mathematics

MDPI

Article

An Intelligent Construction Method for Petrochemical Datasets Based on RLHF and Data De-Identification

Yimin Liu, Qike Ji, Shengbo Lu and Jianliang Chen *

School of Electronic Information, Wuhan University of Science and Technology, Wuhan 430000, China; liuyimin@wust.edu.cn (Y.L.); jiqike@wust.edu.cn (Q.J.);lushengbo@wust.edu.cn (S.L.)
* Correspondence: chenjianliang@wust.edu.cn

Abstract

To mitigate high expert-annotation costs, domain-preference misalignment, and the inherent trade-off between sensitive-data protection and training utility in petrochemical dataset construction, an iterative framework combining human-feedback-aligned reinforcement learning (RLHF) with post hoc data de-identification is proposed. Direct scoring and pairwise preference feedback are generated using two high-capability language models. A reward model is subsequently trained via a joint Bradley–Terry and mean-squared-error loss, followed by three rounds of closed-loop proximal policy optimization (PPO) constrained by a fixed supervised fine-tuning reference model. Retained high-quality samples are then processed through a four-stage post-RLHF de-identification pipeline. Experimental results demonstrate that the PPO-V3 model achieves a reward score increase of 1.183 over the baseline alongside a 96.2% pairwise win rate, while the sensitivity-aware adaptive differential privacy with context-aware token-level injection (SA-ADP-CTI) post-RLHF de-identification method attains a composite score of 0.9636. The reliability of both the automated feedback and privacy-preservation mechanisms is further validated through blind expert review and manual spot checks.

Keywords: RLHF; petrochemical dataset; data de-identification; named entity recognition

MSC: 68T05; 68T07; 68T50; 68P27; 68T09

1. Introduction

Against the backdrop of deep integration between industrial intelligence and the digital economy, intelligent upgrades in equipment operation and maintenance (O&M), production control, and fault management are being rapidly accelerated across the petrochemical industry—a core pillar of modern process manufacturing. High-quality, domain-specific datasets serve as an indispensable foundation for training, validating, and deploying advanced artificial intelligence (AI) models designed for petrochemical equipment fault diagnosis, predictive maintenance, and intelligent O&M decision support [1,2].

Despite their critical importance, the systematic collection and standard construction of petrochemical data are severely constrained by inherent data characteristics, including extreme heterogeneity, geographical dispersion, deep domain specialization, and high privacy sensitivity [3,4]. Structurally, these data span multi-source textual and tabular repositories, including equipment operation manuals, historical maintenance and inspection records, and real-time O&M work orders. Semantically, highly specialized content is embedded within these sources, encompassing complex equipment parameters, underlying chemical

Check for updates

Academic Editor: Raymond Lee

Received: 24 July 2026
Revised: 18 August 2026
Accepted: 27 August 2026
Published: 2 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

process mechanisms, and stringent safety protocols. Furthermore, sensitive organizational information, such as plant identifiers, proprietary company names, and critical process performance indicators, is frequently intertwined throughout the data corpus [3,4].

In high-risk, continuous-flow production environments, incorrect model predictions triggered by noisy, inconsistent, or uncurated training data can lead to catastrophic production disruptions, equipment damage, or severe safety incidents. Consequently, the synthesis and curation of datasets that simultaneously satisfy the demanding metrics of high quality, scalable annotation, and rigorous privacy compliance remain a paramount bottleneck obstructing the broader intelligent transformation of the process industry [5].

Substantial progress has recently been achieved in both domain-specific dataset construction and language model alignment. Regarding dataset creation, annotation costs have been significantly mitigated through the utilization of large language models (LLMs) for automated data labeling and synthetic data generation [6,7]. Furthermore, recent empirical evidence, such as the LIMA framework, demonstrates that data quality exerts a far greater influence on downstream performance than sheer dataset size, thereby highlighting the critical imperative of curated, high-quality alignment corpora [8].

To align generative models with human preferences, Reinforcement Learning from Human Feedback (RLHF) has emerged as a foundational methodology [9]. Under this framework, human preferences are explicitly quantified via a learned reward model, after which the target model is iteratively refined using policy optimization algorithms such as Proximal Policy Optimization (PPO) [10]. In general-domain settings, this standard RLHF paradigm has been shown to substantially enhance model adherence to expert human criteria [11]. More recently, the versatility of RLHF has been broadened to specialized operational scenarios; frameworks such as Helios [12], CoRLHF [13], UP-RLHF [14], and MM-RLHF [15] have adapted preference alignment to energy systems, collaborative optimization, multi-modal interactions, and other complex domains.

In parallel, stringently regulated privacy requirements have spurred the development of specialized alignment paradigms. To eliminate reliance on sensitive inputs, privacy-preserving substitution mechanisms have been developed to replace authentic prompt instructions [16]. Concurrently, federated RLHF architectures have been introduced to accomplish collective preference alignment across distributed entities without transmitting raw data [17]. Together, these privacy-aware methodologies establish promising avenues for harmonizing data confidentiality with advanced model alignment.

However, applying these paradigms to construct high-quality, compliant datasets in specialized industrial verticals—such as the petrochemical sector—presents three major challenges. First, existing dataset construction methodologies struggle to reconcile quality, scalability, and root-cause optimization. While purely manual annotation yields high-quality labels, it remains prohibitively expensive and difficult to scale [7]. Conversely, direct generation using general-purpose LLMs followed by simple filtering offers high scalability but is compromised by substantial domain bias and elevated error rates [6]. Active learning provides a trade-off between these two extremes [18]; nevertheless, it functions merely as a sample-level correction mechanism, failing to address root causes at the generative-model level or to establish closed-loop iteration. Second, feedback cost constraints and reward-scale drift co-occur. Due to the limited availability of petrochemical domain experts for large-scale labeling, low-cost feedback sources are required. Meanwhile, reward models are susceptible to absolute-score drift and catastrophic forgetting across multiple iterations [19], thereby invalidating cross-round reward comparisons. Third, the interdependency between de-identification and RLHF has not been systematically investigated. De-identification relying solely on named-entity recognition provides insufficient privacy protection, whereas blanket masking of all entities degrades semantic integrity and

leads to over-de-identification [16,17,20–22]. Furthermore, automated tools exhibit high false-negative rates regarding sensitive information [16]. Applying de-identification prior to RLHF disrupts preference ordering and reward learning, while executing it post-RLHF requires preserving the training utility of the aligned data [21].

To address these critical challenges, an iterative Reinforcement Learning from Human Feedback (RLHF) dataset-construction framework integrated with a post hoc data de-identification pipeline is proposed, utilizing multi-answer question-answering data constructed from real-world corporate petrochemical documentation. A dual-large-language-model (dual-LLM) feedback is introduced to eliminate high human annotation costs and subjective bias; a joint-loss reward model is formulated to mitigate reward-scale drift and catastrophic forgetting across iterative updates; a three-round Proximal Policy Optimization (PPO) closed loop under a fixed reference constraint is designed to suppress reward hacking while progressively sharpening domain-alignment precision; and a post-RLHF privacy governance pipeline is deployed to preserve maximum dataset utility while guaranteeing strict regulatory compliance. The primary contributions of this work are summarized as follows:

- Iterative RLHF Closed Loop for Cost Efficiency and Stability. High feedback costs and reward-scale drift are mitigated through a joint reward model (RM) combining Bradley–Terry (BT) pairwise and mean squared error (MSE) direct-score losses. Preference ranking is prioritized via dynamic loss weighting ($\alpha = 0.2$), while catastrophic forgetting is prevented using a mixed replay strategy, yielding a 1.9 percentage point accuracy gain over BT-only training. A fixed supervised fine-tuning (SFT) anchor ($\beta = 0.05$) is maintained across all PPO rounds to suppress reward hacking. Consequently, PPO-V3 achieves a 1.183 reward gain over the base model and a 96.2% pairwise win rate.
- Post-RLHF De-identification Framework. Preference-ranking distortions caused by early de-identification perturbations are mitigated by establishing an "RLHF-first, de-identification-later" sequence, which improves bilingual evaluation understudy (BLEU-4) by 0.105 over pre-processing. A four-stage framework is developed, integrating RoBERTa-wwm-ext-CRF (conditional random field, CRF) entity recognition (F1 = 0.9779), context-sensitive differential privacy (DP), dynamic authorization, and auditing. An end-to-end composite score of 0.9636 is achieved, outperforming full-entity masking by 0.223.
- Expert Validation and External Benchmarking. Closed-loop evaluation bias is minimized through independent assessments of 500 responses by three domain experts. Strong Pearson correlations are demonstrated between expert ratings and both dual-LLM feedback ($r = 0.78$) and RM-V3 scores ($r = 0.66$), with complete ranking alignment across five models. Furthermore, a 92.1% pass rate from manual audits of 150 de-identified samples confirms the system's external reliability.

2. Methods

2.1. Overall Framework

To systematically resolve three persistent bottlenecks in the construction of petrochemical question-answering datasets—namely, the prohibitive costs associated with expert manual annotation, the severe misalignment between generative outputs and stringent industrial standards, and the fundamental trade-off between sensitive-information protection and downstream model training utility—an integrated data-curation framework combining iterative Reinforcement Learning from Human Feedback (RLHF) with post-RLHF data de-identification is proposed. The end-to-end operational workflow is schematically illustrated in Figure 1.

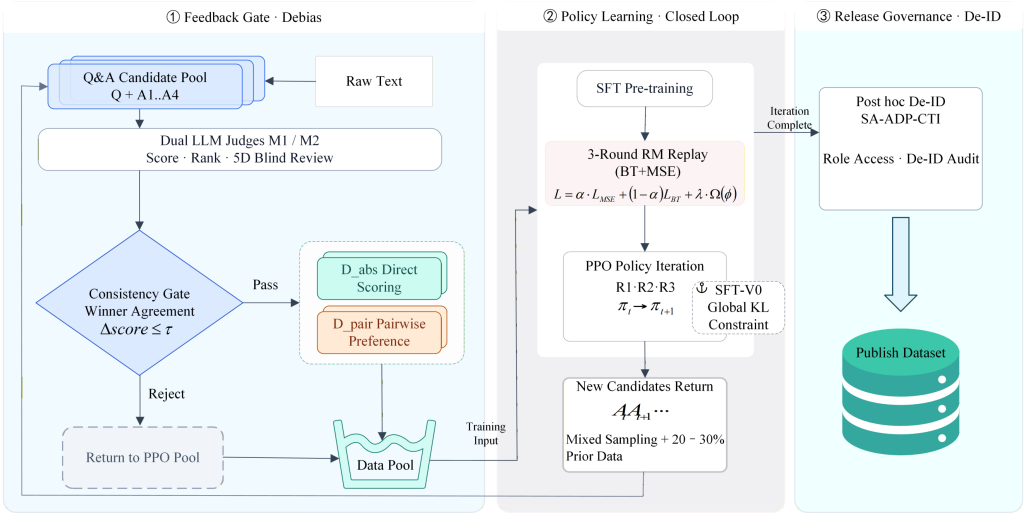


Figure 1. Overall workflow for constructing a domain-specific question-answering dataset.

The proposed architecture is structured into three strictly sequential modules designed to optimize data quality, policy alignment, and privacy preservation in turn. First, high-capability large language models (LLMs) are deployed within the feedback-data processing and sample-construction module to perform multi-dimensional scoring, systematic debiasing, cascaded filtering, and stringent quality control, thereby synthesizing calibrated direct-scoring and pairwise-preference training samples. Second, the iterative Reinforcement Learning from Human Feedback (RLHF) module is executed, wherein a joint-loss reward model is trained on these structured feedback datasets, and subsequent Proximal Policy Optimization (PPO) policy updates are iteratively conducted under a fixed SFT reference constraint to ensure scale stability and prevent reward hacking. Finally, upon the completion of preference alignment, the post-RLHF de-identification module is engaged to implement sensitive-entity recognition, context-aware differential-privacy perturbation, role-authorized dynamic masking, and rigorous privacy auditing; this post hoc sequencing explicitly prevents early perturbation noise from corrupting the reward-learning objective or distorting domain preference orderings. Detailed formulations and operational procedures for these three modules are presented as follows, respectively.

2.2. Feedback-Data Construction Using Two High-Capability Models

Raw samples first undergo two-stage quality filtering based on agreement between the two models' scores and an overall quality threshold. The retained data are then stratified by sample type and data category and placed in the data pool for subsequent training.

2.2.1. Raw Sample Format

Each basic feedback unit is a multi-candidate sample consisting of one question and four candidate answers. A raw sample is represented as follows:

$$S_i = \left\{ Q_i, A_i^1, A_i^2, A_i^3, A_i^4 \right\} \tag{1}$$

Here, Q_i denotes a petrochemical-domain question, and A_i^1–A_i^4 are four candidate answers with graded quality levels: A_i^1 is a high-information-density reference answer that cites specific standards and parameters; A_i^2 is a good answer whose core content is correct but lacks key details; A_i^3 is a partial answer that covers only one aspect; and A_i^4 is a weak answer that is directionally correct but expressed only in general terms. This graded structure provides a natural quality signal for constructing pairwise preferences.

Table 1 presents a representative raw sample from the constructed dataset, illustrating the graded quality levels, the scores independently assigned by the two models, and the feedback derived after filtering.

Table 1. An illustrative multi-answer sample with dual-model scores.

Element	Content	s_1	s_2	Final
Q	What are the main differences between centrifugal and reciprocating compressors in the selection of operating-condition monitoring parameters?	—	—	—
$A^{(1)}$	Reciprocating units emphasize rod-drop (axial sinking), lubricating-oil header pressure, packing temperature, and suction/discharge pressure and temperature; centrifugal units focus on axial displacement, dry-gas-seal primary seal-gas flow rate, flare pressure and flow rate, and hydrogen purity and suction-scrubber level for the recycle-hydrogen compressor. Both monitor bearing temperature and vibration, but with different priorities.	9.5	9.0	0.925
$A^{(2)}$	Both unit types track bearing temperature and vibration, but reciprocating units weight rod drop, lubricating-oil header pressure, and packing temperature more heavily, whereas centrifugal units weight axial displacement, dry-gas-seal parameters, and several process parameters.	8.5	8.5	0.850
$A^{(3)}$	The differences mainly lie in monitoring rod drop, packing, and lubricating oil for reciprocating units versus axial displacement and dry-gas-seal-related parameters for centrifugal units.	7.5	7.5	0.750
$A^{(4)}$	Different units have different focuses; although reciprocating and centrifugal units share common parameters, their key monitoring contents differ.	6.5	4.0	0.525

As Table 1 shows, the four candidates form a strictly descending quality gradient. $A^{(1)}$ enumerates the specific monitored parameters of both unit types and receives the highest scores; $A^{(2)}$ states the shared parameters and the differing emphases correctly but omits specifics such as the flare system and hydrogen purity; $A^{(3)}$ covers only the parameter categories without naming any concrete quantity; and $A^{(4)}$ merely restates that the two differ without any domain content. The sample passes all filters in Section 2.2.3: the inter-model discrepancies of the medium- and high-quality candidates (0.5, 0.0, 0.0) are well below the threshold of 1.5, and the relaxed threshold of 2.5 for the low-quality candidate is exactly met ($\text{diff}_i^{(4)} = 2.5$), illustrating why a looser bound is necessary for weak answers. After averaging and normalization, the sample yields four direct-score supervision tuples $(Q, A^{(k)}, s^{(k)})$ with $s = 0.925/0.850/0.750/0.525$; together with the preference pairs extracted from the candidate ranking (e.g., $A^{(1)} \succ A^{(4)}$, $A^{(2)} \succ A^{(4)}$), on which both models agree about the winner label, it supplies both sample types defined in Equations (3) and (4).

2.2.2. Independent Scoring by Two High-Capability Models

Two heterogeneous high-capability language models—M_1 (Claude Sonnet 4.6) and M_2 (GPT-5.4 thinking)—serve as feedback sources and independently perform two evaluation tasks. The first is direct scoring: for each A_i^k, model m assigns an overall quality score $s_{m,i}^k$ on a 1–10 scale and a confidence score $c_{m,i}^k$, where $m \in \{1, 2\}$. The second is pairwise ranking: for any candidate pair $\left(A_i^j, A_i^k\right)$, the model outputs the tuple $\{\text{winner}, s_A, s_B, \text{conf}\}$.

The evaluation dimensions, in descending order of importance, are correctness (40%), completeness (30%), and usefulness (30%). Outputs strictly follow JavaScript Object Notation (JSON) format, and scores are assigned in increments of 0.5.

2.2.3. Debiasing, Filtering, and Feedback-Sample Definition

To reduce systematic single-model bias and noise, the raw multi-answer samples undergo cascaded filtering and quality filtering after independent scoring by the two high-capability models. The filtering procedure comprises four components.

Inter-model score-discrepancy filtering. For each A_i^k, calculate the absolute difference between the two models' scores:

$$\text{diff}_i^{(k)} = \left| s_{1,i}^{(k)} - s_{2,i}^{(k)} \right| \tag{2}$$

For medium- and high-quality candidate answers ($k \in \{1,2,3\}$), apply $\text{diff}_i^{(k)} \leq 1.5$; for low-quality candidate answers ($k = 4$), relax the threshold to $\text{diff}_i^{(k)} \leq 2.5$. This is because the two models inherently disagree more when evaluating low-quality answers, and an overly strict threshold would substantially reduce recall.

Confidence filtering. Require the confidence assigned by each model to satisfy $c_{m,i}^{(k)} \geq 0.7$ and the mean confidence across the two models to satisfy $\bar{c}_i^{(k)} \geq 0.75$.

Preference-consistency validation. Retain a pairwise sample only when M_1 and M_2 agree completely on the winner label; discard samples with conflicting judgments or ties.

Question-level data splitting. The training and validation sets are split 8:2 at the question level, ensuring that all pairwise combinations derived from the same question remain in one split and preventing inflated validation metrics caused by same-question leakage.

After filtering, the feedback data comprise two sample types:

$$D_{\text{rate}} = \left\{ \left(Q_i, A_i^{(k)}, \text{final_score}_i^{(k)} \right) \right\} \tag{3}$$

$$D_{\text{pref}} = \left\{ \left(Q_i, A_i^{(j)}, A_i^{(k)} \right) \right\} \tag{4}$$

Here, D_{rate} denotes the direct-score supervision set used to train the reward model, and D_{pref} denotes the pairwise-preference ranking supervision set. In R1, the reward-model training and validation data are constructed separately from direct-score samples and preference samples. The same qualified answer may appear in both types of samples, thereby improving the utilization of feedback data.

2.3. Iterative RLHF Optimization

Using the direct-score and preference-ranking samples constructed in Section 2.2, we model RLHF optimization as a coupled closed loop of reward modeling and policy optimization. The reward model r_ϕ converts feedback from the two high-capability LLMs into an optimizable scalar reward signal, while the PPO policy model π_θ undergoes multiple policy updates under that signal and a fixed Kullback–Leibler (KL) constraint, progressively approaching domain-expert preferences in petrochemical question-answering.

2.3.1. Joint-Loss Reward Model

The reward model $r_\phi(\cdot)$ uses the SFT-V0 model produced during SFT as its base. The language model's final layer is replaced with a scalar output head to form an AutoModelForSequenceClassification (num_labels = 1) architecture. LoRA adapters (rank = 8, alpha = 16, dropout = 0.05) are applied to projection matrices W_q and W_v. Only the LoRA parameters are trained to ensure efficient graphics processing unit (GPU) memory use and training. Given a question–answer pair (x, y), the reward function is defined as follows:

$$r_\phi(x, y) = \mathbf{w}^\top \mathbf{h}_\phi([x; y]) + b \tag{5}$$

Here, $\mathbf{h}_\phi(\cdot)$ is the hidden state output by the backbone at the final sequence position, and $\mathbf{w}, b$ are the parameters of the linear head.

For the two sample types constructed in Section 2.2, we define a joint training objective that combines the Bradley–Terry pairwise loss with a mean squared error loss for direct scores. This enables the reward model to establish an absolute scoring coordinate system while learning relative preference rankings. The Bradley–Terry pairwise loss is applied to $\mathcal{D}_{\text{pref}}$:

$$\mathcal{L}_{\text{BT}} = -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{pref}}} \log \sigma\big(r_\phi(x,y_w) - r_\phi(x,y_l)\big) \tag{6}$$

where y_w is the preferred answer, y_l is the non-preferred answer, and $\sigma(\cdot)$ is the sigmoid function. The mean squared error loss is applied to $\mathcal{D}_{\text{rate}}$, with each score s normalized as $\frac{s}{10}$:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{(x,y,s)\sim\mathcal{D}_{\text{rate}}} \left(r_\phi(x,y) - \frac{s}{10}\right)^2 \tag{7}$$

The total objective is a weighted combination with L2 regularization:

$$\mathcal{L}_{\text{total}} = \alpha\,\mathcal{L}_{\text{MSE}} + (1-\alpha)\,\mathcal{L}_{\text{BT}} + \lambda\,\Omega(\phi) \tag{8}$$

where $\alpha \in [0,1]$ is the loss-balancing coefficient, and $\Omega(\phi)$ is the L2 regularization term for the LoRA parameters. This joint loss gives the RM both relative-ranking capability and an absolute scoring anchor, avoiding the "correct ranking but drifting reward scale" problem that arises when only BT loss is used.

To mitigate reward-scale drift and catastrophic forgetting during iterative RM training, we adopt a three-round reward-data strategy characterized by equal-weight initialization in R1, ranking-dominant optimization in R2/R3, incremental LoRA training, replay of preference and direct-score samples, and reuse of a fixed direct-score validation set. Accordingly, α was set to 0.5 in R1 and to 0.2 in R2 and R3, shifting the objective from equal weighting in R1 to ranking-dominant training in the later rounds.

The hybrid objective in Equation (8) admits a two-level theoretical characterization. In the space of reward outputs, each pairwise summand of $\mathcal{L}_{\text{BT}}$ is the softplus of an affine function—convex and smooth with second derivative at most $1/4$—so $\mathcal{L}_{\text{BT}}$ is convex with a Lipschitz-continuous gradient, while $\alpha\,\mathcal{L}_{\text{MSE}}$ is strongly convex on the direct-score coordinates; the unregularized hybrid is thus a convex, smooth, lower-bounded function of the reward outputs. This structure also clarifies the role of the MSE term: the Bradley–Terry likelihood is invariant to additive reward shifts and identifies rewards only up to an additive gauge—the formal origin of the "correct ranking but drifting reward scale" pathology—whereas the quadratic anchor pins the reward to the normalized expert-score scale and restores identifiability, so the joint objective is a scale-anchored Bradley–Terry maximum-likelihood estimator with a unique population minimizer under realizability. In the parameter space, $\mathcal{L}_{\text{total}}(\phi)$ is non-convex but smooth (only the LoRA increments are trained, so the induced Jacobian is bounded) and coercive ($\lambda\,\Omega(\phi)$ grows without bound), so minimizers exist. For smooth, lower-bounded, possibly non-convex objectives, the AdamW optimizer used for reward-model training (Section 3.2) converges to a first-order stationary point at the classical $\mathcal{O}\left(1/\sqrt{T}\right)$ rate [23,24], consistent with the stable convergence observed under all three loss configurations in Section 4.4.

2.3.2. PPO Policy Optimization

We model text generation as a Markov decision process:

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle \tag{9}$$

Here, the state $s_t \in \mathcal{S}$ is the concatenation of the question and the generated answer prefix, i.e., $s_t = [Q; y_{<t}]$; the action $a_t \in \mathcal{A}$ selects the next token from vocabulary $\mathcal{V}$; the transition function P performs deterministic concatenation, $s_{t+1} = [s_t; a_t]$; the policy $\pi_\theta(a \mid s)$ is the generation model to be optimized; and the discount factor is $\gamma = 0.99$. The RM supplies a terminal reward only when generation ends, upon reaching the end of sequence (EOS) or the maximum length:

$$r_t = \begin{cases} 0, & t < T \\ r_\phi(Q, y_{1:T}), & t = T \end{cases} \tag{10}$$

A reward is issued only at the end of the sequence; the immediate reward for every intermediate token is 0.

Policy optimization uses PPO's clipped surrogate objective:

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(\rho_t(\theta)\hat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{11}$$

where:

$$\rho_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \tag{12}$$

The advantage $\hat{A}_t$ is estimated using generalized advantage estimation (GAE):

$$\hat{A}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}, \quad \delta_t = r_t + \gamma V_\psi(s_{t+1}) - V_\psi(s_t) \tag{13}$$

We set $\lambda = 0.95$. The value function V_ψ shares the policy backbone and is updated by minimizing the temporal-difference mean squared error.

To prevent reward hacking against the RM and divergence from the original semantic space, we add a token-level KL-divergence penalty to the reward signal:

$$\tilde{r}_t = r_t - \beta\,\text{KL}(\pi_\theta(\cdot \mid s_t) \,\|\, \pi_{\text{ref}}(\cdot \mid s_t)) \tag{14}$$

Here, the KL coefficient is $\beta = 0.05$, and target_kl $= 0.1$ is used as a soft constraint. All three rounds share the same fixed reference model, SFT-V0:

$$\pi_{\text{ref}}^{(R1)} = \pi_{\text{ref}}^{(R2)} = \pi_{\text{ref}}^{(R3)} = \pi_{\text{SFT-V0}} \tag{15}$$

This design prevents the reference model from drifting with the PPO policy from the previous round and keeps the scale of the KL constraint consistent across all three rounds. It therefore improves the comparability of reward distributions, length statistics, and convergence behavior across rounds. Rewards are also clipped to $[-5.0, +5.0]$ to prevent extreme RM outputs on out-of-distribution samples.

These two mechanisms jointly bound the policy deviation from the fixed reference $\pi_{\text{ref}} = \pi_{\text{SFT-V0}}$. At the token level, the clipped surrogate objective restricts the per-update probability ratio to $\rho_t(\theta) \in [1-\epsilon, 1+\epsilon]$, so a single policy update can alter the log-probability of any token by at most $\log(1+\epsilon) - \log(1-\epsilon)$. At the trajectory level, the KL penalty $\beta\,\text{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})$ with $\beta = 0.05$ and the soft constraint target_kl $= 0.1$ monitor the same deviation on average: once the expected per-token KL approaches 0.1, the penalty term dominates the reward and halts further drift, confining each round to a trust region of approximately 0.1 nats per token around $\pi_{\text{SFT-V0}}$. The fixed reference shared by all rounds keeps this bound anchored across the closed loop, which is the mechanism-level guarantee that successive policies remain comparable and free from cumulative drift.

2.3.3. Three-Round Iterative Closed Loop

The three-round PPO closed loop is shown in Figure 1. R1 is initialized from SFT-V0 and trains RM-V1/PPO-V1. In R2 and R3, the preceding policy generates additional answers, new preference pairs are selected, and training continues from the previous round's RM LoRA weights, ultimately producing PPO-V3.

This closed loop can be viewed as a multi-round policy-iteration system with a fixed reference. In each round, the reward model is progressively refined to capture the new long-tail distribution generated by the preceding policy, while the policy model continues to approach domain-expert preferences in the petrochemical domain under an updated reward signal. Because the reference model $\pi_{\text{SFT-V0}}$ and KL coefficient β remain fixed throughout, results from all three rounds can be compared under a common constraint, reducing the risks of cumulative drift and reward hacking in multi-round RLHF.

2.4. Post-RLHF De-Identification

2.4.1. Rationale for the Processing Order

Placing de-identification after RLHF is a key design choice. The rationale is that placeholder substitution and numerical perturbation alter local text statistics, thereby affecting semantic similarity between neighboring textual units and the consistency of reward-model scores. If de-identification is performed first, the reward model learns preferences in a perturbed semantic space, causing the PPO policy to converge toward an incorrect objective.

To test this sequencing constraint, we establish pre- and post-de-identification control pipelines and compare their training utility and privacy performance under identical RLHF hyperparameters and de-identification strength.

2.4.2. Four-Stage Post-RLHF De-Identification Framework

We design a four-stage post-RLHF de-identification framework, as shown in Figure 2.

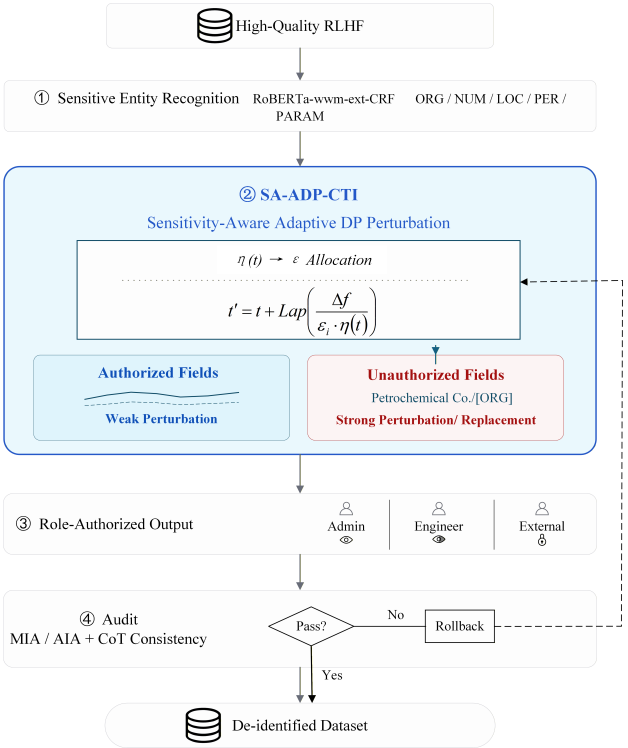


Figure 2. Four-stage post-reinforcement learning from human feedback (RLHF) de-identification framework.

Stage 1: Sensitive-entity recognition. Sensitive-entity recognition is the entry point to post-RLHF de-identification. Following the begin–inside–outside (BIO) tagging scheme, we define five sensitive-entity types: enterprise name (ORG), internal identifier (NUM), geographic location (LOC), person identity (PER), and process parameter (PARAM). RoBERTa-wwm-ext-CRF serves as the recognition module and comprises a Chinese RoBERTa-wwm-ext encoder for contextual representation and a CRF decoder that constrains entity-boundary transitions. Section 4.2 reports the dataset size, baseline models, and recognition performance.

Stage 2: Context-sensitive adaptive differential-privacy perturbation. For fields such as process parameters that are sensitive yet must retain interval characteristics, we develop SA-ADP-CTI (Sensitivity-Aware Adaptive Differential Privacy with Context-aware Token-level Injection). Its core idea is to adapt the perturbation magnitude of each entity token to contextual sensitivity rather than applying uniform global noise to all tokens. For each numerical token t to be perturbed, its value after ε-differentially private perturbation is defined as follows:

$$\tilde{t} = t + \mathrm{Lap}\left(\frac{\Delta f}{\varepsilon_i \cdot \eta(t)}\right) \tag{16}$$

Here, $\mathrm{Lap}(\cdot)$ denotes Laplace noise, Δf is the upper bound on local sensitivity, and $\eta(t) \in (0, 1]$ is the contextual sensitivity weight. In this work, $\eta(t)$ is given a closed-form calibration: each token t is associated with an entity class e recognized by the named entity recognition (NER) module and an access role c, whose calibrated weights are $s(e) \in [0.68, 0.92]$ (entity sensitivity, Table 2) and $\rho(c) \in \{0.18, 0.58, 1.00\}$ (role risk), respectively, and

$$\eta(t) = \frac{\varepsilon_i}{\varepsilon_i + \rho(c) \cdot s(e)} \tag{17}$$

Two limiting cases characterize Equation (17). First, for low-risk roles querying low-sensitivity operational entities (e.g., an internal operator reading process parameters), $\rho(c)s(e) \ll \varepsilon_i$, so $\eta(t)$ approaches 1 and the noise scale $\Delta f/(\varepsilon_i \eta(t))$ shrinks toward $\Delta f/\varepsilon_i$, yielding only weak perturbation that preserves interval characteristics; at the selected operating point ($\varepsilon = 2.0$), this corresponds to $\eta \approx 0.93$ for an operator and $\eta \approx 0.81$ for a process engineer querying a PARAM entity. Second, for role–entity combinations with no access authorization (e.g., an external researcher requesting enterprise names), the authorization layer enforces the $\eta \to 0$ limit rather than evaluating Equation (17): the noise scale diverges, and random perturbation degenerates into deterministic placeholder replacement. In the engineering implementation, the trigger probability of placeholder replacement for an authorized-visible entity equals $1 - \eta(t)$, so Equation (17) calibrates both the perturbation magnitude and the masking probability within a single consistent quantity. The calibrated sensitivity values $s(e)$ are listed in Table 2. The mechanism is compared with conventional de-identification, k-anonymity generalization, full-entity masking, partial-entity masking, and global DP perturbation. Section 4.2 reports the results.

The overall privacy guarantee follows from the sequential composition property of differential privacy: for one record containing n independently perturbed sensitive tokens, the composed mechanism satisfies $\varepsilon_{\text{total}} = \sum_{i=1}^{n} \varepsilon_i$. In the implementation, a per-record budget is allocated uniformly across the sensitive entities recognized in the record, $\varepsilon_i = \varepsilon/n$, so the reported operating point $\varepsilon = 2.0$ is consumed entirely within the record ($\varepsilon_{\text{total}} = 2.0$) and does not accumulate across records, since perturbations of different records are independent. The role of $\eta(t)$ in this accounting is to scale the noise magnitude at each token—the effective per-token budget is $\varepsilon_i \eta(t)$ in the denominator of Equation (16)—while the composition bound is governed by the ε_i allocated in the ledger; $\eta < 1$ can only add noise beyond the guaranteed level, never reduce the composed guarantee below $\sum_i \varepsilon_i$.

Table 2. Calibrated entity-sensitivity values $s(e)$ used in Equation (17).

Entity Class e	**ORG**	**NUM**	**LOC**	**PER**	**PARAM**
$s(e)$	0.68	0.72	0.88	0.92	0.82

Stage 3: Dynamic de-identification based on role authorization. Downstream users of petrochemical datasets have different access privileges, including operators, process engineers, and external researchers. A single de-identification level cannot meet the semantic needs of all roles. We therefore design a dynamic de-identification mechanism that adjusts the visibility of sensitive fields in the same sample according to the user's access role; Section 4.2 compares the different strategies.

Stage 4: De-identification quality auditing and rollback. To prevent over- or under-de-identification, the final stage simultaneously monitors the utility–privacy trade-off, membership-inference attack (MIA) and attribute-inference attack (AIA) risks, and semantic integrity through spot checks. Samples that fail the audit are rolled back to the previous stage for readjustment of perturbation parameters or role permissions.

3. Experimental Settings

3.1. Dataset

The experimental dataset was derived from four sets of equipment-related business documents from a petrochemical enterprise and contains 2600 raw records. Each record follows the multi-answer sample structure defined in Section 2.2, comprising one question Q, four candidate answers $A^{(1)} \sim A^{(4)}$, and independent scores from M_1 (Claude Sonnet 4.6) and M_2 (GPT-5.4 thinking). Questions are categorized as factual-definition, procedural, analytical-judgment, and explanatory-principle questions, corresponding to four typical petrochemical question-answering scenarios: what, how to proceed, how to judge, and why.

Generation of questions and candidate answers. Each document collection is split into sections (sections shorter than 100 characters are merged into their neighbors), yielding 316 chunks; M_1 then generates questions for each chunk in proportion to its length; and M_2 generates the four candidate answers of each question in a deliberately graded quality order, from a grounded, information-dense reference answer ($A^{(1)}$) to a directionally correct but vague one ($A^{(4)}$).

Following the cascaded filtering and quality-cleaning rules described in Section 2.2.3, 102 samples with inter-model score gaps above the corresponding consistency threshold and 49 samples with overall scores below 7.0 are removed, leaving 2449 samples in the valid data pool. For the 151 removed samples, only the necessary fields are retained, and they are added to the global PPO prompt pool as supplements. The valid data pool is partitioned into a fully isolated 500-sample global test set, a 150-sample local validation set, a 1000-sample R1 training pool, a 400-sample R2 holdout pool, and a 399-sample R3 holdout pool. The R2 and R3 pools remain completely isolated until their respective rounds begin. The global PPO prompt pool contains $1000 + 151 = 1151$ questions and is shared across all three rounds. All models are ultimately evaluated on the same 500-sample global test set to ensure a fair comparison.

Partition integrity was additionally verified on the final pools: the test set shares no question identifier with the union of all training-side pools (training, validation, holdout, and PPO prompt pools), and a character 3 g Jaccard scan over all cross-partition question pairs found a maximum similarity of 0.528, with no pair exceeding 0.6; no near-duplicate questions therefore leak across partitions.

3.2. Models and Baselines

Using Qwen1.5-1.8B-Chat as the base model, the experiment constructs five comparison baselines:

Base: the original Qwen1.5-1.8B-Chat without fine-tuning, which serves as the lower-bound baseline for the pipeline.

SFT: the model obtained by merging the LoRA adapters after three epochs of supervised fine-tuning on the R1 training pool; it serves as both the policy initialization and the KL-reference anchor for RLHF.

PPO-V1, PPO-V2, and PPO-V3: models produced by the three RLHF rounds, respectively; PPO-V3 is the final model.

All models are trained on a single NVIDIA GeForce RTX 5060 GPU with FP16 and gradient checkpointing enabled. SFT, RM, and PPO all use incremental LoRA fine-tuning (rank = 8, alpha = 16, dropout = 0.05) to improve GPU-memory efficiency and training stability.

All three stages use the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$): SFT and RM are trained with the Hugging Face Trainer, and PPO with the trl PPOTrainer. For learning-rate scheduling, SFT uses linear decay with a 3%-of-steps warmup (the Trainer default scheduler), RM-V1/V2/V3 use cosine decay with a 3% warmup, and the PPO learning rate is held constant at 1×10^{-6} across all three rounds (no decay). For batch sizes, SFT uses a per-device batch size of 1 with 4 gradient-accumulation steps, RM uses 2 with 16 accumulation steps (an effective batch size of 32), and PPO uses a batch size and mini-batch size of 1. Weight decay is 0.0 for SFT and 0.01 for RM and is disabled for PPO. The global random seed is fixed to 42 for SFT, RM, and PPO. The maximum gradient norm is clipped to 1.0 for RM and PPO.

3.3. Training Hyperparameters

SFT uses a learning rate of 2×10^{-5} for 3 epochs with max_length = 512. RM-V1 uses 2×10^{-5} for 3 epochs, whereas RM-V2 and RM-V3 use 1×10^{-5} for 2 epochs. PPO uses a learning rate of 1×10^{-6}, with 200 updates per round, max_new_tokens = 180, temperature = 0.7, top_p = 0.9, KL coefficient $\beta = 0.05$, target_kl = 0.1, $\gamma = 0.99$, $\lambda = 0.95$, and reward clipping to $[-5, +5]$.

The target modules for the PPO policy's LoRA adapters are expanded from $\{W_q, W_v\}$ in the RM to $\{W_q, W_k, W_v, W_o, W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}\}$, a total of seven projection matrices, providing greater policy capacity. Each PPO round performs 200 update steps, for 600 updates in total. The PPO prompt pool contains 1151 questions (see Section 3.1). Table 3 reports the sample sizes and loss configurations of the three RM training rounds.

Table 3. Reward model (RM) training data and loss configurations across three rounds.

Stage	α	Loss Weighting	Pref. Training Samples	Direct-Score Training Samples	Incremental-Training Init.
RM-V1	0.5	BT 0.5 + MSE 0.5	1000	3200	SFT-V0
RM-V2	0.2	BT 0.8 + MSE 0.2	1548	1589	RM-V1 LoRA
RM-V3	0.2	BT 0.8 + MSE 0.2	1598	1597	RM-V2 LoRA

The 1000 reported in Section 3.1 is the question-level pool size; the direct-score samples (3200) and preference pairs (1000) listed in this table are answer-/pair-level counts derived from that pool after quality filtering and pairing. R2 and R3 follow the same derivation from their respective holdout pools.

R1 additionally includes an 800-sample direct-score validation set and a 250-pair preference validation set. The R1 direct-score validation set is reused throughout R2 and R3 to monitor reward-scale drift.

3.4. Evaluation Metrics

We evaluate the models along three dimensions. The primary metric is the domain-preference reward: using the final RM-V3 as the common evaluator, each model-generated answer in the test set receives a scalar reward $r_\phi(Q, y)$, denoted by $\bar{r}_{\mathrm{V3}}$, which measures the alignment between the generated content and the latest preference distribution. The second metric is the pairwise win rate: for models M_a and M_b, their outcome on the same question Q is defined as follows:

$$\mathrm{win}(M_a, M_b \mid Q) = \begin{cases} +1, & r_\phi(Q, y_a) - r_\phi(Q, y_b) > \tau \\ 0, & \left| r_\phi(Q, y_a) - r_\phi(Q, y_b) \right| \le \tau \\ -1, & \text{otherwise} \end{cases} \tag{18}$$

With threshold $\tau = 1 \times 10^{-6}$, the win rate is the proportion of +1 outcomes among the 500 questions, characterizing the relative performance of two models more robustly than a single mean reward. The third dimension is text similarity: BLEU-4 and recall-oriented understudy for gisting evaluation (ROUGE-L) measure, respectively, n-gram precision and longest-common-subsequence coverage between the generated and reference answers and are treated as auxiliary metrics.

We also report average answer length as a measure of redundancy reduction. In RLHF, compressing answers can lower BLEU-4 even when semantic alignment improves because surface-form differences increase. We therefore treat RM-V3 reward and pairwise win rate as primary metrics and BLEU-4 and ROUGE-L as auxiliary metrics.

4. Experimental Results and Analysis

4.1. Overall RLHF Results

Figure 3 shows training over the three PPO rounds, and Table 4 reports the overall performance of the five models on the 500-sample global test set.

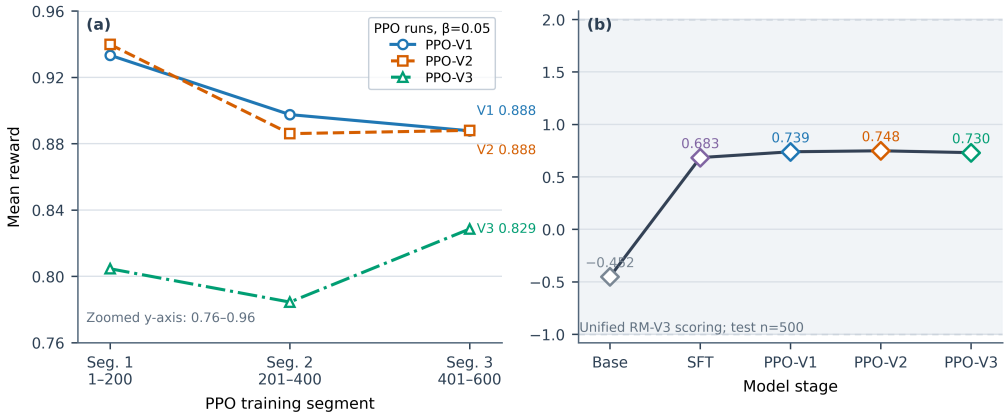


Figure 3. Mean-reward curves over three rounds of proximal policy optimization (PPO) training: (**a**) mean reward of each PPO run across the three training segments; (**b**) unified RM-V3 reward of the five model stages on the 500-sample global test set.

Table 4. Overall performance of the five models on the 500-sample global test set.

Model	BLEU-4	ROUGE-L	RM-V3 Reward	Average Length
Base	0.0446	0.1405	−0.4523	294.0
SFT	0.0753	0.1903	+0.6832	135.1
PPO-V1	0.0719	0.1919	+0.7386	105.7
PPO-V2	0.0623	0.1866	+0.7481	85.1
PPO-V3	0.0586	0.1841	+0.7302	79.4

Table 4 reports the final automatic metrics for each model, complementing the training-dynamics results in Figure 3.

To directly quantify the pairwise relationship between successive models—including the PPO-V2 versus PPO-V3 comparison—Table 5 reports the round-by-round win/tie/loss counts, win rates, mean per-question reward margins, and two-sided sign-test p-values on the 500-question test set.

Table 5. Round-by-round pairwise comparisons on the 500-question global test set (RM-V3 as the common evaluator; tie threshold $\tau = 1 \times 10^{-6}$; n = number of decisive questions).

Comparison	Win	Tie	Loss	Win Rate	Mean Margin	n	p (Sign Test)
SFT vs. Base	477	0	23	0.954	+1.135	500	$<10^{-40}$
PPO-V1 vs. SFT	257	14	229	0.514	+0.055	486	0.221
PPO-V2 vs. PPO-V1	228	52	220	0.456	+0.009	448	0.741
PPO-V3 vs. PPO-V2	172	133	195	0.344	−0.018	367	0.251
PPO-V3 vs. Base	481	0	19	0.962	+1.182	500	$<10^{-40}$

Three observations follow. First, the quality gain of the pipeline is concentrated in the SFT stage: SFT defeats Base on 95.4% of questions ($p < 10^{-40}$), whereas no successive PPO round produces a statistically significant pairwise improvement (all $p \geq 0.22$). Second, the PPO-V3 versus PPO-V2 comparison shows 172 wins, 133 ties, and 195 losses (34.4% win rate) with a mean per-question margin of only -0.018; the two-sided sign test over the 367 decisive questions does not reject parity ($p = 0.251$), and the tie share rises to 26.6%, the highest of any adjacent-round comparison. Statistically, therefore, the two models are indistinguishable in answer quality. Third, the mean-reward decline of PPO-V3 in Table 4 coexists with the largest redundancy reduction: average length falls from 135.1 (SFT) to 85.1 (PPO-V2) and 79.4 (PPO-V3), a 41.2% compression relative to SFT. Because shorter surface forms mechanically lower the reward on marginal questions, reward scale alone is insufficient for model comparison in this regime; the appropriate reading of the three-round loop is quality consolidation under the fixed SFT reference anchor, with progressive redundancy compression, rather than monotonic reward growth.

To complement these tests with explicit uncertainty statements, non-parametric bootstrap 95% confidence intervals (10,000 resamples over the 500 test questions, percentile method) were computed for the mean RM-V3 rewards, the pairwise win rates, and the mean per-question margins. For the mean rewards in Table 4, the intervals are $[-0.4993, -0.4049]$ for Base, $[+0.6383, +0.7291]$ for SFT, $[+0.7030, +0.7743]$ for PPO-V1, $[+0.7119, +0.7837]$ for PPO-V2, and $[+0.6942, +0.7662]$ for PPO-V3. Two conclusions follow. First, the headline gains are statistically robust: the SFT and Base intervals are entirely disjoint, the PPO-V3 versus Base win rate lies in $[0.944, 0.978]$, and its mean per-question margin lies in $[+1.1315, +1.2337]$. Second, the intervals of PPO-V1, PPO-V2, and PPO-V3 overlap almost completely, and the PPO-V3 versus PPO-V2 mean margin lies in $[-0.0423, +0.0068]$, which crosses zero—consistent with the sign tests in Table 5. These intervals quantify resam-

pling uncertainty at the evaluation level; the seed-level variance of the full training loop is discussed in Section 4.5.

Figure 3 and Table 4 reveal three aspects of progressive improvement. First, SFT is a crucial pre-requisite for effective RLHF: relative to Base, SFT substantially improves BLEU-4, ROUGE-L, and RM-V3 reward and achieves a 95.4% pairwise win rate, indicating that supervised fine-tuning first shifts the model from a general-dialogue distribution toward the petrochemical question-answering distribution. Second, three PPO rounds continue to improve domain alignment and reduce redundancy: PPO-V2 achieves the highest RM-V3 reward, while average answer length decreases from 135 in SFT to 79, indicating that reward optimization simultaneously improves domain alignment and information density. Third, the slight decline in PPO-V3 reward relative to PPO-V2 reflects diminishing marginal returns at later iteration stages, as the RM-V3 evaluator becomes more stringent and the model approaches the optimization limit supported by the current feedback pool.

Overall, relative to Base, PPO-V3 improves BLEU-4 by 31.4% and ROUGE-L by 31.0%, raises the RM-V3 reward by 1.183, and reduces answer length by 73.0%, while attaining a 96.2% pairwise win rate. These results demonstrate the effectiveness of the proposed dual-LLM feedback, joint-loss reward model, and iterative PPO closed loop for petrochemical question-answering.

4.2. Validation of the De-Identification Framework

To evaluate the "RLHF first, de-identification later" sequencing design and the four-stage post-RLHF de-identification framework described in Section 2.4, we conduct several experiments on a 1500-sample petrochemical de-identification dataset and a sensitive-sample processing pipeline.

4.2.1. Comparison of Sensitive-Entity Recognition Models

On hard_test, we compare a rule-based baseline, BERT-Softmax, BERT-CRF, BiLSTM-CRF, GlobalPointer, and combinations of different pre-trained models with CRF. The experiments show that RoBERTa-wwm-ext-CRF achieves an F1 score of 0.9779 on hard_test, an absolute improvement of 31.3 percentage points over the rule-based baseline and 7.3 percentage points over BiLSTM-CRF. These results indicate that whole-word-masking pre-training and CRF boundary constraints better identify compound Chinese sensitive entities and provide high-recall entity boundaries for subsequent perturbation and de-identification.

4.2.2. De-Identification-Sequence Comparison

We construct post-RLHF de-identification Pipeline A and pre-RLHF de-identification pipeline B and compare their training utility and privacy performance under identical RLHF hyperparameters and de-identification strength. The results are shown in Table 6.

Table 6. End-to-end comparison of post-RLHF de-identification (Pipeline A) and pre-RLHF de-identification (Pipeline B).

Pipeline	BLEU-4	ROUGE-L	Utility	Sensitive Residual Rate	Over-de-id Rate	RM-V3 Reward
Pipeline A (Post hoc)	0.9582	0.9720	0.9722	0.0916	0.0352	2.2905
Pipeline B (Pre-de-id)	0.8531	0.8898	0.9093	0.0000	0.6800	1.6067

The post-RLHF pipeline outperforms the pre-RLHF de-identification pipeline on all four utility metrics: Pipeline A achieves a pairwise win rate of 0.6780 over Pipeline B and a mean-reward advantage of $+0.6838$. Although Pipeline B reduces the sensitive-information residual rate to zero, its over-de-identification rate reaches 0.68. Before RLHF, it replaces

many domain terms unrelated to privacy with placeholders, causing the reward model to learn preferences in a perturbed semantic space and the policy to converge toward an incorrect objective. This result quantitatively validates the rationale in Section 2.4.1: de-identification perturbations must be applied after RLHF to preserve reward learning in the original semantic space.

4.2.3. Privacy–Utility Pareto Trade-Off

On the same 1500-sample dataset, we compare the risk–utility distributions of Global DP, Entity-level DP, Rule-based masking, and SA-ADP-CTI under different privacy budgets. The results are shown in Figure 4, which plots privacy risk R on the horizontal axis and composite utility U on the vertical axis, together with the non-dominated frontier, to identify candidate operating points under the objective of lower risk and higher utility.

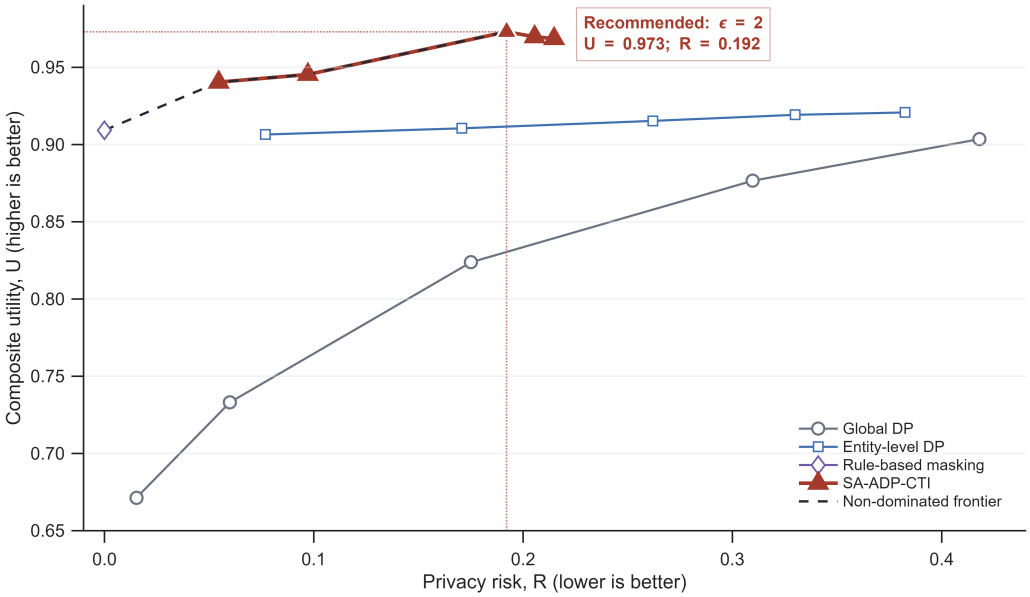


Figure 4. Sensitivity-Aware Adaptive Differential Privacy with Context-aware Token-level Injection (SA-ADP-CTI) privacy–utility trade-off curve.

As shown in Figure 4, rule-based masking has the lowest privacy risk, but its composite utility is fixed and cannot adequately satisfy authorized-retention requirements. Global DP improves utility as the privacy budget increases but simultaneously increases privacy risk, whereas Entity-level DP yields only limited utility gains. By contrast, SA-ADP-CTI reaches the non-dominated frontier in the low-to-moderate-risk region and achieves the recommended operating point $U = 0.973$ and $R = 0.192$ at $\varepsilon = 2$, showing that it can control the residual risk of sensitive values while preserving key semantic information.

4.2.4. Comparison of Dynamic Role-Based De-Identification Strategies

On the same sensitive-sample processing pipeline, we compare dynamic role-based de-identification, full-strength de-identification, partial de-identification, and no de-identification. Dynamic role-based de-identification achieves the best balance among sensitive-information residual rate, retention of authorized information, and semantic utility. Both the residual rate and the over-de-identification rate are zero, and the length-retention rate is 0.9664. Relative to full-strength de-identification, BLEU-4 improves by 38.0% (0.5994 $\rightarrow$ 0.8271). This result shows that role-authorized visibility control can provide finer-grained output-side protection without sacrificing domain semantics.

In addition, the sensitive-information reproduction probe shows that, without de-identification protection, PPO-V2 reproduces content from 620 sensitive contexts at a rate of

0.7943 (492 hits). After dynamic de-identification is introduced, the residual rate falls to zero and the success rate reaches 1.0. This shows that even after RLHF has completed quality alignment, output-side de-identification remains an indispensable security safeguard.

4.2.5. De-Identification Quality Audit

Privacy-budget trade-off testing shows that SA-ADP-CTI achieves the highest composite utility of 0.9730 at the selected operating point while maintaining a privacy-protection score of 0.8079. Membership-inference and attribute-inference attack scores fall to 0.7949 and 0.4823, respectively, substantially below the original-text values of 0.9999 and 0.8238. Across 1500 end-to-end experiments, both chain-of-thought (CoT) consistency and domain-semantic retention meet the audit thresholds, and no sample enters the rollback process (zero residual-sensitive items across all 1500 outputs; worst-case CoT consistency of 0.951 and semantic utility of 0.905 remain above the thresholds).

Table 7 reports the sensitivity of SA-ADP-CTI to the privacy budget ε over the grid $\varepsilon \in \{0.5, 1.0, 2.0, 4.0, 8.0\}$, together with the global-DP baseline under the same grid. The utility of SA-ADP-CTI peaks at $\varepsilon = 2.0$ (0.9730) and degrades gracefully toward both ends (0.9404 at $\varepsilon = 0.5$; 0.9686 at $\varepsilon = 8.0$), while privacy protection decreases monotonically from 0.9454 to 0.7852 as ε grows; the selected operating point $\varepsilon = 2.0$ therefore sits at the knee of the utility–privacy trade-off rather than at an isolated optimum. The global-DP baseline, in contrast, forfeits utility at small ε (0.6713 at 0.5) because uniform noise damages non-sensitive content, confirming that the adaptive allocation—rather than the budget value itself—drives the advantage.

Table 7. Sensitivity of SA-ADP-CTI and the global differential privacy (DP) baseline to the privacy budget ε (semantic utility and privacy protection; 1500-sample de-identification corpus).

	SA-ADP-CTI		Global DP	
ε	**Utility**	**Privacy**	**Utility**	**Privacy**
0.5	0.9404	0.9454	0.6713	0.9847
1.0	0.9454	0.9028	0.7331	0.9401
2.0	0.9730	0.8079	0.8238	0.8249
4.0	0.9698	0.7944	0.8766	0.6903
8.0	0.9686	0.7852	0.9035	0.5820

4.3. Manual Evaluation and Ablation Validation

To address the evaluation circularity of "dual-LLM feedback → RM training → RM-V3 self-evaluation," we conduct two small-scale manual evaluations. Petrochemical-domain experts serve as independent external evaluators to independently assess the credibility of the feedback sources and the engineering compliance of the de-identification framework.

4.3.1. Agreement Between Dual-LLM Feedback and Human Experts

We draw a stratified sample of 100 questions from the 500-sample global test set. Three engineers with experience in petrochemical equipment operation and maintenance or process engineering independently score the Base, SFT, PPO-V1, PPO-V2, and PPO-V3 answers for correctness, completeness, and usefulness.

Inter-expert agreement (Krippendorff's $\alpha = 0.74$) reaches a "good" level, indicating a well-designed scoring protocol and stable expert judgments. The Pearson correlation between mean expert scores and dual-LLM scores is 0.78, and Spearman's rank correlation with RM-V3 scores is 0.62. The five models have identical expert- and RM-V3-based rankings. These results quantitatively support the feasibility of replacing expensive expert annotation with dual-LLM feedback, as proposed in Section 2.2.

4.3.2. Manual Review of De-Identified Samples

A stratified sample of 150 outputs from the four-stage de-identification framework is reviewed by two petrochemical engineers and one data-compliance specialist for semantic integrity and privacy adequacy.

Manual review yields a mean semantic integrity score of 4.35/5, an intraclass correlation coefficient (ICC) of 0.76, and a 92.8% pass rate at the threshold of $\geq$4. The majority-vote pass rate for privacy adequacy is 92.1%, which corresponds to rollback rates of 7.2% (semantic integrity) and 7.9% (privacy adequacy); all failed samples are returned to Stages 2–3 for perturbation-parameter or permission readjustment and are re-audited before release.

4.4. Ablation Study

To evaluate the joint-loss reward-model design proposed in Section 2.3.1, we ablate the reward-model loss function in R1. With the preference-training set (1000 pairs), direct-score training set (3200 samples), preference-validation set (250 pairs), random seed (seed = 42), and number of training steps held constant, three configurations are compared: BT pairwise loss only ($\alpha = 0$), MSE direct-score loss only ($\alpha = 1$), and the proposed joint loss ($\alpha = 0.5$). The resulting RMs are then used to train PPO-V1 under exactly the same PPO hyperparameters, and preference-ranking accuracy and the PPO-V1 mean reward are compared. The results are shown in Table 8.

Table 8. Ablation of the reward-model loss function in R1.

Loss Configuration	Preference-Ranking Accuracy	PPO-V1 Mean Reward
BT only ($\alpha = 0$)	0.812	0.882
MSE only ($\alpha = 1$)	0.764	0.864
Joint loss ($\alpha = 0.5$)	0.831	0.933

The joint loss performed best on both metrics. Relative to BT-only training, it increased preference-ranking accuracy by 1.9 percentage points and the PPO-V1 mean reward by 0.051. Relative to MSE-only training, the corresponding improvements were 6.7 percentage points and 0.069.

BT and MSE played complementary roles in the joint objective. The BT-only configuration (Acc = 0.812) substantially outperformed the MSE-only configuration (Acc = 0.764) in preference ranking, indicating that the BT loss is better suited to learning relative preferences. However, because BT alone provides no absolute-score reference, its reward scale is more susceptible to drift, and its PPO-V1 mean reward remained below that of the joint-loss configuration. BT therefore supplies fine-grained relative ranking, whereas MSE anchors the absolute reward scale; combining them improves both ranking accuracy and downstream policy reward.

4.5. Limitations of the Proposed Framework

Although the above experimental results confirm the effectiveness of the proposed framework on real petrochemical question-answering records, several limitations remain that warrant further improvement. First, the scale and source coverage of the evaluation data are limited: the current dataset comprises only 2600 raw records originating from a single petrochemical enterprise, and the post hoc de-identification pipeline is validated on merely 1500 samples. Although the dual-LLM feedback and the three-round closed loop already deliver quantifiable gains on this corpus, the generalization of the framework to multi-enterprise, cross-regional, and even cross-industrial corpora (e.g., chemicals, energy, and metallurgy) has not yet been empirically verified; differences in terminology, safety conventions, and data-privacy regulations across enterprises may simultaneously affect

the transferability of the reward model and the recall of sensitive-entity recognition. It should be emphasized that the above limitation is primarily confined to the dimension of data coverage and can be readily addressed within the existing closed-loop architecture by extending data sources, expanding data scale, and incorporating cross-enterprise and cross-industrial evaluation corpora. In this sense, transfer testing across enterprises would carry diagnostic value for the framework itself: verifying whether the reward model remains aligned under different terminology and safety conventions would delimit the boundary of data-centric generalization, while measuring the recall of sensitive-entity recognition under different privacy regulations would quantify the robustness of the de-identification pipeline, both determine whether the closed loop can be reused as-is, recalibrated, or retrained when ported to a new enterprise. Furthermore, all training stages use a single random seed (42); the seed-level variance of the full SFT–RM–PPO loop has therefore not been characterized. Evaluation-side uncertainty is quantified instead through the bootstrap confidence intervals and pairwise sign tests reported in Section 4.1, and multi-seed repetition of the full pipeline is left as future work.

5. Conclusions

To address high expert-annotation costs, domain-preference misalignment, and the conflict between sensitive-data security and dataset utility in process industries, this paper proposed a privacy-preserving iterative RLHF framework that organizes human-preference alignment and post hoc data de-identification into a single sequential closed loop. The framework replaces large-scale expert annotation with independent scoring and pairwise preferences from two high-capability LLMs, providing a low-cost and auditable feedback source for reward learning. A joint-loss reward model, iterated over three rounds under a fixed SFT reference anchor, aligns model outputs with domain-expert preferences while preserving cross-round comparability and progressively compressing redundancy. The SA-ADP-CTI post-RLHF pipeline then applies token-level adaptive perturbation and role-authorized masking to sensitive entities, preserving dataset utility under privacy-compliance requirements. The "alignment-first, de-identification-later" sequencing shields the reward-learning objective from early perturbation noise, and the experimental results on petrochemical question-answering records support the effectiveness of the framework as a whole. The evaluation of this study is confined to data from a single enterprise and a single backbone scale, and generalization to cross-enterprise and cross-industrial settings has not yet been empirically verified (Section 4.5). Future work will explore scaling to larger foundation architectures, integrating offline alignment paradigms (e.g., DPO), and extending the framework to broader continuous-process domains such as energy and metallurgy.

Author Contributions: Conceptualization, Y.L.; methodology, Y.L.; software, Y.L.; validation, Q.J. and Y.L.; formal analysis, Q.J. and Y.L.; investigation, Q.J. and Y.L.; data curation, Y.L.; writing—original draft, Y.L.; writing—review and editing, S.L. and J.C.; visualization, Q.J. and Y.L.; supervision, S.L. and J.C.; project administration, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to commercial and privacy restrictions.The raw business documents of the petrochemical enterprise are not publicly available due to commercial and privacy restrictions.

Acknowledgments: During the preparation of this manuscript, the authors used GLM-5.2 for the purposes of language polishing and sentence structure optimization. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RLHF	Reinforcement Learning from Human Feedback
PPO	Proximal Policy Optimization
SFT	Supervised Fine-Tuning
RM	Reward Model
BT	Bradley–Terry
MSE	Mean Squared Error
LoRA	Low-Rank Adaptation
LLM	Large Language Model
NER	Named Entity Recognition
CRF	Conditional Random Field
RoBERTa	Robustly Optimized BERT Pre-Training Approach
DP	Differential Privacy
KL	Kullback–Leibler
BLEU	Bilingual Evaluation Understudy
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SA-ADP-CTI	Sensitivity-Aware Adaptive Differential Privacy with Context-aware Token-level Injection

References

1. Yang, C. Research and exploration on preventive maintenance strategies for rotating equipment in refining and chemical enterprises based on AI. *Pet. Process. Petrochem.* **2026**, *57*, 81–87.
2. Luo, M.J.; Zhao, Y.P.; Zhang, M.X.; Shi, X.G.; Lan, X.Y. High-quality datasets for petrochemical industry in era of large language models: Challenges and opportunities. *CIESC J.* **2026**, *77*, 435–447.
3. Singh, S.K. ChemSafeAI+: A machine learning driven dynamic safety and optimization framework for chemical process industries. *Preprints* **2026**, 2026010507. [CrossRef]
4. Wu, Q.; Yao, D.; Liu, Y.; Xu, T.; Ying, Z. Research on construction methods of high-quality multimodal datasets in industrial manufacturing. In *Proceedings of the 2025 IEEE International Conference on Blockchain, Zhengzhou, China, 30 October–2 November 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 241–246.
5. Liu, Z.; Liu, S.; Min, J.; Zhang, J.; Cen, J.; Han, P.; Hu, S.; Meng, Z.; He, X.; Zhou, D. InspecSafe-V1: A multimodal benchmark for safety assessment in industrial inspection scenarios. *arXiv* **2026**, arXiv:2601.21173. [CrossRef]
6. Yu, J.; Sun, S.; Hu, X.; Yan, J.; Yu, K.; Li, X. Improve LLM-as-a-judge ability as a general ability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025), Suzhou, China, 4–9 November 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 14099–14115.
7. Tan, Z.; Li, D.; Wang, S.; Beigi, A.; Jiang, B.; Bhattacharjee, A.; Karami, M.; Yu, Z.; Li, J.; Liu, H. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), Miami, FL, USA, 12–16 November 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 930–957.
8. Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. LIMA: Less is more for alignment. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 55006–55021. [CrossRef]
9. Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4299–4309.
10. Zheng, R.; Dou, S.; Gao, S.; Hua, Y.; Shen, W.; Wang, B.; Liu, Y.; Jin, S.; Liu, Q.; Zhou, Y.; et al. Secrets of RLHF in large language models part I: PPO. *arXiv* **2023**, arXiv:2307.04964. [CrossRef]
11. Chaudhari, S.; Aggarwal, P.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; Narasimhan, K.; Deshpande, A.; Castro da Silva, B. RLHF deciphered: A critical analysis of reinforcement learning from human feedback for LLMs. *ACM Comput. Surv.* **2025**, *58*, 53:1–53:37. [CrossRef]
12. Jiang, H.Y.; Zeng, F.J.; Qu, B.; Lin, X.J.; Zhong, W. Helios: A foundational language model for smart energy knowledge reasoning and application. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*; Demberg, V., Inui, K., Marquez, L., Eds.; ACL: Rabat, Morocco, 2026; pp. 3208–3220.

13. Liu, Q.; Song, Z.; Liang, Y.; Xie, Z.; Zhang, S.; Zhang, J.; Li, Y. CoRLHF: Reinforcement learning from human feedback with cooperative policy-reward optimization for LLMs. *Expert Syst. Appl.* **2026**, *301*, 130113. [CrossRef]
14. Zhai, Y.; Lei, Y.; Zhang, H.; Yu, Y.; Xu, K.; Feng, D.; Ding, B.; Wang, H. Uncertainty-penalized reinforcement learning from human feedback with diversified reward LoRA ensembles. *Inf. Process. Manag.* **2026**, *63*, 104548. [CrossRef]
15. Zhang, Y.F.; Yu, T.; Tian, H.; Fu, C.; Li, P.; Zeng, J.; Xie, W.; Shi, Y.; Zhang, H.; Wu, J.; et al. MM-RLHF: The next step forward in multimodal LLM alignment. *arXiv* **2025**, arXiv:2502.10391.
16. Yu, D.; Kairouz, P.; Oh, S.; Xu, Z. Privacy-preserving instructions for aligning large language models. *arXiv* **2024**, arXiv:2402.13659. [CrossRef]
17. Fan, F.X.; Tan, C.; Ong, Y.S.; Wattenhofer, R.; Ooi, W.-T. FedRLHF: A convergence-guaranteed federated framework for privacy-preserving and personalized RLHF. *arXiv* **2024**, arXiv:2412.15538. [CrossRef]
18. Belakaria, S.; Kazdan, J.; Marx, C.; Cundy, C.; Neiswanger, W.; Koyejo, S.; Engelhardt, B.E.; Ermon, S. Sharpe ratio-guided active learning for preference optimization in RLHF. *arXiv* **2025**, arXiv:2503.22137.
19. Wang, B.; Zheng, R.; Chen, L.; Liu, Y.; Dou, S.; Huang, C.; Shen, W.; Jin, S.; Zhou, E.; Shi, C.; et al. Secrets of RLHF in large language models part II: Reward modeling. *arXiv* **2024**, arXiv:2401.06080.
20. Guo, R.; Zhang, H.H. Chinese medical named entity recognition based on RoBERTa and adversarial training. *J. East China Univ. Sci. Technol.* **2023**, *49*, 144–152.
21. Saeidian, S.; Yavuzyılmaz, A.; Grosse, L.; Schuppe, G.F.; Gouverneur, A.; Oechtering, T.J.; Skoglund, M. A tight context-aware privacy bound for histogram publication. *IEEE Signal Process. Lett.* **2025**, *32*, 4169–4173. [CrossRef]
22. Manzanares-Salor, B.; Sánchez, D. Enhancing text anonymization via re-identification risk-based explainability. *Knowl. Based Syst.* **2025**, *310*, 112945. [CrossRef]
23. Zaheer, M.; Reddi, S.J.; Sachan, D.; Kale, S.; Kumar, S. Adaptive methods for nonconvex optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; pp. 9815–9825.
24. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.