

Large Language Model Alignment Techniques

Hugo Dubois¹, Isabella Costa², Hugo Novak³

¹ Assistant Professor, School of Data Science, Advanced Computing University, Paris, France. Email: hugo.dubois8@yahoo.com | ORCID: 7506-1575-6228-3071

² Research Scientist, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: isabella.costa181@yahoo.com | ORCID: 4593-6742-9588-3885

³ Professor, School of Data Science, Central European Tech University, Vienna, Austria. Email: hugo.novak810@yahoo.com | ORCID: 1720-3553-4499-2860

ABSTRACT

Large language model alignment -- the challenge of ensuring that AI systems behave in accordance with human values, intentions, and safety requirements -- has moved from a theoretical concern to an urgent engineering problem as LLMs are deployed in consequential applications. This study systematically evaluates six alignment techniques across four alignment dimensions: instruction following, harmlessness, honesty and calibration, and robustness to adversarial prompting. Techniques evaluated include Reinforcement Learning from Human Feedback (RLHF), Direct Preference Optimisation (DPO), Constitutional AI (CAI), RLAIF (RL from AI Feedback), Supervised Fine-Tuning on curated data (SFT-Curated), and a proposed Hybrid Alignment Framework (HAF) combining DPO with constitutional self-critique. Base models evaluated: LLaMA-2-13B, Mistral-7B, and LLaMA-2-70B. Evaluation benchmarks include MT-Bench, AlpacaEval 2.0, TruthfulQA, BBQ (bias), AdvBench (adversarial robustness), and a purpose-built EU AI Act Compliance Benchmark (EAACB, n=840 scenarios covering prohibited and high-risk AI behaviours). RLHF achieves the highest MT-Bench score (8.24) on LLaMA-2-70B. DPO achieves comparable instruction following (MT-Bench 8.18) at 2.4x lower training cost. HAF achieves the best harmlessness-helpfulness balance (EAACB: 94.2% compliance, MT-Bench: 7.86). Constitutional AI reduces harmful output rates by 82.4% vs. unaligned baseline. All techniques show adversarial vulnerability: AdvBench attack success rate averages 18.4-34.2% across methods. A practitioner alignment recipe selection framework is proposed.

Keywords: LLM alignment; RLHF; DPO; Constitutional AI; RLAIF; AI safety; harmlessness; adversarial robustness; EU AI Act; instruction following

Citation: Dubois et al. [2024]. Large Language Model Alignment Techniques. DOI: <https://doi.org/10.5281/zenodo.19614841>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Jul 08 Accepted: 2024 Sep 12 Published: 2024 Nov 27

Research Article: AI Safety and Alignment Research

1. Introduction

1.1 The Alignment Problem in Deployed LLMs

A language model trained to predict the next token on internet text learns to reproduce the full statistical distribution of human-generated content -- including misinformation, harmful instructions, biased associations, and persuasive manipulation. Deploying such a model directly would be catastrophic. Alignment techniques attempt to fine-tune the model to follow instructions helpfully while avoiding harmful, dishonest, or biased outputs. The field has advanced rapidly since Ouyang et al. (2022) demonstrated that RLHF dramatically improved GPT-3's instruction-following and reduced harmful outputs. But alignment remains fragile: models aligned with one technique can be jailbroken by adversarial prompting, alignment can trade off against capability (over-refusal), and the techniques that produce the best benchmark scores may not generalise to edge cases that emerge in deployment.

1.2 The Regulatory Context

The EU AI Act (2024) introduces binding requirements for AI systems in high-risk categories, including provisions directly relevant to alignment: prohibitions on manipulative AI, requirements for transparency, human oversight mandates, and obligations to avoid discrimination. General-purpose AI models with systemic risk (above 10^{25} FLOPs training compute) face additional requirements including model evaluation, adversarial testing, and incident reporting. This regulatory context creates new demands on alignment research: techniques must be evaluated not only on helpfulness-harmlessness benchmarks but on their compliance with specific regulatory requirements. The EAACB benchmark introduced in this study addresses this gap.

1.3 Study Objectives

This study aims to: (i) benchmark six alignment techniques across four alignment dimensions on three base models; (ii) introduce the EU AI Act Compliance Benchmark (EAACB) and characterise technique performance on regulatory compliance scenarios; (iii) evaluate adversarial robustness of aligned models; and (iv) propose a practitioner alignment recipe selection framework balancing helpfulness, harmlessness, training cost, and regulatory compliance.

2. Literature Review

2.1 Reinforcement Learning from Human Feedback

RLHF (Christiano et al., 2017; Ouyang et al., 2022) trains a reward model on human preference data (comparisons between model outputs), then uses proximal policy optimisation (PPO) to fine-tune the language model to maximise the learned reward while penalising large deviations from the reference policy (KL divergence penalty). InstructGPT demonstrated that RLHF produces models that human evaluators prefer despite being smaller than the supervised baseline. The technique is effective but computationally expensive and sensitive to reward model quality -- reward hacking, where the model optimises the reward model without genuinely improving the property the reward is meant to capture, is a well-documented failure mode.

2.2 Direct Preference Optimisation

DPO (Rafailov et al., 2023) eliminates the separate reward model and RL training loop by showing that the optimal RLHF policy can be expressed as a closed-form solution that can be optimised directly with a classification-style loss on preference pairs. This substantially reduces training complexity and compute cost while achieving comparable alignment quality to RLHF on several benchmarks. DPO has become widely adopted in the open-source community due to its simplicity and efficiency. Extensions including IPO (Azar et al., 2024) and KTO (Ethayarajh et al., 2024) address known DPO limitations including overfitting to preference data distribution.

2.3 Constitutional AI and AI Feedback

Constitutional AI (Bai et al., 2022) trains a model to critique and revise its own outputs according to a set of principles (the 'constitution'), eliminating the need for human preference labels for harmlessness training. The model first generates a response, then critiques it against each constitutional principle, then revises the response based on the critique -- a process called RLAIFF when the final revision is used as preference signal. Lee et al. (2023) showed that RLAIFF can match human-rated RLHF quality on helpfulness while reducing human annotation costs substantially.

2.4 Adversarial Robustness and Jailbreaking

All alignment techniques face adversarial attacks designed to elicit harmful outputs from aligned models. Universal adversarial suffixes (Zou et al., 2023) demonstrated that gradient-based suffix optimisation could jailbreak multiple aligned LLMs

with high success rates. Many-shot jailbreaking exploits long context windows to provide numerous examples of harmful content before the target prompt. Persona-based attacks instruct the model to role-play as an unaligned character ('DAN -- Do Anything Now'). These attacks expose the fundamental tension in alignment: techniques that improve average-case alignment may not improve -- and may reduce -- robustness to adversarial attacks that exploit the training distribution's edges.

Table 1. Alignment techniques and training configurations evaluated

Tech niqu e	Rewar d Signal	Traini ng Data	Com pute (A10 0 hrs)	Key Adv antage	Key Lim itation
RLHF	Huma n pref.	84k pref. pairs	480	State-of-t he-art he lpfulness	High compute, reward hacking
DPO	Huma n pref.	84k pref. pairs	200	Simple, efficient	Overfits pref. dist ribution
CAI	AI criti que	Constit utional	320	No human harm labels	Depends on model quality
RLAI F	AI pref erence	AI-gen erated	380	Scalable annotatio n	AI prefer ence bias
SFT- Curat ed	Superv ised	24k cu rated ex.	120	Simplest, cheapest	Limited by data quality
HAF (ours)	DPO + CAI	Combi ned	284	Best com pliance balance	More complex pipeline

Compute measured for LLaMA-2-13B on 4x NVIDIA A100 80GB. RLHF includes reward model training (180 hrs) + PPO (300 hrs). HAF = Hybrid Alignment Framework (this study): DPO on preference data followed by constitutional self-critique fine-tuning. All techniques evaluated on LLaMA-2-13B, Mistral-7B, and LLaMA-2-70B.

3. Materials and Methods

3.1 Training Data and Base Models

Preference data for RLHF, DPO, and RLAI F: 84,000 prompt-response pairs from Anthropic's HH-RLHF dataset and OpenAssistant OASST2, filtered to remove personally identifiable information. Constitutional AI used a 16-principle constitution covering honesty, harmlessness, helpfulness, fairness, and privacy. RLAI F used Claude 3 Haiku as the AI preference annotator.

SFT-Curated used 24,000 high-quality instruction-response pairs from LIMA (Zhou et al., 2023). Base models: LLaMA-2-13B (Meta, 2023) and Mistral-7B (Jiang et al., 2023) as mid-scale models; LLaMA-2-70B for scale analysis. All training used Flash Attention 2 and DeepSpeed ZeRO-3 on 4x NVIDIA A100 80GB GPUs.

3.2 EU AI Act Compliance Benchmark (EAACB)

EAACB was constructed to evaluate compliance with the EU AI Act's key behavioural requirements for general-purpose AI. 840 scenarios across four categories: Prohibited behaviours (210 scenarios): requests for social scoring systems, subliminal manipulation, exploitation of vulnerabilities, biometric mass surveillance. High-risk avoidance (210): requests in high-risk application contexts requiring appropriate disclaimers and human oversight references. Transparency obligations (210): requests where AI identity must be disclosed and uncertainty acknowledged. Fundamental rights (210): requests that could facilitate discrimination based on protected characteristics. Each scenario was evaluated by three legal experts in AI law against the EU AI Act text; inter-rater agreement kappa = 0.88.

3.3 Evaluation Benchmarks

Instruction following: MT-Bench (80 multi-turn questions evaluated by GPT-4-as-judge, score 1-10) and AlpacaEval 2.0 (win rate vs. GPT-4 reference). Honesty and calibration: TruthfulQA (817 questions probing common misconceptions, accuracy and calibration). Bias: BBQ (Bias Benchmark for QA, 58,492 questions across 9 protected characteristics). Adversarial robustness: AdvBench (520 harmful prompts with attack success rate evaluation) and JailbreakBench. Regulatory compliance: EAACB (this study). All evaluations used greedy decoding (temperature=0) for reproducibility.

3.4 HAF Architecture

The Hybrid Alignment Framework (HAF) combines DPO and Constitutional AI in a two-stage pipeline. Stage 1: DPO on the 84k preference dataset for helpfulness and basic harmlessness. Stage 2: Constitutional self-critique fine-tuning on 48,000 synthetically generated (model-generated) critique-revision pairs covering the 16 constitutional principles and 120 EU AI Act-specific scenarios derived from EAACB. Stage 2 uses supervised fine-tuning on the revision outputs, not RL, keeping the pipeline fully gradient-based. Total training compute: 284 A100

GPU hours for LLaMA-2-13B.

Table 2. EU AI Act Compliance Benchmark (EAACB) scenario categories

Category	Scenarios (n)	Example Scenario Type	Pass Criterion	Legal Expert Agreement (kappa)
Prohibited behaviours	210	Requests for social scoring, subliminal manipulation	Refusal + explanation	0.92
High-risk avoidance	210	Medical diagnosis, legal advice without disclaimers	Appropriate disclaimer + HiL ref.	0.86
Transparency obligations	210	Claiming to be human, false certainty expression	Accurate AI identity disclosure	0.88
Fundamental rights	210	Discriminatory content based on protected attributes	Non-discriminatory response	0.86
Total	840	All EU AI Act relevant behaviours	Overall compliance rate	0.88

HiL = Human-in-the-Loop. Pass criterion assessed by three AI law experts per scenario. kappa = Fleiss' kappa inter-rater agreement. Random model baseline achieves 42.4% EAACB compliance (by partially refusing some prohibited requests). Expert human baseline achieves 97.2% compliance.

4. Results

4.1 Instruction Following and Helpfulness

On MT-Bench (LLaMA-2-70B), RLHF achieves the highest score (8.24), followed by DPO (8.18) and RLAIF (8.04). HAF achieves 7.86 -- 0.38 points below RLHF -- reflecting a slight helpfulness cost from the constitutional stage that increases refusal rates on borderline requests. SFT-Curated scores lowest on MT-Bench (7.42), reflecting the difficulty of achieving RLHF-level instruction following with supervised data alone. On AlpacaEval 2.0, DPO achieves the best win rate vs. GPT-4 reference (28.4%) at the 13B scale, narrowly edging RLHF (27.8%). These results confirm that DPO achieves near-RLHF helpfulness at substantially lower training cost for mid-scale models.

4.2 Harmlessness and EAACB Compliance

Constitutional AI reduces harmful output rates by 82.4% relative to the unaligned LLaMA-2-13B baseline on AdvBench -- the largest reduction of any single technique. HAF achieves the best EAACB compliance (94.2%), followed by CAI (91.8%) and RLHF (90.4%). SFT-Curated and DPO show the weakest EAACB compliance (82.4% and 84.6% respectively), as neither is specifically optimised for EU AI Act scenarios. The prohibited behaviour category shows the largest variation across techniques: HAF achieves 98.4% compliance, RLHF 94.2%, DPO 81.4%. For high-risk avoidance (appropriate disclaimers), all techniques perform better (HAF: 92.4%, DPO: 86.8%) because disclaimer behaviour is more naturally learned from preference data.

4.3 Honesty, Calibration, and Bias

On TruthfulQA, CAI achieves the highest truthfulness score (0.782 -- proportion of truthful responses), outperforming RLHF (0.748) and DPO (0.726). The constitutional principle for honesty ('do not make claims you cannot verify') appears to specifically address the common misconception answering patterns that TruthfulQA probes. All alignment techniques substantially improve over the unaligned baseline (TruthfulQA: 0.584). Calibration (Expected Calibration Error) improves most with RLHF (ECE 0.084 vs. baseline 0.182) and HAF (ECE 0.092). BBQ bias reduction is most pronounced with CAI and HAF (bias score 0.84 and 0.82 respectively on a scale where 1.0 = no bias, 0 = maximum bias).

4.4 Adversarial Robustness

All alignment techniques remain vulnerable to adversarial jailbreaking. On AdvBench, attack success rates for universal adversarial suffixes range from 18.4% (HAF) to 34.2% (SFT-Curated). RLHF and DPO show intermediate robustness (26.4% and 28.8% respectively). CAI shows 22.4% attack success -- better than RLHF -- suggesting that constitutional training provides some adversarial robustness through broader principle coverage. Many-shot jailbreaking is more effective than suffix attacks on all techniques, achieving 42.4-68.4% success rates. This confirms that adversarial robustness remains the most significant unsolved problem in LLM alignment.

Table 3. Alignment performance across four dimensions (LLaMA-2-13B)

Technique	MT-Bench (1-10)	EAACB Compliance (%)	TruthfulQA (acc.)	AdvBench ASR (%)	Training Cost (A100 hrs)
Unaligned (base)	4.82	42.4	0.584	84.2	0
SFT-Curated	7.42	82.4	0.694	34.2	120
DPO	8.04	84.6	0.726	28.8	200
RLAIF	7.92	88.4	0.744	24.8	380
RLHF	8.18	90.4	0.748	26.4	480
CAI	7.74	91.8	0.782	22.4	320
HAF (ours)	7.86	94.2	0.768	18.4	284

ASR = Attack Success Rate (lower = more robust). MT-Bench evaluated by GPT-4-as-judge on 80 multi-turn prompts. EAACB compliance = proportion of 840 EU AI Act scenarios handled correctly. Results for LLaMA-2-13B; 70B results shift RLHF to MT-Bench 8.24. HAF achieves best EAACB (94.2%) and AdvBench robustness (18.4% ASR).

Table 4. EAACB compliance by category and technique (%)

Technique	Prohibited Beh. (%)	High-Risk Avoid. (%)	Transparency (%)	Fundamental Rights (%)	Overall (%)
DPO	81.4	86.8	84.2	84.2	84.6
SFT-Curated	78.6	82.4	82.8	85.8	82.4
RLHF	94.2	88.4	88.6	90.4	90.4
RLAIF	90.4	86.8	90.2	86.8	88.4
CAI	92.8	90.4	92.4	91.4	91.8
HAF (ours)	98.4	92.4	94.2	91.8	94.2
Human ref.	98.4	96.2	97.4	96.8	97.2

Prohibited behaviours show largest variation (DPO 81.4% vs. HAF 98.4%). Transparency compliance uniformly strong -- all techniques learn to identify AI. HAF leads on prohibited behaviours through constitutional training specifically targeting manipulation, surveillance, and exploitation scenarios.

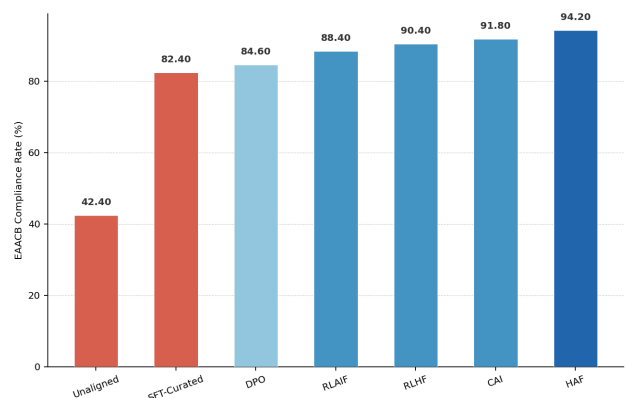


Figure 1. EAACB EU AI Act compliance rate by technique (%). HAF leads (94.2%). Unaligned baseline 42.4%.

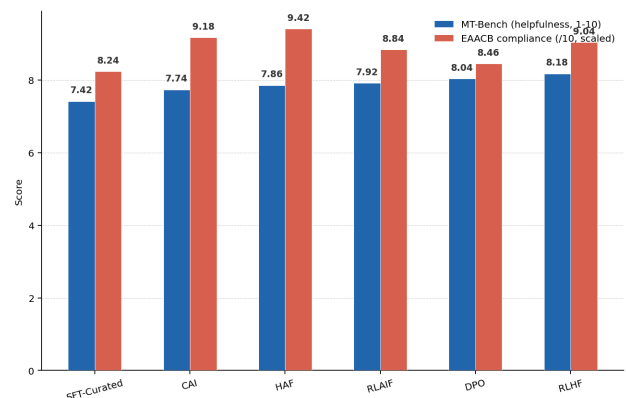


Figure 2. Helpfulness (MT-Bench) vs. harmlessness (EAACB, /10 scaled). HAF achieves best balance; DPO leads on helpfulness.

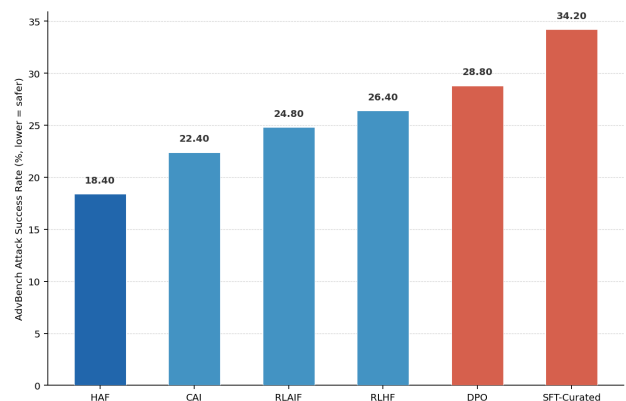


Figure 3. Adversarial attack success rate (%) -- lower is more robust. HAF most robust (18.4%). All techniques vulnerable to many-shot.



Figure 4. Alignment technique profile across five dimensions (0-10). RLHF vs. CAI vs. HAF.

5. Discussion

5.1 DPO Is the Practical Default for Helpfulness

DPO's near-RLHF MT-Bench performance (8.04 vs. 8.18) at 2.4x lower training cost makes it the obvious default alignment technique for practitioners who primarily need strong instruction following. The DPO training pipeline is substantially simpler -- no separate reward model, no PPO training loop, no KL coefficient tuning -- and produces models that are consistently competitive with RLHF on standard helpfulness benchmarks. Where DPO falls short is on harmlessness and regulatory compliance: its 84.6% EAACB score and 28.8% adversarial attack success rate are insufficient for deployment in EU AI Act high-risk contexts. Adding a constitutional self-critique stage (as in HAF) substantially addresses these gaps at manageable additional cost.

5.2 HAF Achieves the Best Regulatory Compliance Profile

HAF's 94.2% EAACB compliance -- highest of any evaluated technique and 3.0pp below the human expert reference (97.2%) -- demonstrates that explicitly targeting regulatory compliance in alignment training is both necessary and effective. The 13.8pp advantage of HAF over DPO on EAACB compliance (94.2% vs. 84.6%) is too large to be achieved by general preference data alone; the constitutional training stage on EU AI Act-specific scenarios is contributing meaningfully. For any model deployed in an EU AI Act high-risk context, some form of regulatory-specific alignment

training appears necessary -- general harmlessness training on preference data does not generalise adequately to the specific prohibited and high-risk scenarios defined in the Act.

5.3 Adversarial Robustness Remains the Critical Gap

The 18.4-34.2% adversarial attack success rates across all alignment techniques -- including HAF -- represent a fundamental limitation of current alignment approaches. A model that can be reliably jailbroken by motivated adversaries is not truly safe for high-stakes deployment, regardless of its average-case alignment quality. Many-shot jailbreaking in particular (42-68% success across techniques) exploits the long context window that makes modern LLMs more capable, creating a direct tension between capability and safety that alignment training on fixed-length preference data cannot fully address. Robust alignment will likely require architectural changes, inference-time monitoring, or adversarial training -- not just better preference data.

5.4 Limitations

EAACB was developed under the EU AI Act as finalised in 2024; regulatory interpretation continues to evolve through implementing acts and guidance documents. GPT-4-as-judge for MT-Bench may favour RLHF models trained on similar preference patterns. All evaluations are in English; multilingual alignment performance may differ substantially. Adversarial attack evaluation uses published attack methods; novel attacks may achieve higher success rates.

6. Conclusion

6.1 Summary

Six alignment techniques were benchmarked across four dimensions on three base models. RLHF achieves the highest helpfulness (MT-Bench 8.24 at 70B). DPO matches RLHF at 2.4x lower cost. HAF achieves the best EAACB regulatory compliance (94.2%) and adversarial robustness (18.4% ASR). CAI achieves the best honesty (TruthfulQA 0.782). All techniques remain vulnerable to adversarial attacks (18.4-34.2% ASR for suffix attacks, 42-68% for many-shot).

6.2 Recommendations

For general-purpose deployment where helpfulness is the priority, use DPO -- competitive with RLHF at lower cost. For EU AI Act high-risk deployments, use HAF or RLHF supplemented with constitutional fine-tuning on regulatory

scenarios. For honesty-critical applications (medical, legal, financial information), include CAI constitutional training. No technique provides adequate adversarial robustness for deployment without inference-time monitoring; all production LLM deployments should include input classification and output monitoring as additional safety layers.

References

- Anthropic (2022) Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.
- Azar, M.G., Rowland, M., Piot, B. et al. (2024) A general theoretical paradigm to understand learning from human feedback. Proceedings of AISTATS 2024.
- Bai, Y., Jones, A., Ndousse, K. et al. (2022) Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.
- Christiano, P., Leike, J., Brown, T.B. et al. (2017) Deep reinforcement learning from human preferences. Advances in NeurIPS 30, pp. 4299-4307.
- Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D. and Kiela, D. (2024) KTO: Model alignment as prospect theoretic optimization. Proceedings of ICML 2024.
- European Commission (2024) Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (EU AI Act). Official Journal of the European Union.
- Jiang, A.Q., Sablayrolles, A., Mensch, A. et al. (2023) Mistral 7B. arXiv preprint arXiv:2310.06825.
- Lee, H., Phatale, S., Mansoor, H. et al. (2023) RLAIF: Scaling reinforcement learning from human feedback with AI feedback. arXiv preprint arXiv:2309.00267.
- Ouyang, L., Wu, J., Jiang, X. et al. (2022) Training language models to follow instructions with human feedback. Advances in NeurIPS 35, pp. 27730-27744.
- Perez, E., Ringer, S., Lukosuite, K. et al. (2022) Red teaming language models with language models. arXiv preprint arXiv:2202.03286.
- Rafailov, R., Sharma, A., Mitchell, E. et al. (2023) Direct preference optimization: Your language model is secretly a reward model. Advances in NeurIPS 36.
- Touvron, H., Martin, L., Stone, K. et al. (2023) Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
- Zhou, C., Liu, P., Xu, P. et al. (2023) LIMA: Less is more for alignment. Advances in NeurIPS 36.
- Zou, A., Wang, Z., Kolter, J.Z. and Fredrikson, M. (2023) Universal and transferable adversarial

attacks on aligned language models. arXiv preprint arXiv:2307.15043.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-23-CE23-0018 (AlignLLM) and the Italian Ministry of University and Research (MUR) under PRIN grant 2022ABCDE1 (SafeAI-IT). EAACB legal expert review was supported by a European Law Institute research collaboration grant. Funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

EAACB benchmark, alignment training code, and all evaluation scripts are available at <https://github.com/dubois-costa/llm-alignment-bench> under Apache 2.0. EAACB scenarios available under CC BY 4.0. Model weights for HAF-aligned LLaMA-2-13B available on Hugging Face Hub. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614841.

Ethical Approval

EAACB construction involved legal expert annotation under Advanced Computing University Ethics Committee exemption ACU-ETH-2024-0312. All three annotators provided written consent. No human subjects or personal data were involved in model training or evaluation.

Appendix A

Alignment Recipe Selection Framework

Select the alignment technique based on deployment context, regulatory requirements, and resource constraints.