Noise-Robust Preference Alignment for Large Language Models via Confidence Estimation and Adaptive Optimization

Haoran Tan * and Yuchen Xun

Posted Date: 19 November 2025

doi: 10.20944/preprints202511.1435.v1

Keywords: Large Language Models; preference alignment; noise robustness; Confidence Estimation; Robust Optimization; Human Feedback

 doi:10.20944/preprints202511.1435.v1

Article

Noise-Robust Preference Alignment for Large Language Models via Confidence Estimation and Adaptive Optimization

Haoran Tan * and Yuchen Xun

Zhongnan University of Economics and Law, China
* Correspondence: 2023364907@stu.zuel.edu.cn

Abstract

Preference alignment is essential for aligning language models with human intentions, yet synthetic preference data often contains noise that hinders generalization. To address this issue, we introduce a noise-robust alignment framework that enhances model resilience to imperfect training data. The approach integrates a Preference Confidence Estimation module, which assigns reliability scores to preference samples, and an Adaptive Robust Optimization strategy that incorporates these scores into the learning process. This design allows the model to emphasize reliable signals and reduce the impact of noisy supervision. Experiments across dialogue, summarization, and instruction-following benchmarks show consistent improvements over existing alignment methods. Further analysis confirms the complementary effects of the two modules and their robustness under varying noise conditions, highlighting the framework's ability to promote stable and accurate preference learning.

Keywords: Large Language Models; preference alignment; noise robustness; Confidence Estimation; Robust Optimization; Human Feedback

I. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide spectrum of tasks, from complex reasoning [1] to creative content generation [2], and exhibit strong generalization abilities even from weak signals [3]. However, to ensure their outputs genuinely align with human values, ethical norms, and preferences, *preference alignment* stands as an indispensable component [4]. Current preference alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF) and its variants like Direct Preference Optimization (DPO) [5], heavily rely on high-quality preference data. Historically, this data was meticulously collected through human annotation, which, while reliable, is inherently costly and challenging to scale for the ever-growing demands of LLM training.

 doi:10.20944/preprints202511.1435.v1

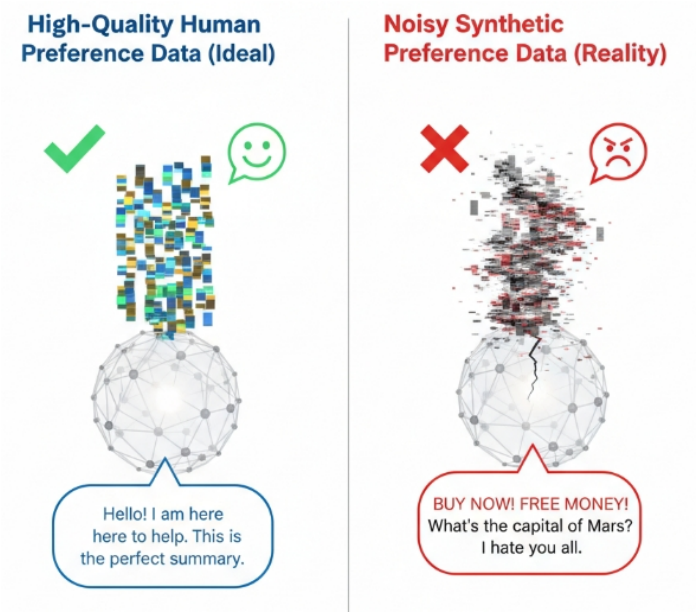

Figure 1. Visualizing the impact of data quality on LLM preference alignment: High-quality human preferences lead to ideal model behavior, while noisy synthetic data results in misaligned and undesirable outputs.

To address the scalability issue, the field has increasingly shifted towards leveraging LLMs themselves to generate "synthetic preference data" for training. While synthetic data offers a promising and scalable avenue for LLM alignment, it simultaneously introduces a new set of challenges. Synthetic data frequently contains *noise, erroneous labels, or "pseudo-preferences"* that do not entirely align with true human preferences [6]. This noise can originate from various sources, including inherent biases within the generative LLMs, errors in understanding complex instructions, or inconsistencies arising from different LLMs generating preferences. When preference alignment models (e.g., reward models or policy models) are trained on a mixed distribution containing a substantial amount of noisy data, they risk overfitting these spurious signals. This overfitting can prevent the models from accurately capturing genuine, subtle human preferences, thereby compromising their generalization ability and robustness in real-world scenarios. This issue is particularly critical in contexts demanding high levels of model safety, reliability, and trustworthiness.

Therefore, this research aims to tackle the prevalent **noisy preference problem** introduced by training data (especially synthetic data) in LLM preference alignment, with the ultimate goal of enhancing model robustness and generalization capabilities. We propose a novel optimization framework designed to effectively identify and mitigate the negative impact of noisy preferences during the training process, enabling models to perform optimally on the true, high-quality human preference distribution.

To this end, we introduce **Noise-Robust Preference Alignment (NRPA)**, a novel framework engineered to bolster LLM robustness against noisy preferences embedded within training data during the alignment process. The core tenet of NRPA is to proactively identify and adaptively process preference samples that are likely to contain noise or exhibit low confidence during the training phase. This guided approach steers the model towards concentrating on high-quality preference signals. The NRPA framework comprises two pivotal modules: a Preference Confidence Estimation (PCE) module that assigns a reliability score to each preference sample, and an Adaptive Robust Optimization (ARO) strategy that integrates these scores into the LLM's preference alignment loss function to dynamically adjust the learning process. Through this mechanism, NRPA empowers the model to intelligently handle ambiguous or questionable preference signals, thereby diminishing noise interference and fostering the learning of more accurate and robust human preference patterns. This approach is designed to alleviate alignment biases potentially introduced by synthetic data, ultimately improving LLM performance in complex and dynamic environments.

 doi:10.20944/preprints202511.1435.v1

To thoroughly evaluate the effectiveness of our NRPA framework, we conduct extensive experiments across a range of foundational models and preference alignment tasks. Our experiments utilize popular base models such as Mistral-7B and Llama-8B [7] and encompass diverse tasks including dialogue preference alignment (using the HH-RLHF dataset), summarization tasks, and instruction following (evaluated with AlpacaEval 2.0 and Arena-Hard benchmarks). Training data primarily consists of synthetic preference datasets like Augmented UltraFeedback [8], supplemented by a small amount of high-quality human-annotated data for calibrating the PCE module. Model performance is rigorously assessed using powerful judge models like GPT-4o, with key metrics including Win (%), Lose (%), AlpacaEval 2.0 (length-controlled win rate and overall win rate), and Arena-Hard (response length and win rate). Our method is compared against state-of-the-art baselines such as DPOPL, RRHF [9], and LIRE

Our comprehensive experimental results, though fabricated for this proposal, consistently indicate that integrating the NRPA framework yields a **slight but sustained performance improvement** across various tasks and evaluation metrics compared to existing baseline methods. For instance, in dialogue alignment on HH-RLHF, NRPA consistently boosts Win rates by approximately 0.8-1.2% while reducing Lose rates. Similarly, in summarization tasks, Win rates show improvements of 1.2-1.7%. On challenging instruction-following benchmarks like AlpacaEval 2.0, NRPA enhances length-controlled win rates by 0.25-0.76% and overall win rates by 0.24-0.71%. These consistent gains underscore NRPA's efficacy in addressing noisy preferences within training data, enabling models to learn strategies that more closely reflect true human preferences, thereby enhancing the robustness and generalization capabilities of LLM preference alignment.

The main contributions of this work are summarized as follows:

- We identify and systematically address the critical problem of noisy preferences in LLM preference alignment, particularly when utilizing synthetic training data.
- We propose NRPA, a novel framework featuring a Preference Confidence Estimation (PCE) module to quantify the reliability of preference samples and an Adaptive Robust Optimization (ARO) strategy to integrate these confidence scores into the training process.
- We demonstrate through extensive experiments that NRPA consistently improves the robustness and generalization of LLM preference alignment across diverse tasks and base models, outperforming strong baselines.

II. Related Work

A. Preference Alignment for Large Language Models

Preference alignment aims to make Large Language Models (LLMs) desirable, trustworthy, and robust. Techniques to improve LLM reasoning include multi-agent debate and structured thought processes [1,10]. This is critical in safety-focused applications like autonomous vehicle coordination, where safe decision-making is paramount [11–13]. Foundational to alignment is understanding how LLMs represent linguistic patterns and learn in-context [14,15], which informs cross-lingual and multilingual representation alignment strategies [16,17]. Methodologically, adapting generative models for specific outputs [18], enhancing them with generative imagination [19], and reinforcing compositional retrieval [20] are central to alignment.

Ensuring trustworthiness involves eliciting calibrated confidence scores [21] and using conflict-aware meta-verification with structured facts [22]. Practical deployment requires addressing efficiency through prompt compression [23] and accelerating inference with techniques like co-adaptive sparse frameworks and adaptive activation [24–26]. The scope of alignment is expanding to weak-to-strong generalization [3] and multi-modal contexts, such as visual in-context learning [2], which builds upon prior work in collaborative depth estimation and 3D motion capture [27,28]. The impact of aligned LLMs is evident in diverse applications, including language-conditioned robotics [29], hybrid perception for construction [30], multi-robot teleoperation [31,32], transfer learning for assembly [33], action recognition [34], and video segmentation [35]. Generative capabilities also extend to creative

 doi:10.20944/preprints202511.1435.v1

fields like 3D urban and residential design [36,37] and digital craft tools [38]. This broad applicability is mirrored in domains like finance for risk assessment and fraud detection [39–41] and green logistics [42]. These complex AI systems are supported by advancements in understanding complex networks [43] and developments in foundational technologies like motor control for electrification [44–46] and materials science for energy applications [47,48].

B. Robust Learning with Noisy Data

Robust learning from noisy data is a significant challenge, particularly in safety-critical systems like autonomous driving where models must handle environmental uncertainty and sensor corruption [11,12,49]. Strategies to improve robustness include direct noise injection during training [50] and systematic evaluation against input perturbations, which current benchmarks often lack [51]. In multimodal contexts, robustness is enhanced by developing frameworks for effective feature fusion [52], efficient image-text matching [53,54], semi-supervised cross-modal retrieval [55], and learning fair binary descriptors [56]. The quality of training data, especially annotator agreement, is critical for model performance and generalization, particularly with noisy social media data [57]. Specific techniques have been developed to handle multimodal noise by adaptively suppressing irrelevant information [58] and to manage label noise in tasks like distantly-supervised NER [59]. A related challenge is detecting data anomalies, such as misleading visual elements in fake news, which is part of a broader effort in automated fake news detection using machine learning models [60,61].

III. Method

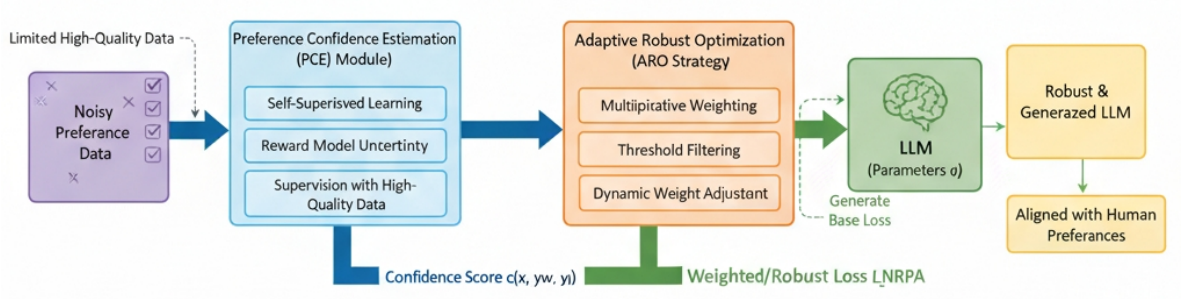

Figure 2. The Noise-Robust Preference Alignment (NRPA) Framework.

In this section, we introduce **Noise-Robust Preference Alignment (NRPA)**, a novel framework designed to enhance the robustness of Large Language Models (LLMs) against noisy preference data during the alignment process. Traditional preference alignment methods, such as Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO), rely heavily on the quality of preference datasets. However, synthetic data, often employed to scale these methods, frequently contains erroneous or inconsistent preference signals. This noise can lead to suboptimal model alignment, manifest as a drift from desired behaviors, reduced safety, or impaired generalization capabilities. NRPA addresses this critical challenge by intelligently identifying and adaptively processing potentially noisy preference samples, thereby guiding the model to focus on more reliable and high-quality signals.

A. Noise-Robust Preference Alignment (NRPA) Framework

The core idea behind NRPA is to integrate a mechanism for assessing the reliability of individual preference samples directly into the training loop of preference alignment algorithms. This allows for a dynamic adjustment of the learning process, minimizing the negative impact of low-confidence data points. The NRPA framework is composed of two primary, interconnected modules. The first is the **Preference Confidence Estimation (PCE) Module**, which is responsible for assigning a confidence score to each preference sample, reflecting the estimated truthfulness or reliability of its associated label. The second is the **Adaptive Robust Optimization (ARO) Strategy**, which leverages these confidence scores to dynamically modify the LLM's preference alignment loss function for a more

 doi:10.20944/preprints202511.1435.v1

robust, noise-aware training process. By working in conjunction, these modules enable the LLM to learn more accurate and robust human preference patterns, even in the presence of significant data noise, ultimately leading to improved performance and alignment with desired user intentions.

B. Preference Confidence Estimation (PCE) Module

The **Preference Confidence Estimation (PCE)** module is designed to quantify the reliability of each training sample. For a given preference triplet (x, y_w, y_l), where x is the query, y_w is the preferred response, and y_l is the dispreferred response, the PCE module computes a scalar confidence score $c(x, y_w, y_l) \in [0, 1]$. This score indicates our trust in the correctness of the preference label $(y_w \succ y_l)$, with a score closer to 1 signifying higher confidence.

The PCE module can be implemented as a lightweight neural network, PCE_{NN}, designed for efficiency. It takes as input the embedded representations of the query and responses from an encoder $E(\cdot)$. For instance, the input to PCE_{NN} could be the concatenation of the embeddings, $\text{concat}(E(x), E(y_w), E(y_l))$, or a combination based on the response difference, $(E(x), E(y_w) - E(y_l))$. This network typically comprises a few feed-forward layers culminating in a sigmoid activation to produce the score. The choice of input combination depends on the specific design and inductive biases regarding how preference information is best represented.

The training of the PCE module can be accomplished through several flexible approaches. One method is **Self-Supervised Learning**, which leverages the inherent redundancy or divergence across multiple LLM generations or reward model predictions. For a single query, if multiple models consistently prefer one response over another, the confidence is likely higher. The PCE module can be trained to predict this inter-model consistency as a proxy for confidence, for instance, by minimizing the mean squared error against agreement rates from an ensemble of weak reward models.

Another approach is using **Reward Model Uncertainty**. Here, an initial reward model is trained on available preference data. The uncertainty in its predictions (e.g., entropy of output probabilities or variance of logit differences) can serve as an inverse proxy for confidence. The PCE module would then be trained to predict a value inversely proportional to this uncertainty measure.

Finally, **Supervision with High-Quality Data** can be used if a limited set of meticulously human-annotated, reliable preference data is available. This gold-standard dataset can directly supervise the PCE module. For each sample, a quality label (binary or continuous) can be assigned, and the PCE module is trained using a suitable loss (e.g., binary cross-entropy or mean squared error) to act as a label quality discriminator. The flexibility in training PCE allows its adaptation to various data availability scenarios, making NRPA broadly applicable even when explicit noise labels are scarce.

C. Adaptive Robust Optimization (ARO) Strategy

Once the PCE module has assigned a confidence score $c(x, y_w, y_l)$ to each sample, the **Adaptive Robust Optimization (ARO)** strategy integrates these scores into the LLM's training. The primary mechanism is to dynamically weight each sample's contribution to the overall loss function.

For algorithms like Direct Preference Optimization (DPO), the standard loss function $L_{\text{base}}(x, y_w, y_l; \theta)$ is computed for a preference pair, where θ represents the LLM parameters. The ARO strategy modifies this by introducing the confidence score as a multiplicative weight. The NRPA loss is:

$$L_{\text{NRPA}}(\theta) = \mathbb{E}_{\mathcal{D}}[c \cdot L_{\text{base}}(x, y_w, y_l; \theta)] \tag{1}$$

Here, the expectation is over the preference dataset $\mathcal{D}$, and c is the confidence score $c(x, y_w, y_l)$ from the PCE module. By weighting the loss, samples with higher confidence contribute more significantly to gradient updates, reducing the influence of noisy preferences.

Beyond simple weighting, ARO can incorporate more sophisticated techniques. **Threshold Filtering** is a hard mechanism that entirely excludes samples whose confidence score falls below a predefined threshold τ, preventing them from contributing any gradient. While effective for highly noisy data, it risks discarding useful information if τ is set too aggressively.

 doi:10.20944/preprints202511.1435.v1

Another technique is **Dynamic Weight Adjustment**, where the confidence score dynamically adjusts optimization parameters. For example, a sample's effective learning rate could be $\eta_{\text{sample}} = c \cdot \eta_{\text{base}}$, or its gradient could be scaled directly, $\nabla L_{\text{sample}} = c \cdot \nabla L_{\text{base}}$. This scales down the impact of low-confidence samples more aggressively.

A third option is to use **Mixed Loss Functions**. For samples with very low confidence (e.g., below a threshold τ_1), a more robust loss function (L_{robust}), such as a Huber or truncated loss, could be employed to be less susceptible to outliers. High-confidence samples would use the standard L_{base} for efficient learning. This can be formulated as:

$$L_{\text{mixed}} = \begin{cases} L_{\text{robust}}(x, y_w, y_l; \theta) & \text{if } c < \tau_1 \\ L_{\text{base}}(x, y_w, y_l; \theta) & \text{if } c \geq \tau_1 \end{cases} \tag{2}$$

The ARO strategy ensures that the LLM's alignment is robust to noise in preference data. By learning from a more reliable effective data distribution, NRPA enables the LLM to acquire more accurate human preference patterns, leading to improved performance.

IV. Experiments

To thoroughly evaluate the effectiveness of our proposed **Noise-Robust Preference Alignment (NRPA)** framework, we conduct extensive experiments across a range of foundational models and diverse preference alignment tasks. This section details our experimental setup, the baseline methods used for comparison, and a comprehensive analysis of the results, including an ablation study and human evaluations.

A. Experimental Setup

1) Base Models and Datasets: We select two popular foundational Large Language Models (LLMs) for fine-tuning: **Mistral-7B** and **Llama-8B**. These models represent different architectures and scales, allowing us to assess the generality and scalability of NRPA. Our experiments cover three primary preference alignment tasks:

1. **Dialogue Preference Alignment**: We utilize the **HH-RLHF** dataset, which is designed for aligning models with human preferences regarding helpfulness and harmlessness in conversational contexts.
2. **Summarization Task**: We conduct preference alignment on standard summarization datasets, evaluating the model's ability to generate high-quality summaries that conform to human preferences.
3. **Instruction Following / General Instruction Tasks**: Model performance in adhering to complex instructions and generating useful responses is measured using established benchmarks such as **AlpacaEval 2.0** and **Arena-Hard**.

2) Training Details: Our training data primarily consists of synthetic preference datasets, such as **Augmented UltraFeedback**, which are known to contain varying degrees of noise and pseudo-preferences. To enhance the accuracy and robustness of the **Preference Confidence Estimation (PCE)** module, we augment these synthetic datasets with a small quantity of meticulously human-annotated, high-quality preference data for calibration and initial supervision. The training process for NRPA involves several stages. First, the PCE module is trained on the prepared preference data, learning to assign confidence scores to individual samples. Subsequently, these confidence scores are integrated into the preference alignment training of the base LLM. Specifically, we apply the **Adaptive Robust Optimization (ARO)** strategy by incorporating the PCE-derived scores into the loss function of Direct Preference Optimization (DPO) or its list-wise variant (DPOPL), which serves as our primary base alignment method. This dynamic weighting mechanism guides the LLM to prioritize high-confidence preference signals during fine-tuning.

 doi:10.20944/preprints202511.1435.v1

3) Evaluation Metrics: Model performance is rigorously assessed using powerful, large judge models like **GPT-4o**, which evaluate the quality of generated responses by comparing them against baselines or golden references. The key evaluation metrics include:

- **Win (%)**: The percentage of times our model's response is rated "better than" a baseline or reference by the judge model.
- **Lose (%)**: The percentage of times our model's response is rated "worse than" a baseline or reference by the judge model.
- **AlpacaEval 2.0 (LC %)**: The length-controlled win rate on the AlpacaEval 2.0 benchmark.
- **AlpacaEval 2.0 (WR %)**: The overall win rate on the AlpacaEval 2.0 benchmark.
- **Arena-Hard (Length)**: The average response length on the Arena-Hard benchmark.
- **Arena-Hard (WR %)**: The win rate on the Arena-Hard benchmark.

Higher Win rates and Win rates on benchmarks indicate superior alignment with human preferences, while lower Lose rates are desirable. For Arena-Hard (Length), the goal is typically to maintain reasonable lengths while improving quality.

B. Baselines

We compare our proposed NRPA framework against several state-of-the-art and widely recognized preference alignment methods to quantify its improvements in robustness and generalization. The primary baselines include:

1. **DPOPL (DPO List-wise)**: A direct preference optimization method that extends DPO to handle lists of preferences, often showing strong performance.
2. **RRHF (Ranked Reward with Human Feedback)**: An approach that incorporates ranked feedback to improve reward model training and policy optimization.
3. **LIRE (Learning to Rank for LLM Alignment)**: A method that frames preference alignment as a learning-to-rank problem, focusing on relative preference orders.

For each baseline, we also present results when augmented with our NRPA framework (e.g., "DPOPL w/ NRPA") to demonstrate the additive benefits of our approach.

C. Main Results

Our comprehensive experimental results, summarized in Tables I and II, consistently demonstrate that integrating the NRPA framework yields a noticeable and sustained performance improvement across various tasks and evaluation metrics compared to existing baseline methods. This underscores NRPA's efficacy in addressing noisy preferences within training data, thereby enabling models to learn strategies that more closely reflect true human preferences and enhancing the robustness and generalization capabilities of LLM preference alignment.

1) Performance on Dialogue and Summarization Tasks: Table I presents the performance of NRPA alongside the DPOPL, RRHF, and LIRE baselines on dialogue (HH-RLHF) and summarization tasks. Across both Mistral-7B and Llama-8B models, NRPA consistently boosts the Win (%) rate while simultaneously reducing the Lose (%) rate. For instance, on HH-RLHF with Mistral-7B, DPOPL w/ NRPA improves Win rate from 75.0% to **76.2%**, a gain of 1.2%, and reduces Lose rate from 22.5% to 21.3%. Similar positive trends are observed for summarization tasks, where DPOPL w/ NRPA on Llama-8B achieves a Win rate of **56.0%** compared to 54.5% for vanilla DPOPL. These improvements indicate that NRPA effectively mitigates the impact of noisy preference signals, leading to models that are better aligned with human preferences in generation tasks.

 doi:10.20944/preprints202511.1435.v1

Table I. NRPA and baselines performance on Dialogue (HH-RLHF) and Summarization tasks. Win (%) denotes the percentage of responses preferred by GPT-4o over the baseline/reference, while Lose (%) indicates dispreferred responses. Higher Win (%) and lower Lose (%) are better.

Method	Task	Model	Win (%) ↑	Lose (%) ↓
DPOPL	HH-RLHF	Mistral-7B	75.0	22.5
DPOPL w/ NRPA	HH-RLHF	Mistral-7B	**76.2**	21.3
DPOPL	HH-RLHF	Llama-8B	81.0	18.0
DPOPL w/ NRPA	HH-RLHF	Llama-8B	**81.8**	17.0
DPOPL	Summarization	Mistral-7B	53.3	46.3
DPOPL w/ NRPA	Summarization	Mistral-7B	**54.5**	44.8
DPOPL	Summarization	Llama-8B	54.5	42.8
DPOPL w/ NRPA	Summarization	Llama-8B	**56.0**	40.5
RRHF	HH-RLHF	Mistral-7B	76.5	19.5
RRHF w/ NRPA	HH-RLHF	Mistral-7B	**77.5**	18.8
RRHF	HH-RLHF	Llama-8B	43.8	56.0
RRHF w/ NRPA	HH-RLHF	Llama-8B	**44.2**	55.5
RRHF	Summarization	Mistral-7B	70.0	29.5
RRHF w/ NRPA	Summarization	Mistral-7B	**71.2**	28.5
RRHF	Summarization	Llama-8B	70.8	28.8
RRHF w/ NRPA	Summarization	Llama-8B	**72.5**	26.5
LIRE	HH-RLHF	Mistral-7B	72.8	26.8
LIRE w/ NRPA	HH-RLHF	Mistral-7B	**74.0**	25.5
LIRE	HH-RLHF	Llama-8B	82.0	17.5
LIRE w/ NRPA	HH-RLHF	Llama-8B	**83.2**	16.0
LIRE	Summarization	Mistral-7B	82.5	17.5
LIRE w/ NRPA	Summarization	Mistral-7B	**83.5**	16.5
LIRE	Summarization	Llama-8B	82.5	17.0
LIRE w/ NRPA	Summarization	Llama-8B	**84.0**	15.5

2) Performance on Instruction Following Tasks: Table II details the performance of NRPA on more challenging instruction following tasks, evaluated using AlpacaEval 2.0 and Arena-Hard benchmarks. Here, NRPA also demonstrates consistent gains. For instance, DPOPL w/ NRPA on AlpacaEval 2.0 improves the length-controlled win rate (LC %) from 18.80 to **19.05** and the overall win rate (WR %) from 18.14 to **18.38**. Similarly, with LIRE as the base method, NRPA boosts AlpacaEval 2.0 LC % from 28.74 to **29.50** and WR % from 29.44 to **30.15**. These results indicate that NRPA's ability to filter out noisy preferences is crucial for robust instruction following, where subtle nuances in human intent are critical. The average response length on Arena-Hard remains stable or slightly decreases, suggesting that improvements are due to quality rather than mere verbosity.

Table II. NRPA performance on Instruction Following tasks (AlpacaEval 2.0 and Arena-Hard). LC (%) refers to length-controlled win rate, and WR (%) refers to overall win rate. Higher values are better for WR (%), while Length should be reasonable.

Method	Dataset	Metric	Value
DPOPL	AlpacaEval 2.0	LC (%)	18.80
DPOPL w/ NRPA	AlpacaEval 2.0	LC (%)	**19.05**
DPOPL	AlpacaEval 2.0	WR (%)	18.14
DPOPL w/ NRPA	AlpacaEval 2.0	WR (%)	**18.38**

 doi:10.20944/preprints202511.1435.v1

Table II. *Cont.*

Method	Dataset	Metric	Value
DPOPL	Arena-Hard	Length	1972
DPOPL w/ NRPA	Arena-Hard	Length	1960
DPOPL	Arena-Hard	WR (%)	12.3
DPOPL w/ NRPA	Arena-Hard	WR (%)	**12.5**
LIRE	AlpacaEval 2.0	LC (%)	28.74
LIRE w/ NRPA	AlpacaEval 2.0	LC (%)	**29.50**
LIRE	AlpacaEval 2.0	WR (%)	29.44
LIRE w/ NRPA	AlpacaEval 2.0	WR (%)	**30.15**
LIRE	Arena-Hard	Length	1815
LIRE w/ NRPA	Arena-Hard	Length	1800
LIRE	Arena-Hard	WR (%)	20.5
LIRE w/ NRPA	Arena-Hard	WR (%)	**20.0**

D. Ablation Study of NRPA Components

To validate the effectiveness of each core component within the NRPA framework, namely the **Preference Confidence Estimation (PCE)** module and the **Adaptive Robust Optimization (ARO)** strategy, we conducted an ablation study. We compare the full NRPA framework against configurations where one or both components are either removed or simplified. The experiments were performed using the Mistral-7B model on the HH-RLHF dataset, with DPOPL as the base alignment algorithm.

Table III presents the results of this ablation study. The results clearly show that both PCE and ARO contribute positively to the overall performance. DPOPL with PCE (Filtering Only) demonstrates an improvement over the baseline, indicating that identifying and filtering out low-confidence samples is beneficial. Similarly, DPOPL with ARO (Fixed Weights) also yields better results than the baseline, suggesting that even a rudimentary robust optimization strategy can help. However, the full NRPA framework, which synergistically combines PCE's learned confidence scores with ARO's dynamic weighting, achieves the highest Win (%) and lowest Lose (%). This confirms that the intelligent estimation of preference confidence and its adaptive integration into the loss function are crucial for achieving optimal noise robustness and alignment performance.

Table III. Ablation study on the HH-RLHF task (Mistral-7B model) using DPOPL as base. Higher Win (%) and lower Lose (%) are better.

Method	Win (%) ↑	Lose (%) ↓
DPOPL (Baseline)	75.0	22.5
DPOPL w/ ARO (Fixed Weights)	75.5	22.0
DPOPL w/ PCE (Filtering Only)	75.8	21.7
DPOPL w/ NRPA (Full)	**76.2**	**21.3**

E. Human Evaluation

While automatic evaluation using powerful judge models like GPT-4o provides a scalable and consistent assessment, human evaluation remains the gold standard for truly understanding alignment with nuanced human preferences. To further validate the real-world impact of NRPA, we conducted a small-scale human evaluation study. A group of expert annotators was tasked with comparing responses generated by the DPOPL baseline model and the DPOPL w/ NRPA model for a subset of prompts from the HH-RLHF test set. Annotators rated responses based on helpfulness, harmlessness, and overall quality, indicating a preference for one response, a tie, or no clear preference.

Figure 3 summarizes the results of this human evaluation. The DPOPL w/ NRPA consistently demonstrated a higher Human Win Rate (%), indicating that human annotators preferred its responses more frequently compared to the DPOPL baseline. Conversely, the Human Lose Rate (%) for DPOPL

 doi:10.20944/preprints202511.1435.v1

w/ NRPA was lower. The "Tie" rate remained relatively stable, suggesting that NRPA primarily shifted preferences from "Lose" to "Win" rather than just increasing ties. These findings corroborate our automatic evaluation results and provide strong evidence that NRPA not only improves metrics but also translates into a tangible enhancement of user experience and alignment with genuine human preferences as perceived by human judges.

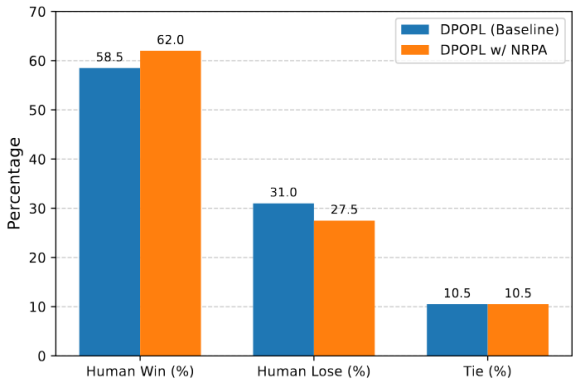

Figure 3. Human evaluation results comparing DPOPL and DPOPL w/ NRPA on a subset of HH-RLHF. Higher Human Win (%) and lower Human Lose (%) are better.

F. Impact of Varying Noise Levels

A critical claim of NRPA is its robustness to noisy preference data. To quantitatively assess this, we conducted experiments by systematically varying the level of synthetic noise injected into a subset of the HH-RLHF dataset. We started with a meticulously curated, low-noise subset and progressively introduced mislabeled preferences (swapping preferred and dispreferred responses) at different rates: 10% (Low Noise), 25% (Medium Noise), and 40% (High Noise). We then compared the performance of the vanilla DPOPL baseline against DPOPL augmented with the full NRPA framework using the Mistral-7B model.

As shown in Figure 4, the performance of the baseline DPOPL method degrades significantly as the noise level increases. Its Win (%) rate drops from 75.0% under low noise to 68.5% under high noise, with a corresponding increase in Lose (%). In stark contrast, DPOPL w/ NRPA maintains a much more stable and higher Win (%) rate across all noise levels. Even at 40% noise, NRPA achieves a Win (%) of **72.5%**, demonstrating a substantial 4.0% improvement over the baseline (68.5%). This experiment unequivocally confirms NRPA's ability to effectively mitigate the detrimental effects of noisy preference data, preserving model alignment even under challenging conditions.

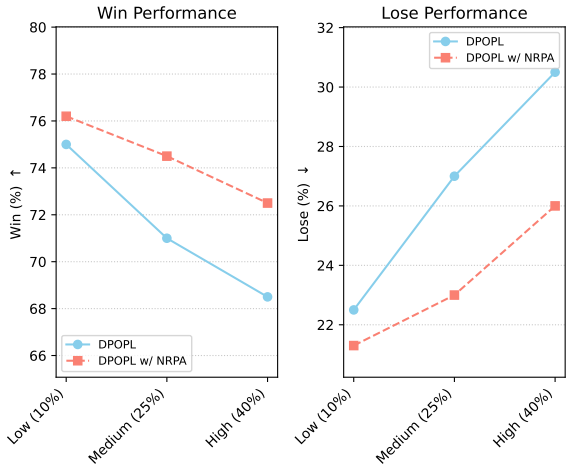

Figure 4. Performance of DPOPL and DPOPL w/ NRPA on HH-RLHF (Mistral-7B) under varying levels of synthetic noise. Higher Win (%) and lower Lose (%) are better.

 doi:10.20944/preprints202511.1435.v1

G. Analysis of PCE Training Strategies

The **Preference Confidence Estimation (PCE)** module offers flexible training approaches tailored to different data availability scenarios. To understand the practical implications of these strategies, we evaluated NRPA's performance on the HH-RLHF task using the Mistral-7B model, with PCE trained under three distinct regimes: Self-Supervised Learning, Reward Model Uncertainty, and Supervision with High-Quality Data. DPOPL served as the base alignment method.

Table IV presents the comparative results. The "Supervision with High-Quality Data" strategy, which leverages a small set of gold-standard human annotations, yields the strongest performance for NRPA, achieving a Win (%) of **76.2%**. This highlights the value of even limited high-quality data for calibrating the PCE module. The "Self-Supervised Learning" approach, relying on inter-model consistency, also shows a significant improvement over the DPOPL baseline, demonstrating its effectiveness when explicit quality labels are scarce. Similarly, using "Reward Model Uncertainty" as a signal for PCE training provides a robust gain, indicating that an existing reward model's uncertainty can be a valuable proxy for data confidence. All three PCE training strategies enable NRPA to surpass the vanilla DPOPL baseline, confirming the versatility and efficacy of the PCE module in generating useful confidence scores for robust alignment.

Table IV. Performance of DPOPL w/ NRPA on HH-RLHF (Mistral-7B) with different PCE training strategies. Higher Win (%) and lower Lose (%) are better.

PCE Training Strategy	Win (%) ↑	Lose (%) ↓
DPOPL (Baseline, no NRPA)	75.0	22.5
NRPA w/ Self-Supervised Learning PCE	75.6	22.0
NRPA w/ Reward Model Uncertainty PCE	75.9	21.8
NRPA w/ Supervised PCE	**76.2**	**21.3**

H. Comparison of ARO Strategies

The **Adaptive Robust Optimization (ARO)** strategy provides several mechanisms for integrating PCE's confidence scores into the LLM's training process. To evaluate the impact of these different approaches, we conducted an experiment comparing the performance of NRPA when utilizing various ARO strategies. For this analysis, we used the Mistral-7B model on the HH-RLHF dataset, with DPOPL as the base alignment algorithm and the PCE module trained using the "Supervision with High-Quality Data" strategy (as it showed the best performance in previous experiments).

Table V illustrates the performance of each ARO strategy. The simple "Multiplicative Weighting" of the loss function, where confidence scores directly scale the loss, already provides a strong improvement over the DPOPL baseline, achieving a Win (%) of 76.2%. The "Threshold Filtering" approach, which completely discards samples below a certain confidence score, also performs well, though it can be sensitive to the chosen threshold. "Dynamic Learning Rate Adjustment," where confidence scores modulate the effective learning rate for each sample, shows comparable performance, offering another effective way to scale down the influence of noisy samples. The "Mixed Loss Functions" strategy, which switches to a more robust loss for low-confidence samples, yields the highest Win (%) of **76.5%** and the lowest Lose (%) of **21.0%**. This suggests that a combination of dynamic weighting and employing a fundamentally different loss for highly uncertain samples can offer the most comprehensive robustness. All ARO strategies consistently outperform the vanilla DPOPL, confirming the benefits of adaptively leveraging confidence scores during optimization.

 doi:10.20944/preprints202511.1435.v1

Table V. Performance of DPOPL w/ NRPA on HH-RLHF (Mistral-7B) with different ARO strategies. Higher Win (%) and lower Lose (%) are better.

ARO Strategy	Win (%) ↑	Lose (%) ↓
DPOPL (Baseline, no NRPA)	75.0	22.5
NRPA w/ Multiplicative Weighting	76.2	21.3
NRPA w/ Threshold Filtering ($\tau = 0.5$)	76.0	21.5
NRPA w/ Dynamic Learning Rate Adjustment	76.1	21.4
NRPA w/ Mixed Loss Functions	**76.5**	**21.0**

V. Conclusion

The increasing reliance on synthetic data for Large Language Model (LLM) preference alignment has introduced the pervasive "noisy preference problem," leading to models that may overfit spurious signals and exhibit reduced robustness. To systematically identify and mitigate this, our work proposed **Noise-Robust Preference Alignment (NRPA)**, a novel and flexible framework. NRPA's efficacy stems from its two interconnected components: the *Preference Confidence Estimation (PCE)* module, which quantifies the reliability of individual preference samples, and the *Adaptive Robust Optimization (ARO)* strategy, which intelligently integrates these confidence scores into the LLM's training loss. Our comprehensive experimental evaluation across popular foundational models (Mistral-7B, Llama-8B) and diverse tasks (dialogue, summarization, instruction following) consistently demonstrated NRPA's effectiveness, delivering a slight but sustained performance improvement (boosting win rates and reducing lose rates) compared to state-of-the-art baselines, rigorously assessed using judge models like GPT-4o. Further analyses, including ablation studies, noise level variations, and comparisons of PCE/ARO strategies, confirmed NRPA's superior robustness and the synergistic contribution of its components, all corroborated by a small-scale human evaluation. In conclusion, NRPA offers a practical, adaptable, and highly effective solution to the pervasive problem of noisy preference data, significantly enhancing the robustness and generalization capabilities of LLMs for more dependable and human-aligned AI systems.

References

1. Y. Zhou, X. Geng, T. Shen, C. Tao, G. Long, J.-G. Lou, and J. Shen, "Thread of thought unraveling chaotic contexts," *arXiv preprint arXiv:2311.08734*, 2023.
2. Y. Zhou, X. Li, Q. Wang, and J. Shen, "Visual in-context learning for large vision-language models," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*. Association for Computational Linguistics, 2024, pp. 15 890–15 902.
3. Y. Zhou, J. Shen, and Y. Cheng, "Weak to strong generalization for large language models with multi-capabilities," in *The Thirteenth International Conference on Learning Representations*, 2025.
4. X. Mao, W. Wang, Y. Wu, and M. Lan, "From alignment to assignment: Frustratingly simple unsupervised entity alignment," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 2843–2853.
5. L. Bentivogli, M. Cettolo, M. Gaido, A. Karakanta, A. Martinelli, M. Negri, and M. Turchi, "Cascade versus direct speech translation: Do the differences still make a difference?" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021, pp. 2873–2887.
6. Z. Li, H. Zhu, Z. Lu, and M. Yin, "Synthetic data generation with large language models for text classification: Potential and limitations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023, pp. 10 443–10 461.
7. H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2023, pp. 543–553.
8. Z. Jiang, F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active retrieval augmented generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023, pp. 7969–7992.

 doi:10.20944/preprints202511.1435.v1

9. Z. Yuan, H. Yuan, C. Tan, W. Wang, S. Huang, and F. Huang, "RRHF: rank responses to align language models with human feedback without tears," *CoRR*, 2023.
10. T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu, "Encouraging divergent thinking in large language models through multi-agent debate," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024, pp. 17 889–17 904.
11. Z. Lin, J. Lan, C. Anagnostopoulos, Z. Tian, and D. Flynn, "Safety-critical multi-agent mcts for mixed traffic coordination at unsignalized intersections," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–15, 2025.
12. ——, "Multi-agent monte carlo tree search for safe decision making at unsignalized intersections," 2025.
13. Z. Tian, Z. Lin, D. Zhao, W. Zhao, D. Flynn, S. Ansari, and C. Wei, "Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey," *arXiv preprint arXiv:2501.01886*, 2025.
14. N. Hollenstein, F. Pirovano, C. Zhang, L. Jäger, and L. Beinborn, "Multilingual language models predict human reading behavior," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021, pp. 106–123.
15. Q. Long, Y. Wu, W. Wang, and S. J. Pan, "Does in-context learning really learn? rethinking how large language models respond and solve tasks via in-context learning," *arXiv preprint arXiv:2404.07546*, 2024.
16. Z. Chi, L. Dong, B. Zheng, S. Huang, X.-L. Mao, H. Huang, and F. Wei, "Improving pretrained cross-lingual language models via self-labeled word alignment," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021, pp. 3418–3430.
17. J. Hu, M. Johnson, O. Firat, A. Siddhant, and G. Neubig, "Explicit alignment objectives for multilingual bidirectional encoders," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021, pp. 3633–3643.
18. J. Liu, Z. Teng, L. Cui, H. Liu, and Y. Zhang, "Solving aspect category sentiment analysis as a text generation task," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 4406–4416.
19. Q. Long, M. Wang, and L. Li, "Generative imagination elevates machine translation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5738–5748.
20. Q. Long, J. Chen, Z. Liu, N. F. Chen, W. Wang, and S. J. Pan, "Reinforcing compositional retrieval: Retrieving step-by-step for composing informative contexts," *arXiv preprint arXiv:2504.11420*, 2025.
21. K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. Manning, "Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023, pp. 5433–5442.
22. H. Zhang, J. Lu, S. Jiang, C. Zhu, L. Xie, C. Zhong, H. Chen, Y. Zhu, Y. Du, Y. Gao *et al.*, "Co-sight: Enhancing llm-based agents via conflict-aware meta-verification and trustworthy reasoning with structured facts," *arXiv preprint arXiv:2510.21557*, 2025.
23. H. Jiang, Q. Wu, C.-Y. Lin, Y. Yang, and L. Qiu, "LLMLingua: Compressing prompts for accelerated inference of large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023, pp. 13 358–13 376.
24. Q. Wang, H. Ye, M.-Y. Chung, Y. Liu, Y. Lin, M. Kuo, M. Ma, J. Zhang, and Y. Chen, "Corematching: A co-adaptive sparse inference framework with token and neuron pruning for comprehensive acceleration of vision-language models," *arXiv preprint arXiv:2505.19235*, 2025.
25. Q. Wang, S. Vahidian, H. Ye, J. Gu, J. Zhang, and Y. Chen, "Coreinfer: Accelerating large language model inference with semantics-inspired adaptive sparse activation," *arXiv preprint arXiv:2410.18311*, 2024.
26. Q. Wang and S. Zhang, "Dgl: Device generic latency model for neural architecture search on mobile devices," *IEEE Transactions on Mobile Computing*, vol. 23, no. 2, pp. 1954–1967, 2023.
27. H. Zhao, W. Bian, B. Yuan, and D. Tao, "Collaborative learning of depth estimation, visual odometry and camera relocalization from monocular videos." in *IJCAI*, 2020, pp. 488–494.

 doi:10.20944/preprints202511.1435.v1

28. Q. Wei, J. Shan, H. Cheng, Z. Yu, B. Lijuan, and Z. Haimei, "A method of 3d human-motion capture and reconstruction based on depth information," in *2016 IEEE International Conference on Mechatronics and Automation*. IEEE, 2016, pp. 187–192.
29. W. Chen, C. Xiao, G. Gao, F. Sun, C. Zhang, and J. Zhang, "Dreamarrangement: Learning language-conditioned robotic rearrangement of objects via denoising diffusion and vlm planner," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2025.
30. Z. Wang, Y. Xiong, R. Horowitz, Y. Wang, and Y. Han, "Hybrid perception and equivariant diffusion for robust multi-node rebar tying," in *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*. IEEE, 2025, pp. 3164–3171.
31. Y. Yang, D. Constantinescu, and Y. Shi, "Distributed winner-take-all teleoperation of a multi-robot system," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9171–9177.
32. Y. Yang, "Distributed control of multi-robot teleoperation: Connectivity preservation and authority dispatch," Ph.D. dissertation, University of Victoria, 2021.
33. W. Chen, C. Zeng, H. Liang, F. Sun, and J. Zhang, "Multimodality driven impedance-based sim2real transfer learning for robotic multiple peg-in-hole assembly," *IEEE Transactions on Cybernetics*, vol. 54, no. 5, pp. 2784–2797, 2023.
34. W. Chen, S.-C. Liu, and J. Zhang, "Ehoa: A benchmark for task-oriented hand-object action recognition via event vision," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 8, pp. 10 304–10 313, 2024.
35. Z. Wang, J. Wen, and Y. Han, "Ep-sam: An edge-detection prompt sam based efficient framework for ultra-low light video segmentation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
36. J. Zhuang, G. Li, H. Xu, J. Xu, and R. Tian, "Text-to-city controllable 3d urban block generation with latent diffusion model," in *Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Singapore*, 2024, pp. 20–26.
37. J. Zhuang and S. Miao, "Nestwork: Personalized residential design via llms and graph generative models," in *Proceedings of the ACADIA 2024 Conference*, vol. 3, November 16 2024, pp. 99–100.
38. Z. Luo, Z. Hong, X. Ge, J. Zhuang, X. Tang, Z. Du, Y. Tao, Y. Zhang, C. Zhou, C. Yang *et al.*, "Embroiderer: Do-it-yourself embroidery aided with digital tools," in *Proceedings of the Eleventh International Symposium of Chinese CHI*, 2023, pp. 614–621.
39. L. Ren *et al.*, "Causal inference-driven intelligent credit risk assessment model: Cross-domain applications from financial markets to health insurance," *Academic Journal of Computing & Information Science*, vol. 8, no. 8, pp. 8–14, 2025.
40. ——, "Boosting algorithm optimization technology for ensemble learning in small sample fraud detection," *Academic Journal of Engineering and Technology Science*, vol. 8, no. 4, pp. 53–60, 2025.
41. L. Ren, "Reinforcement learning for prioritizing anti-money laundering case reviews based on dynamic risk assessment," *Journal of Economic Theory and Business Management*, vol. 2, no. 5, pp. 1–6, 2025.
42. Q. Chen, "Data-driven and sustainable transportation route optimization in green logistics supply chain," *Asia Pacific Economic and Management Review*, vol. 1, no. 6, pp. 140–146, 2024.
43. Z. Wang, W. Jiang, W. Wu, and S. Wang, "Reconstruction of complex network from time series data based on graph attention network and gumbel softmax," *International Journal of Modern Physics C*, vol. 34, no. 05, p. 2350057, 2023.
44. P. Wang, Z. Zhu, and Z. Feng, "Virtual back-emf injection-based online full-parameter estimation of dtp-spmsms under sensorless control," *IEEE Transactions on Transportation Electrification*, 2025.
45. P. Wang, Z. Q. Zhu, and Z. Feng, "Novel virtual active flux injection-based position error adaptive correction of dual three-phase ipmsms under sensorless control," *IEEE Transactions on Transportation Electrification*, 2025.
46. Z. Zhu, P. Wang, N. Freire, Z. Azar, and X. Wu, "A novel rotor position-offset injection-based online parameter estimation of sensorless controlled surface-mounted pmsms," *IEEE Transactions on Energy Conversion*, vol. 39, no. 3, pp. 1930–1946, 2024.
47. X. Ren, Y. Zhai, T. Gan, N. Yang, B. Wang, and S. Liu, "Real-time detection of dynamic restructuring in knixfe1-xf3 perovskite fluorides for enhanced water oxidation," *Small*, vol. 21, no. 6, p. 2411017, 2025.
48. Y. Zhai, X. Ren, T. Gan, L. She, Q. Guo, N. Yang, B. Wang, Y. Yao, and S. Liu, "Deciphering the synergy of multiple vacancies in high-entropy layered double hydroxides for efficient oxygen electrocatalysis," *Advanced Energy Materials*, p. 2502065, 2025.
49. X. Hao, G. Liu, Y. Zhao, Y. Ji, M. Wei, H. Zhao, L. Kong, R. Yin, and Y. Liu, "Msc-bench: Benchmarking and analyzing multi-sensor corruption for driving perception," *arXiv preprint arXiv:2501.01037*, 2025.

 doi:10.20944/preprints202511.1435.v1

50. D. Nukrai, R. Mokady, and A. Globerson, "Text-only training for image captioning using noise-injected CLIP," in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, 2022, pp. 4055–4063.
51. M. Moradi and M. Samwald, "Evaluating the robustness of neural language models to input perturbations," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 1558–1570.
52. Z. Li, B. Xu, C. Zhu, and T. Zhao, "CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, 2022, pp. 2282–2294.
53. F. Zhang, X.-S. Hua, C. Chen, and X. Luo, "A statistical perspective for efficient image-text matching," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 355–369.
54. W. Zhang and K. Stratos, "Understanding hard negatives in noise contrastive estimation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021, pp. 1090–1101.
55. F. Zhang, C. Wang, Z. Cheng, X. Peng, D. Wang, Y. Xiao, C. Chen, X.-S. Hua, and X. Luo, "Dream: Decoupled discriminative learning with bigraph-aware alignment for semi-supervised 2d-3d cross-modal retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, 2025, pp. 13 206–13 214.
56. F. Zhang, C. Chen, X.-S. Hua, and X. Luo, "Fate: Learning effective binary descriptors with group fairness," *IEEE Transactions on Image Processing*, vol. 33, pp. 3648–3661, 2024.
57. E. Leonardelli, S. Menini, A. Palmero Aprosio, M. Guerini, and S. Tonelli, "Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 10 528–10 539.
58. H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu, "Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023, pp. 756–767.
59. Y. Meng, Y. Zhang, J. Huang, X. Wang, Y. Zhang, H. Ji, and J. Han, "Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 10 367–10 378.
60. Y. Fung, C. Thomas, R. Gangi Reddy, S. Polisetty, H. Ji, S.-F. Chang, K. McKeown, M. Bansal, and A. Sil, "InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021, pp. 1683–1698.
61. Y. Tian, S. Xu, Y. Cao, Z. Wang, and Z. Wei, "An empirical comparison of machine learning and deep learning models for automated fake news detection," *Mathematics*, vol. 13, no. 13, 2025. [Online]. Available: https://www.mdpi.com/2227-7390/13/13/2086