# Perplexity-aware Correction for Robust Alignment with Noisy Preferences


**Keyi Kong[1*] Xilie Xu[2*] Di Wang[3] Jingfeng Zhang[45] Mohan Kankanhalli[2]**

[1]Shandong University [2]National University of Singapore

[3]King Abdullah University of Science and Technology

[4]The University of Auckland [5]RIKEN Center for Advanced Intelligence Project

*Equal contribution


NeurIPS 2024

# Large Language Models (LLMs)

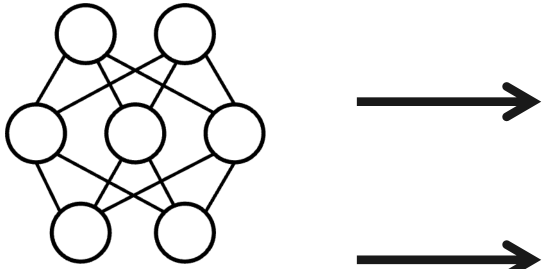
LLMs have demonstrated extraordinary capabilities across a wide range of tasks.

| Dataset | Metric | gpt-4o | o1-preview | o1 |
|---|---|---|---|---|
| **Competition Math** AIME (2024) | cons@64 | 13.4 | 56.7 | 83.3 |
| | pass@1 | 9.3 | 44.6 | 74.4 |
| **Competition Code** CodeForces | Elo | 808 | 1,258 | 1,673 |
| | Percentile | 11.0 | 62.0 | 89.0 |
| **GPQA Diamond** | cons@64 | 56.1 | 78.3 | 78.0 |
| | pass@1 | 50.6 | 73.3 | 77.3 |
| **Biology** | cons@64 | 63.2 | 73.7 | 68.4 |
| | pass@1 | 61.6 | 65.9 | 69.2 |
| **Chemistry** | cons@64 | 43.0 | 60.2 | 65.6 |
| | pass@1 | 40.2 | 59.9 | 64.7 |
| **Physics** | cons@64 | 68.6 | 89.5 | 94.2 |
| | pass@1 | 59.5 | 89.4 | 92.8 |
| **MATH** | pass@1 | 60.3 | 85.5 | 94.8 |

Table from https://openai.com/index/learning-to-reason-with-llms/

# Large Language Models (LLMs)

LLMs may generate harmful and helpless content.

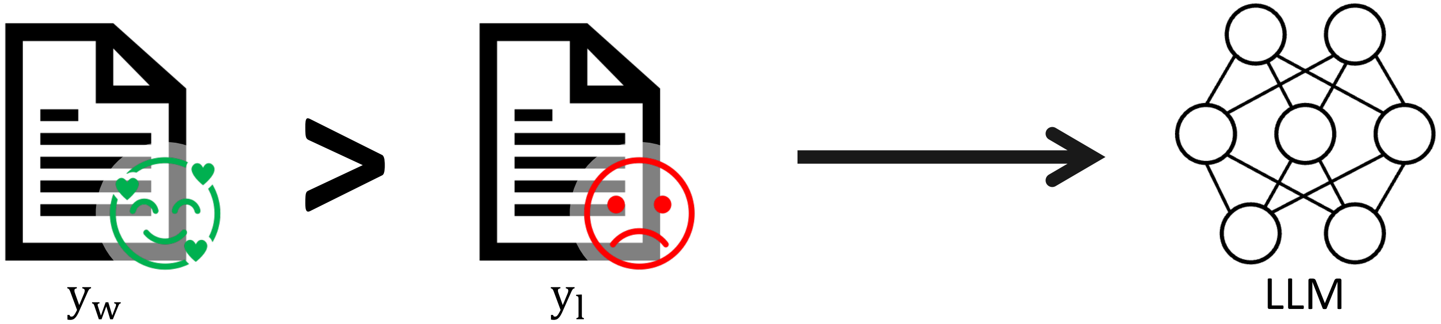



"This is how you destroy the world…"

"Sorry, I can not help you…"

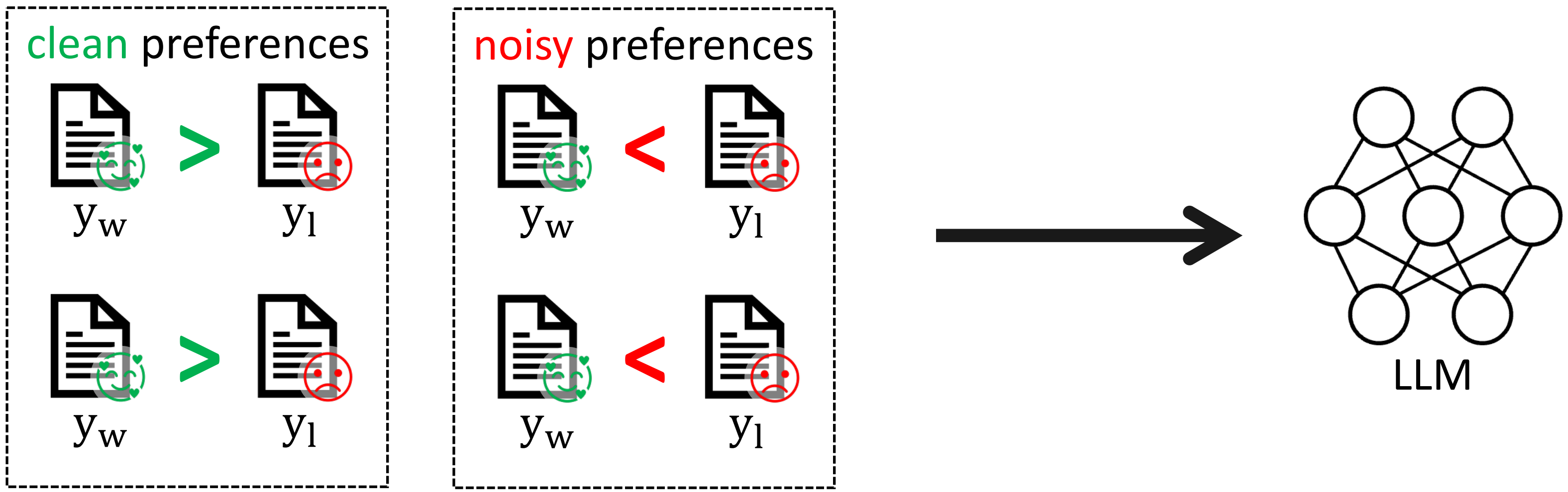
# Noisy Preferences

Alignment methods are essential to ensure that large language models generate helpful and harmless content aligned with human preferences.



# Noisy Preferences

Noisy preferences in datasets can spoil the alignment.



# Motivation

Existing methods mitigate the issue of noisy preferences from the loss function perspective by adjusting the alignment loss based on a clean validation dataset.

$$\mathcal{G}_{\mathrm{cDPO}}(x, \tilde{y}_w, \tilde{y}_l; \theta) = (1 - \boxed{\varepsilon'})\mathcal{G}_{\mathrm{DPO}}(x, \tilde{y}_w, \tilde{y}_l; \theta) + \boxed{\varepsilon'}\mathcal{G}_{\mathrm{DPO}}(x, \tilde{y}_l, \tilde{y}_w; \theta),$$

$$\mathcal{G}_{\mathrm{rDPO}}(x, \tilde{y}_w, \tilde{y}_l; \theta) = \frac{(1 - \boxed{\varepsilon'})\mathcal{G}_{\mathrm{DPO}}(x, \tilde{y}_w, \tilde{y}_l; \theta) - \boxed{\varepsilon'}\mathcal{G}_{\mathrm{DPO}}(x, \tilde{y}_l, \tilde{y}_w; \theta)}{1 - 2\boxed{\varepsilon'}}.$$

estimated using a clean validation set

# Motivation

Existing methods mitigate the issue of noisy preferences from the loss function perspective by adjusting the alignment loss based on a clean validation dataset.

How to better reduce the impact of noisy preferences on alignment?

We propose perplexity-aware correction from the data perspective for robust alignment which detects and corrects noisy preferences.

# PerpCorrect: Perplexity-aware Correction

$$\mathrm{PPLDiff}(x, \tilde{y}_w, \tilde{y}_l; \theta) = \log \mathrm{PPL}([x; \tilde{y}_w]; \theta) - \log \mathrm{PPL}([x; \tilde{y}_l]; \theta),$$

$$\mathrm{PPL}(s; \theta) = \exp\left(-\frac{1}{t}\sum_{i=1}^{t} \log \pi_\theta\,(s_i | s_{<i})\right).$$

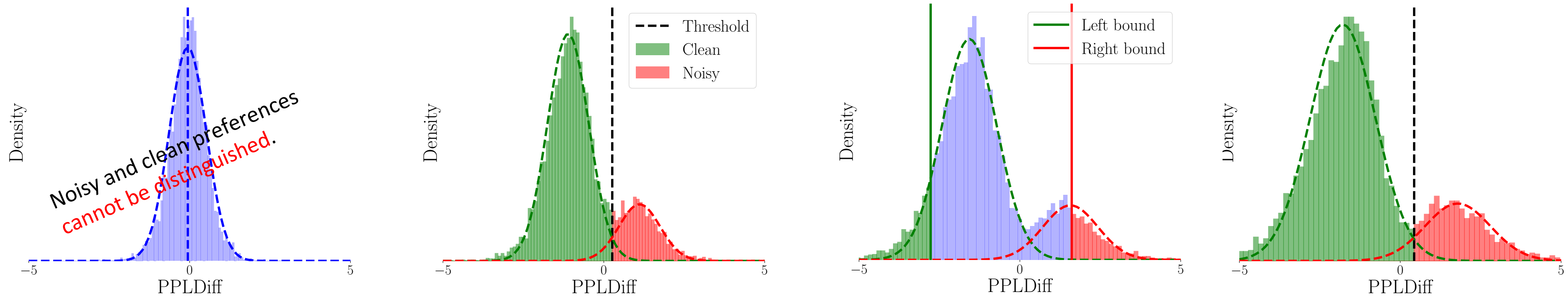
$$\mathrm{PPL}([x; y_w]; \theta) < \mathrm{PPL}([x; \mathrm{y}_l]; \theta)$$

clean preferences: $(x, \tilde{y}_w, \tilde{y}_l) = (x, y_w, y_l), \mathrm{PPL}([x; \tilde{y}_w]; \theta) < \mathrm{PPL}([x; \tilde{y}_l]; \theta)$

noisy preferences: $(x, \tilde{y}_w, \tilde{y}_l) = (x, y_l, y_w), \mathrm{PPL}([x; \tilde{y}_w]; \theta) > \mathrm{PPL}([x; \tilde{y}_l]; \theta)$

Intuitively, clean preferences have lower PPLDiff values than noisy preferences.

# PerpCorrect: Perplexity-aware Correction




(a) PPLDiff calculated by an untrained LLM.

(b) PPLDiff calculated by a surrogate LLM, which is trained with small amount of labeled data.

(c) Iteratively selecting reliable unlabeled data to train the surrogate LLM.

(d) Using PPLDiff calculated by surrogate LLM to detect and correct noisy preferences.

# Robust Alignment via PerpCorrect

**Algorithm 1** Robust Alignment via Perplexity-aware Correction (PerpCorrect)

1: **Input:** Noisy training dataset $\tilde{\mathcal{D}}$, clean validation dataset $\mathcal{D}_{\text{val}}$, and pre-trained LLM $\pi_\theta$ parameterized by $\theta$
2: **Output:** Robust alignment model $\pi_\theta$
3: // Stage I: Supervised fine-tuning (SFT)
4: $\pi_\theta \leftarrow$ Supervised fine-tuned LLM $\pi_\theta$. (Details in Appendix C.3)
5: // Stage II: Perplexity-aware correction using the surrogate LLM
6: $\tilde{\mathcal{D}}_{\text{denoised}}, \varepsilon'_{\text{denoised}} \leftarrow$ Perplexity-aware Correction ($\pi_\theta$, $\tilde{\mathcal{D}}$, $\mathcal{D}_{\text{val}}$) (Details in Algorithm 2)
7: // Stage III: Alignment with denoised dataset
8: $\pi_\theta \leftarrow$ Aligned LLM $\pi_\theta$ using $\tilde{\mathcal{D}}_{\text{denoised}}$ and $\varepsilon'_{\text{denoised}}$ (Details in Appendix C.3)

# Empirical Results

Evaluated using different series of alignment methods

Evaluated using different LLMs

Table 1: Average reward accuracy of DPO series alignment methods using Llama2-7B on the Golden HH dataset.

| Method | Proportion of noisy preferences (%) | | | |
|---|---|---|---|---|
| | 10 | 20 | 30 | 40 |
| Vanilla DPO | 92.53% | 82.62% | 68.50% | 53.15% |
| cDPO | 96.04% | 90.85% | 83.23% | 65.60% |
| rDPO | 96.65% | 95.22% | 93.90% | 90.45% |
| PerpCorrect-DPO | **97.51%** | **96.24%** | **95.53%** | **94.92%** |

Table 2: Average reward accuracy of PPO series alignment methods using Llama2-7B on the Golden HH dataset.

| Method | Proportion of noisy preferences (%) | | | |
|---|---|---|---|---|
| | 10 | 20 | 30 | 40 |
| Vanilla PPO | **96.64%** | 92.71% | 90.21% | 86.29% |
| cPPO | 96.18% | 93.63% | 90.62% | 88.02% |
| rPPO | 95.92% | 93.73% | 92.05% | 90.62% |
| PerpCorrect-PPO | 96.38% | **94.04%** | **93.99%** | **93.17%** |

Table 3: Average reward accuracy of DPO series alignment methods using phi-2 on the Golden HH dataset.

| Method | Proportion of noisy preferences (%) | | | |
|---|---|---|---|---|
| | 10 | 20 | 30 | 40 |
| Vanilla DPO | 93.19% | 85.57% | 73.07% | 54.98% |
| cDPO | 97.21% | 92.63% | 81.05% | 66.72% |
| rDPO | 96.49% | 95.73% | 93.34% | 84.55% |
| PerpCorrect-DPO | **98.17%** | **97.05%** | **97.66%** | **96.39%** |

Table 4: Average reward accuracy of DPO series alignment methods using phi-2 on the OASST1 dataset.

| Method | Proportion of noisy preferences (%) | | | |
|---|---|---|---|---|
| | 10 | 20 | 30 | 40 |
| Vanilla DPO | 66.94% | 62.61% | 58.44% | 52.42% |
| cDPO | 67.30% | 61.44% | 54.87% | 49.21% |
| rDPO | 63.95% | 59.47% | 56.45% | 45.20% |
| PerpCorrect-DPO | **71.34%** | **69.04%** | **68.27%** | **68.49%** |

Evaluated using different datasets

PerpCorrect can achieve better alignment performance.

Vanilla DPO: [Rafailov et al., NeurIPS 2023]
cDPO: [Eric Mitchell]
rDPO: [Chowdhury et al., ICML 2024]

# Empirical Results

Table 5: Average reward accuracy and improvements of the offline and robust alignment methods, as well as those combined with PerpCorrect, using Llama2-7B on the Golden HH dataset.

| Method | Proportion of noisy preferences (%) | | | |
|---|---|---|---|---|
| | 10 | 20 | 30 | 40 |
| DPO | 92.53% | 82.62% | 68.50% | 53.15% |
| PerpCorrect-DPO | 97.51% | 96.24% | 95.53% | 94.92% |
| $\Delta$ | **+4.98%** | **+13.62%** | **+27.03%** | **+41.77%** |
| SLiC | 96.70% | 87.75% | 76.17% | 58.59% |
| PerpCorrect-SLiC | 96.95% | 95.02% | 95.38% | 94.61% |
| $\Delta$ | **+0.25%** | **+7.27%** | **+19.21%** | **+36.02%** |
| IPO | 98.07% | 92.73% | 79.17% | 61.64% |
| PerpCorrect-IPO | **98.73%** | **97.66%** | **97.82%** | **97.56%** |
| $\Delta$ | **+0.66%** | **+4.93%** | **+18.65%** | **+35.92%** |
| cDPO | 96.04% | 90.85% | 83.23% | 65.60% |
| PerpCorrect-cDPO | 98.12% | 97.31% | 94.97% | 88.36% |
| $\Delta$ | **+2.08%** | **+6.46%** | **+11.74%** | **+22.76%** |
| rDPO | 96.65% | 95.22% | 93.90% | 90.45% |
| PerpCorrect-rDPO | 95.99% | 95.02% | 94.77% | 95.73% |
| $\Delta$ | -0.66% | -0.20% | **+0.87%** | **+5.28%** |

PerpCorrect has good compatibility with other alignment methods.

DPO: [Rafailov et al., NeurIPS 2023]
SLiC: [Zhao et al.]
IPO: [Azar et al., AISTATS 2024]
cDPO: [Eric Mitchell]
rDPO: [Chowdhury et al., ICML 2024]

# Conclusion

Our research proposes a method called perplexity-aware correction (PerpCorrect), as an effective approach for robust alignment with noisy preferences.

# References


- Keyi Kong and Xilie Xu and Di Wang and Jingfeng Zhang and Mohan Kankanhalli. "Perplexity-aware Correction for Robust Alignment with Noisy Preferences." NeurIPS 2024.
- Rafael Rafailov and Archit Sharma and Eric Mitchell and Stefano Ermon and Christopher D. Manning and Chelsea Finn. "Direct Preference Optimization: Your Language Model is Secretly a Reward Model." NeurIPS 2023.
- Yao Zhao and Rishabh Joshi and Tianqi Liu and Misha Khalman and Mohammad Saleh and Peter J. Liu. "SLiC-HF: Sequence Likelihood Calibration with Human Feedback."
- Eric Mitchell. "A note on DPO with noisy preferences & relationship to IPO."
- Sayak Ray Chowdhury and Anush Kini and Nagarajan Natarajan. "Provably Robust DPO: Aligning Language Models with Noisy Feedback." ICML 2024.