Aligner, Diagnose Thyself: A Meta-Learning Paradigm for Fusing Intrinsic Feedback in Preference Alignment

Mengyang Li[1] Pinlong Zhao[2] Zhong Zhang[1]

[1]Tianjin Normal University, Tianjin, China
[2]Hangzhou Dianzi University, Hangzhou, China

ICLR 2026

The Challenge: Noisy Preferences in LLM Alignment

- **Background:** Aligning LLMs (e.g., via DPO) relies heavily on preference datasets.
- **The Problem:** Datasets are plagued by *Noisy Preferences (NPs)* due to annotator disagreement, subjectivity, or AI labeling errors.
- **Limitations of Existing Robust Methods:**
  - *Coarse-grained adjustments* (e.g., cDPO, rDPO): Apply uniform global corrections, ignoring instance-specific nuances.
  - *Single-heuristic criteria* (e.g., PerpCorrect): Rely solely on one metric like Perplexity Difference (PPLDiff).
- **Our Argument:** Preference reliability is multifaceted. Relying on a single metric creates blind spots (e.g., fluent but hallucinated text).

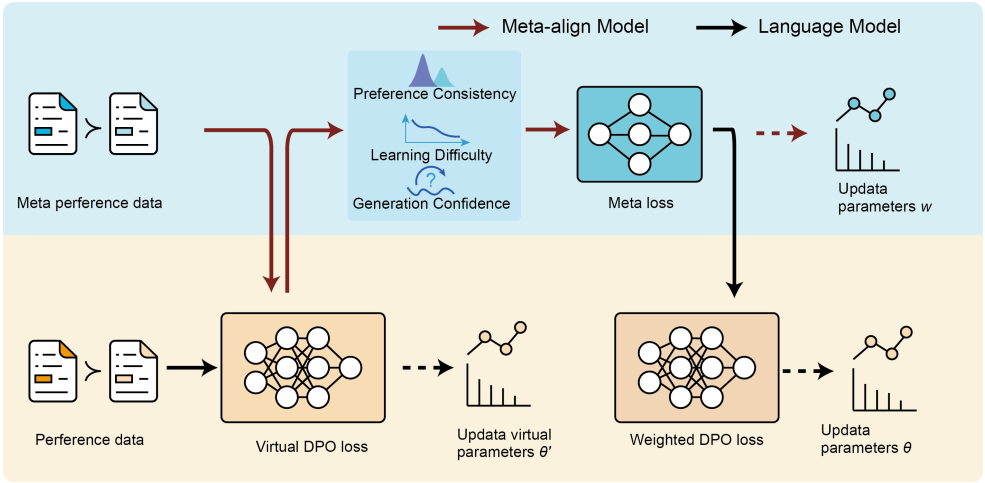
Our Paradigm: "Aligner, Diagnose Thyself"

Instead of relying on single external heuristics, we empower the model to self-diagnose by systematically fusing multiple streams of **intrinsic feedback**.

We construct a dynamic **Intrinsic Diagnostic Vector ($\mathbf{z} \in \mathbb{R}^3$)**:

1. **Preference Consistency (z_{ppl}):** Does likelihood align with the label?
   $\rightarrow$ Measured by dynamic Perplexity Difference (PPLDiff).
2. **Learning Difficulty (z_{loss}):** How easily does the model assimilate this?
   $\rightarrow$ Measured by instance-wise DPO training loss.
3. **Generation Confidence (z_{uncert}):** How certain is the generation?
   $\rightarrow$ Measured by token-level predictive entropy (Uncertainty).

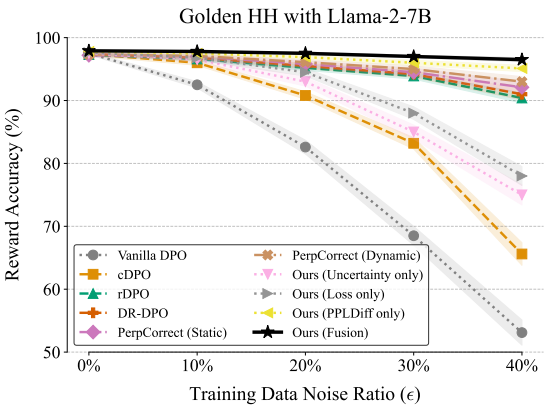
Methodology: Meta-Learning for Fusing Diagnostics



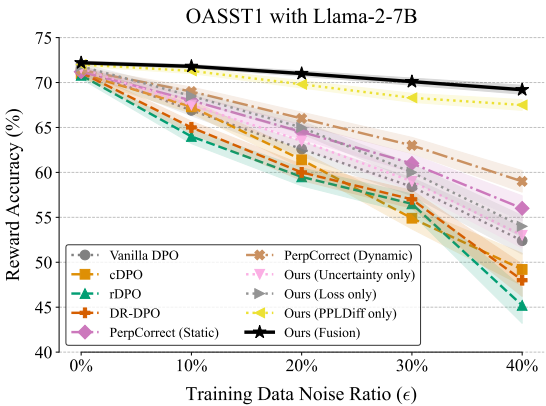
- **Inner Loop:** Meta-learner outputs sample weights based on the diagnostic vector. The main model performs a *virtual update*.

Main Results: State-of-the-Art Robustness

Reward Accuracy under Varying Noise Levels (ϵ)



Golden HH (Llama-2-7B)

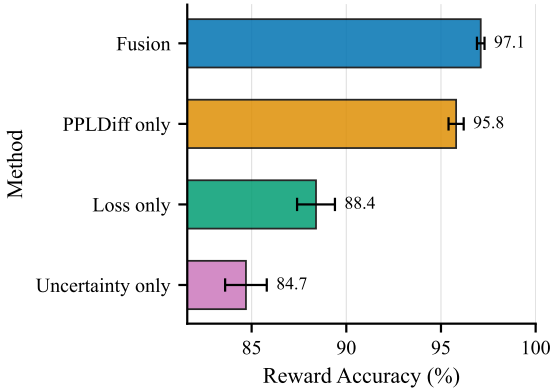


OASST1 (Llama-2-7B)

- **Ours (Fusion)** consistently establishes a new SOTA across all noise conditions.

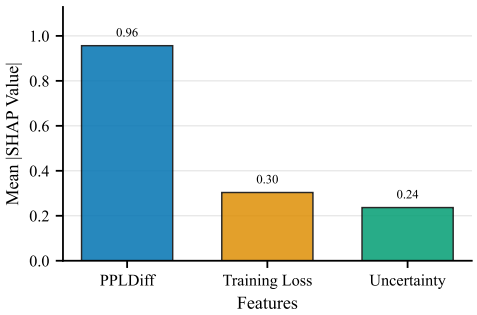
Ablation Study: The Necessity of Diagnostic Fusion

- **Question:** Is fusion necessary, or is PPLDiff enough?
- **Findings:**
  - PPLDiff is the strongest single diagnostic.
  - However, **Ours (Fusion)** significantly surpasses the PPLDiff-only variant.
  - *Insight:* Loss and Uncertainty act as crucial correctives, addressing the inherent blind spots of a single-heuristic approach.

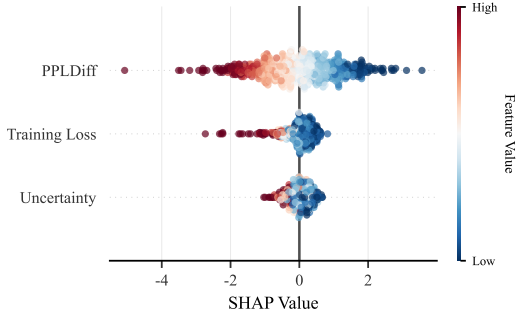


Golden HH test set (30% training noise)

In-depth Analysis: Uncovering the Interplay (SHAP)



(a) Feature Importance



(b) SHAP Value Distribution

- **Hierarchy:** PPLDiff is the primary signal, but Loss and Uncertainty exert substantial influence.
- **Non-linear Interplay:**
  - *Loss* acts as a high-impact flag (high loss strongly reduces weight).
  - *Uncertainty* acts as a nuanced modulator to temper confidence.

Scalability & Conclusion

Validation on Real-World & Large-Scale Data

- **Scalability:** Successfully scaled to **10.8M** samples (StackExchange) with only linear computational overhead.
- **Real-World Noise:** Maintains consistent SOTA improvements on naturally noisy datasets with human disagreement (WebGPT, Chatbot Arena).

Conclusion

- We shift the paradigm from single-heuristic filtering to **dynamic, multi-perspective self-diagnosis**.
- Fusing preference consistency, learning difficulty, and generation confidence provides a principled path towards robust and trustworthy LLMs.

Thank You! / Q&A