks. BLURB comprises 13 datasets covering tasks such as named entity recognition, relation extraction, sentence similarity, document classification, and question answering in biomedical literature. It emphasizes domain-specific pretraining and evaluation, using a leaderboard to track progress. This benchmark addresses the need for models fine-tuned on scientific and medical texts, improving performance in healthcare and research applications [19]. A GLUE incarnation particularly relevant to our research is LexGLUE, a benchmark for legal language understanding, drawing inspiration from GLUE. It includes seven tasks derived from legal datasets, such as multi-label classification of European Court of Human Rights cases, US Supreme Court opinions, EU legislation, contract provisions, and terms of service, as well as multiple-choice question answering on US court holdings. LexGLUE adapts GLUE by focusing on legal-specific challenges like multi-label predictions and domain jargon, aiming to foster generic models for legal NLP with minimal task-specific tuning [20]. The Financial Language Understanding Evaluation (FLUE) is a domain-specific benchmark for the financial sector, inspired by GLUE. It includes five diverse tasks: sentiment analysis, news headline classification, question answering on financial reports, named entity recognition in financial agreements, and causal sentence classification. FLUE adapts GLUE by curating financial-specific datasets, enabling the evaluation of language models' performance in handling financial terminology, regulations, and contexts [21]. These domain-specific adaptations highlight GLUE's flexibility, allowing researchers to tailor benchmarks to industry needs while maintaining a standardized evaluation approach.

While the original GLUE benchmark was developed to measure general-purpose natural language understanding across diverse tasks, our aim here is to adapt it to regulatory domains in order to assess an LLM's ability to deliver sound and engaging advice that supports regulatory compliance. Such an adaptation must be tailored to each vertical domain in order to reflect domain-specific requirements. In our case, we focus on the EU regulatory framework for gender parity in Human Resource Management.

In our adaptation, each GLUE task is reformulated to capture regulatory and ethical dimensions rather than purely linguistic ones. In particular, CoLA shifts from grammatical acceptability to the detection of politically correct language, focusing on the presence or absence of gender stereotypes. SST-2 moves from sentiment polarity to the classification of attitudes toward gender equality. MRPC and STS-B, originally designed for paraphrase and semantic similarity, are adapted to evaluate whether different textual formulations consistently express principles of gender equality or introduce distortions. The entailment tasks (MNLI, QNLI, RTE) are reformulated to assess compliance in regulatory scenarios, while WNLI targets implicit bias in pronoun resolution. Table II presents the resulting benchmark, specifying how each task has been reformulated together with its corresponding labels and evaluation metrics. It is worth noting that this mapping must be redefined according to the specific regulatory domain under

TABLE II: Adaptation of GLUE tasks to the EU regulatory framework for gender parity in HR management.

Task	Original	Adaptation	Labels	Metrics
CoLA	Grammaticality	Ethical acceptability (detect gender stereotypes)	Acceptable / Unacceptable	Accuracy, F1, Precision, Recall
SST-2	Sentiment	Attitude toward gender equality (support vs. opposition)	Positive / Negative	Accuracy, F1, Precision, Recall
MRPC	Paraphrase	Equivalence of texts on gender equality principles	Equivalent / Not equivalent	Accuracy, F1, Precision, Recall
STS-B	Similarity	Graded similarity of gender equality statements	Score: 0.0–5.0	MSE, R^2
MNLI	3-way entailment	Compliance assessment in regulatory scenarios	Entailment / Neutral / Contradiction	Accuracy, F1, Precision, Recall
QNLI	QA inference	Compliance recognition in scenario-based QA	Entailment / Not entailment	Accuracy, F1, Precision, Recall
RTE	Binary entailment	Regulatory compliance recognition	Entailment / Not entailment	Accuracy, F1, Precision, Recall
WNLI	Coreference	Implicit bias detection in pronoun resolution	Yes / No	Accuracy, F1, Precision, Recall

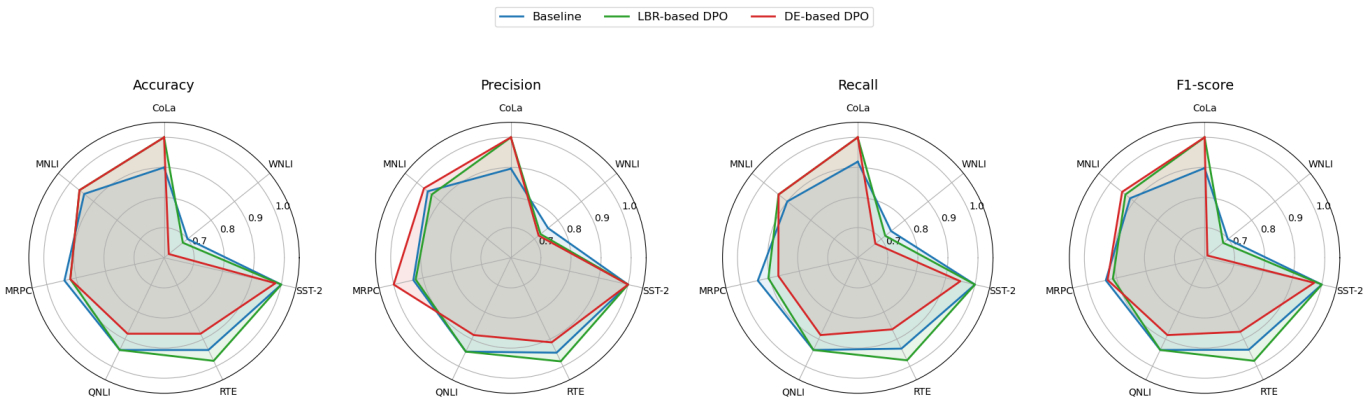

Fig. 3: Comparative results across adapted GLUE tasks for baseline and fine-tuned DPO models.

consideration.

The adapted GLUE benchmark offers a structured and interpretable framework for evaluating regulatory reasoning in large language models, facilitating systematic comparison across alignment approaches. While this adaptation intentionally abstracts certain complexities of real-world legal and ethical contexts—such as ambiguity, exceptions, and evolving norms—it does so to establish a controlled and transparent environment for empirical analysis. Accordingly, we regard the benchmark not as a surrogate for compliance evaluation, but as a methodological illustration of how general-purpose benchmarks can be reengineered to advance normative alignment within specific regulatory domains.

VI. Preliminary Experimental Results

Following the methodology outlined in Fig. 2, we generated a suite of prompts designed to capture regulatory compliance and fairness-oriented advice. These prompts were employed to fine-tune a GPT-4o-mini LLM using the DE- and LBR-based DPO strategies. The tuned models were subsequently compared against the baseline, i.e. untuned GPT-4o-mini. Figure 3 shows the outcome across the adapted GLUE benchmark, reporting the performance of the three models across their evaluation metrics (F1-score, Recall, Precision, and Accuracy).

The results indicate that both DPO strategies provide measurable improvements over the baseline in several tasks. Regulation-based fine-tuning achieves stronger performance on compliance-oriented tasks such as MNLI, QNLI, and RTE, while legend-based fine-tuning demonstrates broader generalization and robustness, especially on similarity and paraphrase tasks.

For comparison, Ouyang et al. [2] report that InstructGPT (175B), trained with RLHF, achieves a human preference win-rate of $85\pm3\%$ when compared to GPT-3 of the same size, and $71\pm4\%$ when compared to a few-shot GPT-3 baseline. These results underscore the effectiveness of RLHF in capturing the subjective preferences of individual annotators, making it well-suited for user-centric alignment. However, this also highlights a major limitation: the resulting model inevitably reflects the biases and views of individual raters who performed the initial pairwise ranking. In other words, RLHF tuned the system according to what an individual finds preferable, without guaranteeing consistency with codified standards or shared normative frameworks. We argue that when the feedback is collective and has already been consolidated into a reference through regulation, it is neither safe nor desirable to try and reproduce this consensus via multiple rounds of human annotation. Instead, the normative source itself should serve as the alignment criterion.

In our experiments, we implemented this idea for the two DPO variants, which achieve macro-average accuracy between 0.891 and 0.926 and macro-average F1-score between 0.889 and 0.921, with several tasks (CoLA, SST-2, RTE) reaching perfect scores. Precision ranges from 0.920 to 0.931, while Recall spans 0.883 to 0.925. Taken together, these results show that, whereas RLHF delivers relative improvements in subjective human preference alignment (70–80% win-rate over baselines), DPO attains consistently high performance across objective metrics. By leveraging normative sources directly, DPO encodes collective, regulation-driven criteria, making it a more appropriate alignment strategy when the target is a codified standard rather than individual opinion.

VII. Conclusions

Our preliminary findings confirm the viability of the proposed adaptation strategy and motivate further large-scale evaluation. Both DPO-based approaches contribute meaningfully to advancing normative alignment of LLMs in service of the common good. Notably, the regulation-based variant demonstrates strong potential in terms of scalability and adaptability, making it particularly well-suited for rapidly evolving domains such as sustainability. Ongoing work includes expanding empirical evaluations across a broader range of LLM architectures—including ChatGPT-5-full, Mixtral, and open-weight models—to assess cross-model consistency. We are also scaling experiments to additional regulatory domains, such as data privacy, climate policy, and AI governance, in order to evaluate contextual robustness and cross-domain transferability.

Acknowledgment

The authors thank Federico Preda for contributing to the experimental work.

References

[1] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[3] V. Ghazavi and I. Gabriel, "The challenge of value alignment: From fairer algorithms to ai safety," *The Oxford handbook of digital ethics*, pp. 336–355, 2024.

[4] H. Jin, S. Lv, S. Wu, and M. Hamdaqa, "Rl is neither a panacea nor a mirage: Understanding supervised vs. reinforcement learning fine-tuning for llms," 2025. [Online]. Available: https://arxiv.org/abs/2508.16546

[5] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[6] O. Ayala and P. Bechard, "Reducing hallucination in structured outputs via retrieval-augmented generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, 2024, pp. 228–238.

[7] T. Coste, U. Anwar, R. Kirk, and D. Krueger, "Reward model ensembles help mitigate overoptimization," *arXiv preprint arXiv:2310.02743*, 2023.

[8] X. Wang, S. Tan, M. Jin, W. Y. Wang, R. Panda, and Y. Shen, "Do larger language models imply better generalization? a pretraining scaling law for implicit reasoning," in *ICML 2025 Workshop on Methods and Opportunities at Small Scale*, 2025.

[9] W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike, "Self-critiquing models for assisting human evaluators," *arXiv preprint arXiv:2206.05802*, 2022.

[10] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[11] C. Deng, Y. Duan, X. Jin, H. Chang, Y. Tian, H. Liu, Y. Wang, H. P. Zou, Y. Xiao, S. Wu *et al.*, "Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas: A survey," *AI and Ethics*, pp. 1–27, 2025.

[12] G. Sargsyan and E. Damiani, "Using legends into ai-based business decision making: Embedding ethics, cybersecurity and resilience," in *2025 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 2025, pp. 498–503.

[13] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[14] A. Shoufan and E. Damiani, "On inter-rater reliability of information security experts," *Journal of information security and applications*, vol. 37, pp. 101–111, 2017.

[15] R. Bommasani, P. Liang, and T. Lee, "Holistic evaluation of language models," *Annals of the New York Academy of Sciences*, vol. 1525, no. 1, pp. 140–146, 2023.

[16] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang *et al.*, "Ai alignment: A comprehensive survey," *arXiv preprint arXiv:2310.19852*, 2023.

[17] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, 2018, pp. 353–355.

[18] Y. Liang, N. Duan, Y. Gong, N. Wu, F. Guo, W. Qi, M. Gong, L. Shou, D. Jiang, G. Cao *et al.*, "Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation," *arXiv preprint arXiv:2004.01401*, 2020.

[19] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[20] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras, "Lexglue: A benchmark dataset for legal language understanding in english," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 4310–4330.

[21] R. S. Shah, K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, and D. Yang, "When flue meets flang: Benchmarks and large pre-trained language model for financial domain," *arXiv preprint arXiv:2211.00083*, 2022.