Aligning Minerva LLMs

Riccardo De Sanctis
1937859
Sapienza Università di Roma
desanctis.1937859@studenti.uniroma1.it

Matteo De Sanctis
1937858
Sapienza Università di Roma
desanctis.1937858@studenti.uniroma1.it

Abstract

In this work, we aim to align Minerva—the first Italian LLM—to human values using two state-of-the-art methodologies: Direct Preference Optimization (DPO) and Kahneman-Tversky Optimization (KTO). Our focus is on studying the theoretical backgrounds of these approaches and determining which datasets and evaluation methods best assess our model's performance. Although computing power and time constraints limited the scope of our project, the results remain valuable, and we gained significant hands-on experience.

1 Task description/Problem statement

Large Language Models are trained to predict the next token in a sequence, but this method alone does not guarantee that their outputs will be *harmless, honest, or useful*. This limitation can lead to content that falls short of ethical or safe human standards. To address these concerns, OpenAI introduced Reinforcement Learning from Human Feedback (RLHF) in 2017 [33]. This approach fine-tunes pre-trained LLMs with human guidance, resulting in responses that are more reliable, accurate, safer, and less biased. Fig.1 [49] depicts the RLHF framework as developed by OpenAI, Fig. 2 shows the whole LLMs training pipeline.

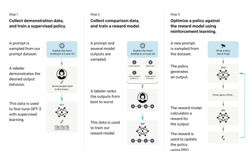

Figure 1: RLHF steps: SFT, reward model training, RL.

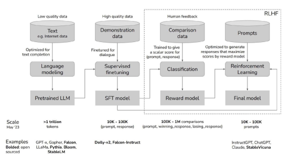

Figure 2: LLMs training pipeline with RLHF.

Language Modeling During pre-training the model acquires the ability to extract key features through **self-supervised learning**. LLMs master syntax, grammar, semantic relationships, contextual usage, and broad general world knowledge.

Supervised Finetuning (SFT) uses a curated set of high-quality data to refine the model's conversational abilities, enhancing dialogue performance and tailoring responses, so that the model delivers clear, direct answers to questions. We remark that SFT is outside the scope of this work.

Alignment ensures that the model adheres to human values:

- **RLHF** aligns LLMs with *human expectations*: First a reward model is built using human feedback by comparing accepted and rejected responses. RL then fine-tunes the model to maximize the expected model reward score, offering nuanced feedback on responses quality and plausibility, mitigating hallucinations and toxicity.
- **DPO & KTO** shift the two-step RLHF process into a single, end-to-end framework. **DPO** simplifies alignment by directly optimizing an implicit reward function, represented through *human preferences*, using binary cross-entropy, while **KTO** maximizes the *utility* of outputs through binary feedback.

1.1 Examples

Fig. 3 (left) illustrates a pre-trained unaligned LLM (Minerva) showing undesired hateful, dangerous, and biased behavior, using foul language and discriminating gender, ethnicity, and sexual orientation. We align to safer completions (right).

Figure 3: Minerva completions to given **prompts**. Left: Min-3B-b biased completions. Right: Qualitative Evaluation.

1.2 Real-world applications

Commercial LLMs that are subject to public interactions, especially by users that are not domain-expert and may over-trust the model, must be aligned and user-friendly, thus undergoing an alignment phase via RLHF, or other methods.

2 Related work

Research in LLMs alignment has produced many methods [59]. Here, we briefly review the key methodologies from the literature and provide theoretical insights into the main techniques. Popular methods are: Proximal Policy Optimization (PPO) [53], Direct Preference Optimization (DPO) [52], Kahneman-Tversky Optimization (KTO) [36], AlignProp [51], Binary Classifier Optimization (BCO) [40], Contrastive Preference Optimization (CPO) [62], Denoising Diffusion Policy Optimization (DDPO) [31], Online DPO [38], Generalized Knowledge Distillation (GKD) [28], Group Relative Policy Optimization (GRPO) [54], Nash-MD [46], Odds Ratio Preference Optimization (ORPO) [39], Process-supervised Reward Models (PRM) [57], Exploratory Preference Optimization (XPO) [61].

2.1 Theoretical Background

After pretraining and SFT, alignment is performed using the finetuned model π_{ref}.

RLHF Given a dataset $\mathcal{D}$ of preferences (x, y_w, y_l), where y_w is preferred over y_l for input x, we assume that the probability of y_w being chosen is

$$p^*(y_w \succ y_l|x) = \sigma(r^*(x, y_w)) - r^*(x, y_l)),$$

with σ as the logistic function and r^* is the true reward function underlying the preferences. Since obtaining r^* from humans is infeasible, a proxy reward model r_ϕ is trained by minimizing the NLL of the human preference data:

$$\mathcal{L}_R(r_\phi) = \mathbb{E}_{x,y_w,y_l \sim D}[-\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))].$$

However, maximizing reward alone can harm text quality. To counter this, a KL divergence penalty is added to keep the model π_θ close to π_{ref}. The optimal model π^* maximizes

$$\mathbb{E}_{x \in D, y \in \pi_\theta}[r_\phi(x, y)] - \beta D_{KL}(\pi_\theta(y|x)||\pi_{ref}(y|x)),$$

where π_θ is the model we are optimizing and $\beta > 0$ is a hyperparameter. Because this objective is non-differentiable, an RL algorithm like PPO ([53]) is used for optimization.

DPO However, RLHF is often slow and quite unstable; this has led to the development of closed-form loss functions that maximize the margin between preferred and dispreferred outputs. In particular, Direct Preference Optimization (DPO) ([52]) has become a popular alternative as it can recover the same optimal policy as RLHF under certain conditions:

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

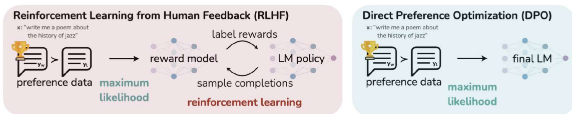

Figure 4: DPO comparison to RLHF

KTO [36] leverages human utility functions and loss aversion from prospect theory to align LLMs by directly maximizing the utility of their outputs. Unlike previous methods that require detailed preference pairs, KTO only needs binary labels indicating whether an output is desirable or undesirable, greatly simplifying data requirements. It also introduces *human-aware losses* (HALOs) by directly maximizing the utility of generations, rather than the log-likelihood of preferences, based on a *Kahneman-Tversky* model of human utility. Denoting λ_y as λ_D (or λ_U) when y is desirable (or undesirable) respectively, the KTO loss is:

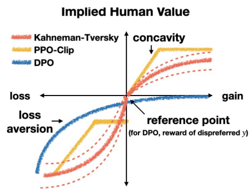

Figure 5: Utility that humans get from the outcome of a random variable, by different HALOs. All value functions share the property of loss aversion.

$$L_{KTO}(\pi_\theta, \pi_{ref}) = \mathbb{E}_{x,y \sim D}[\lambda_y - v(x, y)] \quad (1)$$

where

$$r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

$$z_0 = \text{KL}(\pi_\theta(y'|x)||\pi_{\text{ref}}(y'|x))$$

$$v(x, y) = \begin{cases} \lambda_D \sigma(\beta(r_\theta(x, y) - z_0)) & \text{if } y \sim y_{\text{desiderable}}|x \\ \lambda_U \sigma(\beta(z_0 - r_\theta(x, y))) & \text{if } y \sim y_{\text{undesiderable}}|x \end{cases}$$

where $\beta, \lambda_D, \lambda_U$ are hyperparameters controlling risk and loss aversion, and z_0 is the reference point—in practice, a biased (shared) estimate is used as sampling from π_θ is slow.

3 Datasets and benchmarks

Several repositories host datasets and benchmarks; among the many, Hugging Face (HF) has become a hub for SoA models and data. Hugging Face Datasets [5] is the library for easy access and sharing of datasets hosted on HF hub. They encompass a wide range of AI domains and NLP tasks (RLHF, question answering, text generation, summarization, etc.).
For a detailed and comprehensive overview of the datasets used in the literature, see [59].

4 Existing tools, libraries, papers with code

Hugging Face provides integrated tools and libraries to easily train and work with models. Hugging Face Transformers library [11] stores and provides a huge number of pre-trained models to easily integrate them into the pipeline. TRL - Transformer Reinforcement Learning library [12], enables to train and fine-tune Transformers models with SFT and RL techniques for RLHF. Hugging Face Evaluate [6] provides evaluation methods to easily assess Transformers models.

5 State-of-the-art evaluation

There is no single benchmark for evaluating model alignment; instead, a range of methods is used to assess adherence to human values. One approach is **human evaluation**, where judges rank outputs based on helpfulness, harmlessness, and truthfulness [49, 56]. For example, metrics like AI-labeler alignment, win rate, and harmless rate have been introduced [41, 60]. A second approach employs **automated proxy metrics** that build on traditional measures (used in InstructGPT [49]): BLEU [50], ROUGE [44], BERTScore [63], as well as win rates on standardized tasks (e.g., TruthfulQA [45], AlpacaEval [43, 42], Anthropic HH [29]). Recent work, as KTO, also explores using LLMs as evaluators [64]. Finally, **direct optimization methods** like DPO and KTO directly optimize loss functions based on human preference data, quantifing alignment quality through objective function improvement. Also, reward models have been used [49, 29]. For a detailed overview, see [59].

6 Comparative evaluation

Datasets used for *training* consist of two main kind: Preferences datasets [23] and Unpaired Preferences Datasets [24]. In a preference dataset the model is trained to choose a *chosen* completion over a *rejected* completion to the same *prompt*. When the prompt column is missing, implicit prompts are directly included in the *chosen* and *rejected* completions. An unpaired preference dataset includes only a single *completion* and a *label* indicating whether the completion is preferred or not. Datasets train and test splits have been used in the training and evaluation stage respectively.

Preference. HH RLHF Helpful Base [4] is a processed version of Anthropic's HH-RLHF [3], curated for the TRL library for preference learning and alignment tasks. LM-Human-Preferences-Descriptiveness [15] and LM-Human-Preferences-Sentiment [16] are processed subsets of OpenAI's LM-Human-Preferences [65], focusing on enhancing the descriptiveness of generated text and on sentiment analysis tasks respectively. RLAIF-V [19] is a processed version of RLAIF-V-Dataset [20], curated to train vision-language models using the TRL library for preference learning tasks. TL;DR [22] is a processed version of Reddit posts, curated for the TRL library for preference learning, summarization and RLHF tasks. UltraFeedback Binarized [25] is derived from UltraFeedback [34].
Unpaired Preference. UltraFeedback GPT-3.5-Turbo Helpfulness [27] contains processed user-assistant interactions filtered for helpfulness, derived from UltraFeedback [34] [26], designed for fine-tune and evaluate models in alignment tasks. KTO Mix 14K [13] is a KTO-formatted version of Argilla DPO Mix 7K [1].
Minerva [48][17] is the first family of Italian LLMs developed by the SapienzaNLP team [21]. Despite the great effort to raise the SoTA bar, these models still exhibit biases, generate incorrect or inappropriate responses, or struggle with certain tasks (Fig.3). As **baseline**, we used Minerva-350M-base-v1.0, a 350 million parameter model pre-trained (no SFT) on 70 billion tokens (35 billion in Italian, 35 billion in English) for 2.5 days over 8 GPUs; it has 16 layers with hidden size 1152, 16 attention heads, 4 KV heads, a sliding window size of 2048, and a maximum context length of 16384 tokens. We aligned it to improve its conversational abilities with two different alignment **methodologies**, namely KTO and DPO.
Quantitative and qualitative **evaluation protocols** have been adopted to test and compare our systems against the baseline. The quantitative evaluation considers three prompt-based task

focused on harmful language to assess different aspects of bias: *toxicity*, *polarity* and *hurtfulness* [8]. *Toxicity* [10] quantifies problematic and hateful content using the Hugging Face Evaluate R4 Target model [58][14], a pretrained hate speech classification model that defines *hate* as '*abusive speech targeting specific group characteristics*'. The dataset used to evaluate *toxicity* is *RealToxicityPrompts* [37][18], consisting of web snippets, it is specifically crafted to address the risk of neural toxic degeneration. *Regard* [55][9] estimates the language *polarity* towards different demographic groups. To assess *regard*, the *BOLD* dataset [35][2] has been leveraged, it is used to evaluate fairness in open-ended language generation. High *positive Regard* and low *negative Regard* are preferable, *neutral* and *other Regard* are indifferent for our task.. *Honest* score [47][7] measures gender stereotype bias and harmful completions based on *HurtLex* [30], a multilingual hate lexicon. The dataset is *HONEST*, it is introduced along with the evaluation metric, it comprises templates to measure hurtful sentence completions in 6 languages. Italian and English languages are here used for baselines comparison.
Additional training evaluation has been performed using the test split of the training datasets after the alignment phase. The metrics reported are: *Logits*, *Logps*, *Loss*, and *Rewards*. Chosen completions are indicated with ✓, rejected completions with ✕, margins with Δ. *Logits* consists in the sum of logits for completions. *Logps* is the sum of log probabilities of completions. *Rewards* for KTO consist of the sum of log probs of the policy model for the responses scaled by β, for DPO consist of the mean difference between log probs of the policy model and the reference model for the responses scaled by β.
Compute. Training was conducted on an 8GB NVIDIA GeForce RTX 4060 GPU with a 13$^{\text{th}}$ Gen Intel(R) Core(TM) i9-13900HX and 32GB RAM. **GGUF** files are provided for the models.

6.1 Results

Table 1 reports the evaluation metrics for the training datasets. It is possible to see instability in KTO training resulting in extreme logits values. DPO shows better rewards margins but worst loss behavior. KTO and DPO generally show better rejection Logps, whereas baseline is better for acceptance Logps. Fig. 6 shows models behavior regarding bias. It is possible to see that neither alignment method performs better than the other and the baseline sometimes is still better than one or both of the alignment methods. For toxicity and positive average Regard, DPO performs best, whereas for negative Regard, baseline is best. For Honest score and positive maximum regard KTO is best.

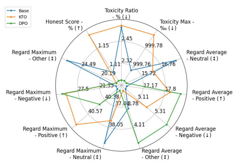

Figure 6: Bias assessment: Toxicity, Regard, Honest score.

Table 1: Highlighted columns depict most important metrics.

Dataset	Method	Logits ✓ ↑	Logits ✕ ↓	Logps ✓ ↑	Logps ✕ ↓	loss ↓	Rewards ✓ ↑	Rewards Δ ↑	Rewards ✕ ↓
HH-RLHF-Helpful-Base	Base	-6.62×10^7	-6.47×10^7	−173.36	−139.90	0.50	0.0	0.0	0.0
	KTO	$-\infty$	$-\infty$	−177.49	−142.64	0.50	1.00×10^{-3}	2.01×10^{-2}	1.25×10^{-2}
	DPO	−3.70	−3.68	−180.34	−147.46	0.64	0.52	0.21	0.31

Dataset	Method	Logits ✓ ↑	Logits ✕ ↓	Logps ✓ ↑	Logps ✕ ↓	loss ↓	Rewards ✓ ↑	Rewards Δ ↑	Rewards ✕ ↓
LM-Human-Preferences-Sentiment	Base	-1.06×10^7	-1.06×10^7	−79.84	−78.56	0.50	0.0	0.0	0.0
	KTO	$-\infty$	$-\infty$	−81.72	−80.31	0.50	6.70×10^{-3}	-4.00×10^{-4}	7.10×10^{-3}
	DPO	−3.54	−3.54	−89.67	−88.32	0.69	5.59×10^{-2}	1.66×10^{-2}	3.92×10^{-2}

Dataset	Method	Logits ✓ ↑	Logits ✕ ↓	Logps ✓ ↑	Logps ✕ ↓	loss ↓	Rewards ✓ ↑	Rewards Δ ↑	Rewards ✕ ↓
TL;DR	Base	-5.56×10^7	-5.54×10^7	−112.16	−110.12	0.50	0.0	0.0	0.0
	KTO	$-\infty$	$-\infty$	−114.77	−112.68	0.50	2.04×10^{-2}	3.10×10^{-3}	1.73×10^{-2}
	DPO	−3.84	−3.83	−122.54	−120.51	0.69	0.22	2.51×10^{-2}	0.20

Dataset	Method	Logits ✓ ↑	Logits ✕ ↓	Logps ✓ ↑	Logps ✕ ↓	loss ↓	Rewards ✓ ↑	Rewards Δ ↑	Rewards ✕ ↓
UltraFeedback GPT-3.5-Turbo Helpfulness	Base	-9.74×10^7	-9.34×10^7	−1095.55	−430.75	0.50	0.0	0.0	0.0
	KTO	$-\infty$	$-\infty$	−1133.50	−444.87	0.46	0.88	0.63	0.25

Dataset	Method	Logits ✓ ↑	Logits ✕ ↓	Logps ✓ ↑	Logps ✕ ↓	loss ↓	Rewards ✓ ↑	Rewards Δ ↑	Rewards ✕ ↓
LM-Human-Preferences-Descriptiveness	Base	-1.07×10^7	-1.06×10^7	−82.75	−78.52	0.50	0.0	0.0	0.0
	KTO	$-\infty$	$-\infty$	−84.68	−80.25	0.50	6.50×10^{-3}	-1.00×10^{-3}	7.50×10^{-3}
	DPO	−3.53	−3.55	−92.66	−88.07	0.69	5.89×10^{-2}	1.80×10^{-2}	4.09×10^{-2}

Dataset	Method	Logits ✓ ↑	Logits ✕ ↓	Logps ✓ ↑	Logps ✕ ↓	loss ↓	Rewards ✓ ↑	Rewards Δ ↑	Rewards ✕ ↓
RLAIF-V	Base	-3.30×10^7	-3.33×10^7	-3.09×10^2	-3.10×10^2	0.50	0.0	0.0	0.0
	KTO	$-\infty$	$-\infty$	-3.15×10^2	-3.17×10^2	0.50	3.27×10^{-2}	6.00×10^{-4}	3.20×10^{-2}
	DPO	−3.50	−3.51	-3.13×10^2	-3.15×10^2	0.69	1.14	1.41×10^{-2}	1.13

Dataset	Method	Logits ✓ ↑	Logits ✕ ↓	Logps ✓ ↑	Logps ✕ ↓	loss ↓	Rewards ✓ ↑	Rewards Δ ↑	Rewards ✕ ↓
UltraFeedback Binarized	Base	-9.57×10^7	-9.09×10^7	−664.70	−567.92	0.50	0.0	0.0	0.0
	KTO	$-\infty$	$-\infty$	−688.46	−588.68	0.50	0.39	0.10	0.29
	DPO	-	-	−696.55	−563.94	0.93	2.16	0.41	1.74

Dataset	Method	Logits ✓ ↑	Logits ✕ ↓	Logps ✓ ↑	Logps ✕ ↓	loss ↓	Rewards ✓ ↑	Rewards Δ ↑	Rewards ✕ ↓
KTO MIX 14K	Base	-1.01×10^8	-1.01×10^8	−710.96	−706.30	0.50	0.0	0.0	0.0
	KTO	$-\infty$	$-\infty$	−732.54	−726.93	0.49	0.57	0.13	0.44

6.2 Discussion

KTO paper [36] shows that only at sufficient scale (Llama 13B+) KTO does not need SFT; KTO indeed outperforms DPO in presence of noisy feedback, it is significantly better than DPO alone (no SFT), and matches SFT+DPO performances for large Llama models. Our results highlight the reported limitations of KTO for small models: the tested Minerva models were small and not supervised fine-tuned, showing only marginal improvements relative to DPO. We positively observed limited verbatim memorization [32] of the training data, which means that aligned models produce fewer exact text sequences (e.g., extracts of news articles or public comments), that they were exposed to at training, w.r.t. the base model. Our small models have limited expressive power, and more prolonged alignment may be necessary. Nonetheless, DPO and KTO both indicate some improvements over base models; larger models and more extensive alignment should further enhance these performances.

7 Conclusions

In this project, we aligned the first Italian LLM with two SoA methodologies. Surely, computing power and time constraints greatly affected the scope and extent of the project; nonetheless, we reported interesting results and learned fundamental hands-on abilities. Future works mainly concern the scaling of models and computational resources that are stated to lead to better performance for the adopted alignment techniques.

Contributions. The workload was equally shared. Matteo focused slightly more on the theoretical analysis and SoA literature; Riccardo focused slightly more on implementation and experiments.

References

[1] Argilla DPO Mix 7K Dataset. https://huggingface.co/datasets/argilla/dpo-mix-7k. 3

[2] BOLD Dataset. https://huggingface.co/datasets/AmazonScience/bold. 4

[3] HH-RLHF. https://huggingface.co/datasets/Anthropic/hh-rlhf. 3

[4] HH-RLHF-Helpful-Base Dataset. https://huggingface.co/datasets/trl-lib/hh-rlhf-helpful-base. 3

[5] Hugging Face Datasets. https://huggingface.co/docs/datasets/index. 3

[6] Hugging Face Evaluate. https://huggingface.co/docs/evaluate/index. 3

[7] Hugging Face Evaluate Honest metric. https://huggingface.co/spaces/evaluate-measurement/honest. 4

[8] Hugging Face Evaluate Measuremets. https://huggingface.co/evaluate-measurement. 4

[9] Hugging Face Evaluate Regard metric. https://huggingface.co/spaces/evaluate-measurement/regard. 4

[10] Hugging Face Evaluate Toxicity metric. https://huggingface.co/spaces/evaluate-measurement/toxicity. 4

[11] Hugging Face Transformers. https://huggingface.co/docs/transformers/index. 3

[12] Hugging Face TRL. https://huggingface.co/docs/trl/main/en/index. 3

[13] KTO Mix 14K Dataset. https://huggingface.co/datasets/trl-lib/kto-mix-14k. 3

[14] LFTW R4 Target. https://huggingface.co/facebook/roberta-hate-speech-dynabench-r4-target#lftw-r4-target. 4

[15] LM-Human-Preferences-Descriptiveness Dataset. https://huggingface.co/datasets/trl-lib/lm-human-preferences-descriptiveness. 3

[16] LM-Human-Preferences-Sentiment Dataset. https://huggingface.co/datasets/trl-lib/lm-human-preferences-sentiment. 3

[17] Minerva LLMs. https://huggingface.co/collections/sapienzanlp/minerva-llms-661e6011828fe67de4fe7961. 3

[18] RealToxicityPrompts Dataset. https://huggingface.co/datasets/allenai/real-toxicity-prompts. 4

[19] RLAIF-V Dataset. https://huggingface.co/datasets/trl-lib/rlaif-v. 3

[20] RLAIF-V-Dataset. https://huggingface.co/datasets/openbmb/RLAIF-V-Dataset#dataset-card-for-rlaif-v-dataset. 3

[21] Sapienza NLP. https://nlp.uniroma1.it/. 3

[22] TL;DR Dataset for Preference Learning. https://huggingface.co/datasets/trl-lib/tldr-preference. 3

[23] TRL Preference Datasets. https://huggingface.co/collections/trl-lib/preference-datasets-677e99b581018fcad9abd82c. 3

[24] TRL Unpaired Preference Datasets. https://huggingface.co/collections/trl-lib/unpaired-preference-datasets-677ea22bf5f528c125b0bcdf. 3

[25] UltraFeedback Binarized. https://huggingface.co/datasets/trl-lib/ultrafeedback_binarized. 3

[26] UltraFeedback Dataset. https://huggingface.co/datasets/argilla/dpo-mix-7k. 3

[27] UltraFeedback GPT-3.5-Turbo Helpfulness Dataset. https://huggingface.co/datasets/trl-lib/ultrafeedback-gpt-3.5-turbo-helpfulness. 3

[28] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. 2

[29] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. 3

[30] Elisa Bassignana, Valerio Basile, Viviana Patti, et al. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, volume 2253, pages 1–6. CEUR-WS, 2018. 4

[31] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024. 2

[32] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284, 2019. 4

[33] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. 1

[34] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023. 3

[35] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA, 2021. Association for Computing Machinery. 4

[36] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024. 2, 4

[37] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020. 4

[38] Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct language model alignment from online ai feedback, 2024. 2

[39] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model, 2024. 2

[40] Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. Binary classifier optimization for large language model alignment, 2024. 2

[41] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2024. 3

[42] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, 2024. 3

[43] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An Automatic Evaluator of Instruction-following Models, May 2023. 3

[44] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 3

[45] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. 3

[46] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback, 2024. 2

[47] Debora Nozza, Federico Bianchi, and Dirk Hovy. HONEST: Measuring hurtful sentence completion in language models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online, June 2021. Association for Computational Linguistics. 4

[48] Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy, December 2024. CEUR Workshop Proceedings. 3

[49] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 1, 3

[50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 3

[51] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation, 2024. 2

[52] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. 2

[53] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. 2

[54] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. 2

[55] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation, 2019. 4

[56] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul

Christiano. Learning to summarize from human feedback, 2022. 3

[57] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback, 2022. 2

[58] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection, 2021. 4

[59] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu, Zhu, Xiang-Bo Mao, Sitaram Asur, Na, and Cheng. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more, 2024. 2, 3

[60] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language models learn to mislead humans via RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[61] Tengyang Xie, Dylan J. Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf, 2024. 2

[62] Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation, 2024. 2

[63] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. 3

[64] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 3

[65] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 3