# ARGUS: Feedback-Reinforced Gradual LLM-Based Framework for Interpretable and Robust Archive Review

Anonymous ACL submission

## Abstract

Automated archive review faces challenges in interpreting domain-specific semantics and ensuring traceable decisions, because existing methods relying on rigid rules or generic language models lack complex context understanding and review transparency. Regarding these issues, we propose ARGUS, a feedback-reinforced gradual framework for archive review. ARGUS uses hierarchical rule-embedded prompts for stepwise inference, feedback-driven sample enhancement via LLM inference logs for robustness, and parameter-efficient fine-tuning via low-rank adaptation. Evaluations on real-world archives and benchmarks show ARGUS achieves 10.5–15.5% higher accuracy than baselines, reduces ASR by 25%, and has been proven to effectively complete review tasks under limited resources.

## 1 Introduction

Archive review is one of the critical tasks in the field of natural language processing(NLP). Recent studies (Vaswani et al., 2017; Wei et al., 2022a; CONNEAU and Lample, 2019) have demonstrated that in the field of natural language processing, large language models (LLMs) have more advantages compared to deep learning (DL). After pre-training on large-scale data, the general features learned by LLMs can be directly transferred to downstream tasks (Chiang and yi Lee, 2022). Even without fine-tuning for specific tasks, they can complete various natural language processing tasks (Brown et al., 2020). Moreover, the inherent explainability (Liang et al., 2023) of LLMs, generating natural-language explanations to enhance inference transparency,effectively addresses the interpretability issue of existing methods. More importantly, the global attention mechanism of the Transformer architecture adopted in LLMs makes it easier to understand complex contexts (Vaswani et al., 2017; Jin et al., 2025), such as the implicit semantic features commonly found in archival texts.

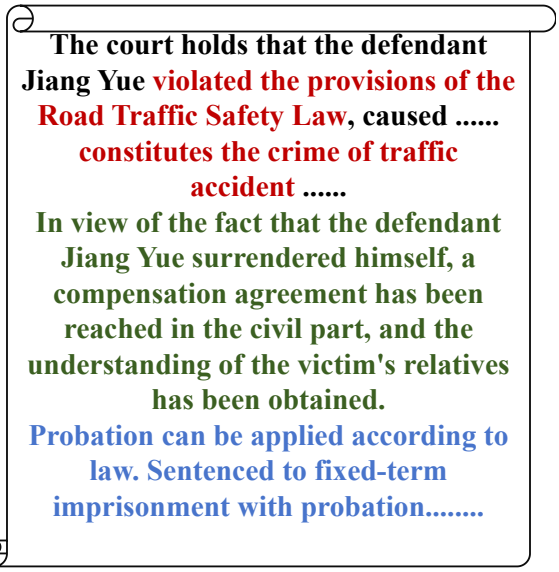


Figure 1: Example of archives. Archival data exhibits high domain specificity (red), and due to the complex contexts often involved in archival documents, they are often obscure (green) and ambiguous (blue).

However, it is hard to apply LLMs for archival review tasks due to the following challenges: 1) **Fine-grained domain knowledge deficiency of LLMs in complex semantic scenarios**: Archival data, as shown in Figure 1, is highly domain-specific and contains complex semantic contexts, but LLMs lack fine-grained domain knowledge (Yang et al., 2023), which makes it difficult to accurately identify sensitive content and compliance boundaries; 2) **General LLMs are prone to perturbation**: general-purpose LLMs have weak anti-interference capabilities and are vulnerable to text adversarial perturbations such as semantic paraphrasing (Xu et al., 2024; Zhang et al., 2020; Peng et al., 2025). s

To address these issues, we propose a Large Language Model-based Feedback-Reinforced Gradual

Framework called **ARGUS**, for Interpretable and Robust Archive Review. ARGUS constructs structured inference prompt templates based on domain features, makes up for the lack of fine-grained domain knowledge of LLMs through rule embedding, and improves the accuracy of identifying sensitive content and compliance boundaries. It also generates training samples by combining model inference log analysis with diffusion models, further enhances model performance through lightweight iterations, and ensures robustness at the same time. We conduct experiments on a real archival dataset and 2 public sensitive datasets. Compared with Pretrained long-sequence language models of BigBird and Longformer, ARGUS achieves 10.5–15.5% higher accuracy than baselines, reduces ASR by 25%. Our main contributions are as follows:

- The first domain-enhanced large language model framework for the task of archival opening review. Under the lightweight fine-tuning strategy of Quantized Low-Rank Adaptation, while taking into account both the review accuracy and robustness, the demand for computing resources is substantially diminished.
- Proposed a "rule embedding-adversarial enhancement" LLM-based automated archival review idea, which combines Hierarchical Rule-Embedded Prompting and Semantic-Preserving Adversarial Generation driven by diffusion models, to address the lack of fine-grained domain knowledge.
- Experiments conducted on real archival datasets show that ARGUS outperforms the baseline model in both accuracy and recall. In adversarial testing, compared with the pretrained model, ARGUS can better suppress misjudgments under three types of perturbation tests.

## 2 Related Work

**Rule-based methods.** In the early practices of archival opening review, rule-based review systems (like keyword matching and regular expressions) relying on static rule libraries struggle with semantic ambiguity and context dependence. This leads to a higher misjudgment rate in complex scenarios. The conflict between their rigid matching mechanisms and the dynamic evolution of language also limits the systems' ability to generalize new variants of sensitive words (Hedda et al., 2017). Hybrid frameworks integrating domain knowledge bases (such as expert systems and knowledge graphs) have steered the review approach towards intelligence. However, constructing logical rules incurs high human costs, and it's difficult to adapt to large-scale heterogeneous data (Zhong et al., 2024). Static knowledge bases can't evolve dynamically, and there's a notable lag in system reconstruction when facing policy updates or new terms.

**DL-based methods.** Supervised machine learning endows computers with context-understanding ability, enabling them to identify text content for review at a deeper level and improving review accuracy (Hutchinson, 2018). For instance, the natural language processing-based clustering model (CPPIM) for automatically detecting personal identity information in unstructured text corpora and the Byte-mLSTM have also achieved excellent results (Kulkarni and Cauveryn, 2021). The Chinese word segmentation model Attention-BiLSTM-CRF, developed by integrating the attention mechanism, LSTM, and Conditional Random Field (CRF), has also been proven effective on SCD text data (Zheng et al., 2023). The combination of XLNet and BiLSTM-CRF in named entity recognition (NER) tasks demonstrates XLNet's superiority in capturing context information and has achieved leading results in NER (Yan et al., 2021).

**LLM-based methods.** Recent related studies have started to explore the application of large language models (LLMs) in detecting and protecting personally identifiable information in archival data, such as detecting and safeguarding personal identity information within archival data. (Yang et al., 2023). In the field of natural language processing, the current focus is on the Transformer architecture (such as BERT(Devlin et al., 2019)). Starting from dialogue understanding tasks, continued pre-training of models on domain-specific datasets is adopted to enhance model performance. For example, continued pre-training is carried out on domain datasets using the Masked Language Model (MLM), Span Boundary Objectives (SBO), and Perturbed Mask Objectives (PMO) (Wu et al., 2021). Pre-trained models learn language knowledge through self-supervised tasks (such as the Masked Language Model and Next Sentence Prediction), and improve dialogue understanding and multi-task processing capabilities through domain adaptive pre-training, but there are still limitations in data dependence and capturing diachronic features (Han et al., 2021). Significant achievements

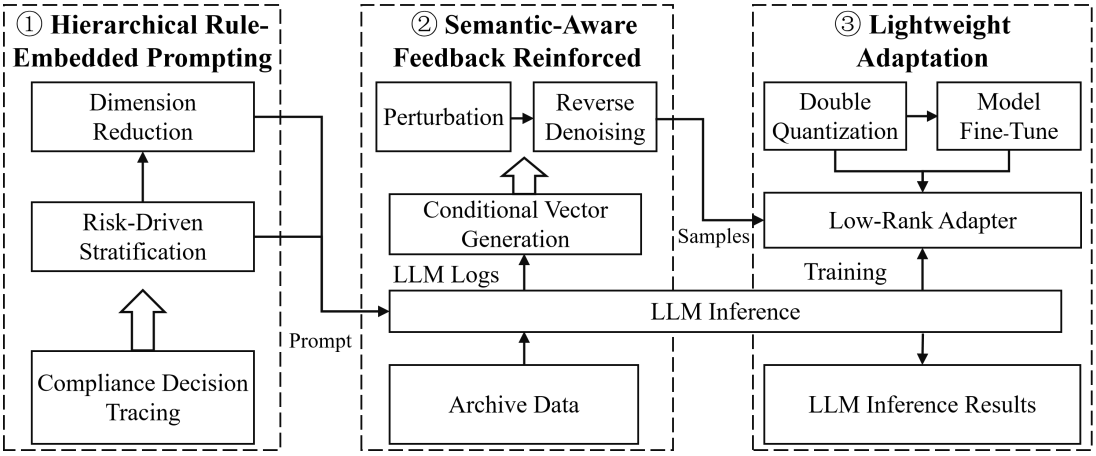


Figure 2: The overview of ARGUS.

have been made in multiple NLP tasks and multimodal tasks, surpassing traditional models. For example, in scenarios such as medical text classification (Singhal et al., 2022) and legal clause parsing (Cui et al., 2023).

## 3 Design

To address the issues of general LLMs lacking fine-grained domain knowledge in complex semantic contexts and being prone to perturbations, ARGUS constructs Hierarchical Rule-Embedded Prompting based on domain features, making up for the lack of fine-grained domain knowledge of LLMs through rule embedding. ARGUS completes Semantic-Aware Feedback Reinforced by combining diffusion models with the analysis of model inference logs. Finally, ARGUS adopts a Lightweight Adaptation strategy to ensure lightweight iteration. The complete overview of ARGUS is shown in the Figure 2.

### 3.1 Hierarchical Rule-Embedded Prompting

Inspired by the Least To Most (Zhou et al., 2023) (Wei et al., 2022b) ,which decomposes complex tasks and makes layer-wise judgments based on conditions, ARGUS proposes Hierarchical Rule-Embedded Prompting, as shown in Figure 3. It embeds prior domain knowledge into task inference and designs hierarchical checking steps. By deeply embedding domain rules into the predefined steps of the prompt and taking into account risk priorities, the model's accuracy in identifying sensitive information in complex texts is effectively improved.

**Rule Dimension Reduction.** By parsing legal texts, high-dimensional and complex rules (such as abstract clauses) are decomposed into atomized Boolean logical conditions (e.g."Incomplete name and containing special characters → Not sensitive"), and redundant rules are eliminated through the Risk Exposure Coefficient to form a streamlined and executable rule system.

**SYSTEM:**
You are a senior examiner responsible for the review work of archives , and the following are the steps of y our review :
[Step 1] Identify personal identification markers ..... directly output "Involves" .
[Step 2] Detect structured address elements .....direct ly output "Involves" .
[Step 3] Detect implicit identity cues, .....mentions.
[Step 4] Directly output not involve .

**USER:**
Here are the contents you need to review:{Input_con tent},The answer must start with either "Involves" or "Does not involve", and then explain the reason.
[Important] Please be sure to complete the output of the judgment conclusion strictly according to the abo ve steps! Omitting any unredacted information may l ead to serious legal consequences!

Figure 3: Prompt used in our methods

**Risk-Driven Stratification.** Based on the reduced dimensionality rule system and risk hierarchy, hierarchical verification is constructed to form a decision path for the selection of risk samples from highest to lowest priority.

- Primary Identifier Integrity Layer (PII-L): Validates the integrity of direct unique identifiers, such as identifying whether personal identity information is complete.

"**sample id**":"SAMP-0001",
"**original text**": "Case Name: Chang'an Research Institute and Li * Economic Dispute Execution Ruling Party: Research Institute, Li * Full Text: .....",
"**errors**":[
{"**type**": "Missing Confusion processing","**position**": [12,22]}
],
"**layer**":{"**focus layer**":"L3","**attack types**":["add_special_chars", "nested_address"],"**priority**": 0.15}


Figure 4: The sample logs of LLMs

- Structured Location Compliance Layer (SLC-L): Verifies the compliance of location identifiers. For example, location identifiers in nonsensitive data must meet the minimum necessary disclosure requirement (location information should not contain obvious numerical features).

- Semantic Obfuscation Compliance Layer (SOC-L): Detects the semantic obfuscation compliance of indirectly linkable identifiers. For example, detect implicit sensitive information clues(such as identity cues), and the key information in sensitive data should be obfuscated (surnames should be retained, while given names should be replaced with special symbols such as '*').

**Compliance Decision Tracing.** The hierarchical review path interacts with the attention mechanism and probabilistic generation strategy of LLMs, ensuring the traceability of review results. Meanwhile, explicit conclusions and inference bases are enforced via declarations of the severity of legal consequences to guide model generation.

### 3.2 Semantic-Aware Feedback Reinforced

Based on the inference logs feedback from large language models, ARGUS constructs feedback-reinforced samples and feeds them back to the base model for fine-tuning. An example of the LLMs inference log is shown in Figure 4. Specifically, ARGUS builds conditional vectors from the logs generated by large LLMs inference and incorporates them into the diffusion generation process. These conditional vectors guide and constrain the noise injection range to avoid semantic distortion while directing model attention to generate feedback-reinforced structured samples $D_{adv}$.

**Conditional Vector Generation.** In the LLMs reasoning logs, apart from the "original text" where model inference errors occur, there are mainly two components: "**errors**" record the error type "type" and error character interval "position" in the LLMs' inference of the "original text", from which an error vector $\mathbf{e} \in \{0,1\}^3$ is constructed; "**layer**" records the targeted generation layer "focus layer", predefined adversarial strategy "attack type", and generation weights for the "original text", from which a hierarchical weight vector $\omega = [\alpha_{PII}, \alpha_{SLC}, \alpha_{SOC}]$ is constructed. Finally, the error vector $\mathbf{e}$ and hierarchical weight vector $\omega$ are concatenated to generate the conditional vector .

$$\mathbf{c} = \mathbf{concat}(\omega, \mathbf{e}) \in \mathbf{R}^6$$

here, $0 < \alpha_i < 0.2$.

**Hierarchical Perturbation.** Based on the conditional vector $\mathbf{c}$ to diagnose the model's cognitive deficiencies, ARGUS uses a hierarchical-structure-sensitive mask matrix to restrict the range of noise injection, thereby directionally perturbing latent semantics:

$$M_c \in \{0,1\}^d$$

and outputs the latent representation $\mathbf{z}_t$ with limited noise:

$$\mathbf{z}_t = \sqrt{\overline{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon} \odot \mathbf{M}_c \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$$

here $\overline{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$, $\beta_i$ is the noise scheduling parameter. By restricting noise to act only on the embedding space of the target hierarchical structure, semantic-aware perturbations are achieved.

**Condition-Guided Reverse Denoising.** Leveraging the Transformer architecture, ARGUS explicitly integrates the structured conditional vector via cross-modal attention. It dynamically adjusts noise exclusively in targeted regions at each step, iteratively generating high-fidelity, diverse augmented samples in an end-to-end manner without external constraints.The cross-modal attention computation is as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{d_k}}\right)\mathbf{V}$$

$d_k$ denotes the feature length of the key vector.

The key innovation of ARGUS lies in explicitly injecting the hierarchical conditional vector into the Transformer's attention by modifying the Key (K) matrix for condition-aware generation.

$$\mathbf{K} = \mathrm{Concat}(\mathbf{z}_t, \mathbf{c})\mathbf{W}_K$$

Here, $\mathbf{W}_K$ is a learnable projection matrix, ensuring the model prioritizes risk-level-related semantic

features during denoising. ARGUS directly associates the conditional vector with generated content via cross-modal attention and performs corresponding dynamic weight adjustments.

$$Logs \xrightarrow{Reinforced} D_{adv}$$

ARGUS leverages the inference logs feedback during the large language model's inference process to gradually generate feedback-reinforced samples through diffusion under the explicit guidance of the conditional vector . The reinforced samples $D_{adv}$, after structured processing, are used as feedback-reinforced structured training data and input back to the large language model for targeted fine-tuning.

### 3.3 Low-Rank Lightweight Adaptation

To balance the effectiveness of fine-tuning with hardware resource constraints under limited conditions, ARGUS employs the Quantized Low-Rank Adaptation strategy to lightweight fine-tune general LLMs based on feedback-reinforced structured fine-tuning samples $D_{adv}$. Lightweight Adaptation not only effectively improves model performance but also ensures the controllability and Interpretability of its output content.

**Rigorous Generation Constraints.** ARGUS employs a unique LLM fine-tuning mechanism with "instruction-input-output" structured samples and feedback reinforcement to strongly constrain generated content (such as suppress hallucinations). The structured fine-tuning samples after feedback reinforcement are shown in the Figure 5.


{
"**instruction**": "You are a senior reviewer responsible for the review work of archives opening, and the following are the steps of your daily review work...",
"**input**": "Case Name: Civil Ruling on Corporate Merger Dispute and Sales Contract Dispute; Party: Shen*...",
"**output**": "**Involved**. The name 'Shen*' contains the special character '*', failing the condition of Step 1. The address '...No. 110' includes detailed residential information without desensitization, satisfying the condition of Step 2."
}


Figure 5: Example of feedback-reinforced structured fine-tuning samples

The feedback-reinforced structured samples align with generation logic in data format, ensuring the fine-tuned model follows "Hierarchical Rule-Embedded" logic for results, facilitating debugging and iteration. Modifying instruction templates enables rapid adaptation to archive audit domains, offering stronger content control than traditional unstructured-data-dependent fine-tuning.

**Double Quantization.** First, perform NF4 quantization on all model parameters and inject trainable low-rank matrices to achieve efficient fine-tuning.For a linear transformation $\mathbf{W}_0 x$,

$$\mathbf{W}_{\text{QLoRA}} x = \underbrace{(\mathbf{W}_0^{\text{4bit}})_{\text{NF4}} x}_{\text{Parameter freeze}} + \alpha \cdot \underbrace{\mathbf{B}\mathbf{A}x}_{\text{Low-rank adapter}}$$

Here, $(\mathbf{W}_0^{\text{4bit}})_{\text{NF4}}$ are the frozen 4-bit quantized base model weights. $\mathbf{A}, \mathbf{B}$ are trainable low-rank matrices with rank $r \ll d$ . The scaling coefficient $\alpha$ controls the contribution strength of the adapter. Then, the quantized weights are encoded via a codebook, and this codebook is quantized again with fewer bits (such as NF4 or FP4). This dual-step compression minimizes resource demands while retaining base model knowledge.

**Hierarchical Adversarial Training.**Inject the adversarial samples $D_{adv}$ generated in Section 3.2 into the training process and introduce a hierarchy-sensitive loss function.

$$\mathcal{L}_{\text{QLoRA}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{adv}}} \left[ \sum_{i=1}^{3} w_i \cdot \mathcal{L}_{\text{CE}}(f_\theta(x^{(i)}), y^{(i)}) \right]$$

Here,$w_i = Softmax\left([w_{PII}, w_{SOC}, w_{SLC}]\right)$, $x^{(i)}$ represent adversarial samples in layer $i$.

## 4 Experiments

This study aims to address the following research questions.

**RQ1**: Does ARGUS exhibit superior performance compared to other baseline methods in archival review?

**RQ2**: Does the "Hierarchical Rule-Embedded Prompting" adopted in ARGUS lead to more accurate archival review performance?

**RQ3**: Does the "Semantic-Aware Adversarial Enhancement" in ARGUS improve model robustness while ensuring performance?

**RQ4**: Can our method reduce costs while maintaining performance to adapt to practical archival review scenarios with limited resources?

### 4.1 Experimental Settings

**Dataset.** We evaluate the performance of our method on 3 datasets: 1) Archives: an archival dataset constructed from unpublished records provided by the Hunan Provincial Archives; 2) Crimes (Zhang et al., 2025): built from publicly available

Chinese court rulings; and 3) ai4privacy: an open-source privacy dataset PII-Masking (ai4Privacy, 2023) designed for training models to remove personally identifiable information (PII) from text.

**Evaluation Metrics.** The authenticity and reliability of any method depend on the comprehensiveness of its evaluation. In the field of archive review, false negatives may lead to the leakage of classified content, false positives may cause unnecessary reviews, and the robustness of the auditing method is also critical to the review task. To accurately evaluate our method, we selected a set of robust evaluation metrics, including accuracy, recall, F1 score, and adversarial sample attack success rate (ASR).

**Implementation Details.** We use a server equipped with 2 NVIDIA V100 GPUs, each with 32GB of memory. In the actual model fine-tuning of the experiment, we adopt the Paged AdamW Optimizer, set the learning rate at 2e-5, train epoch set at 5, and implement an early stopping strategy based on the accuracy of the validation set. For other hyper-parameters, we choose to set the rank $r$ of the low-rank matrix to 8 and the LoRA scaling factor $lora\_alpha$ to 32, which is used to scale the update amplitude of the LoRA weights (Hu et al., 2022). In terms of quantization-related parameters, we select a 4-bit quantization bitwidth to better control the precision of weight quantization and specify the quantization data type as nf4 (Dettmers et al., 2023).

**Baseline.** We compare our methods with the baselines listed as follows,and adaptively trained for the long-text characteristics inherent to archive review tasks:

- Longformer (Beltagy et al., 2020; Askari et al., 2023):leveraging a sparse attention mechanism (local windows + random sampling + global tokens), it optimizes Transformer's quadratic complexity to linear, supporting inputs up to 4,096 tokens. We performed full-parameter training on a specialized archival dataset to enhance its ability to model cross-paragraph dependencies, capturing long-range semantic associations and review rule features in archival texts.

- BigBird (Zaheer et al., 2020):achieving linear-time processing for long sequences via a combination of sliding local windows and task-specific global attention, it was also fully fine-tuned on the archival dataset.

### 4.2 Main Results

The performance comparison between ARGUS and the baseline methods in three datasets is shown in Table 1. To evaluate the effectiveness of ARGUS, the datasets we used cover different text types,and we conducted numerous experiments.

ARGUS outperforms all benchmarks in terms of precision and recall on all datasets. On the real-world archive dataset Archives, when using the DeepSeek-14B model, ARGUS achieves approximately a 10% improvement in inference accuracy compared to the baseline model, with a recall rate of 88.45%. This indicates that ARGUS exhibits better archive sensitivity than the baseline. This achievement is attributed to ARGUS's rule-embedded prompt optimization method and the "Semantic-Aware Feedback Reinforced" model enhancement strategy.

The inference process in ARGUS simultaneously generates explanatory text, allowing output results to be directly interpretable without additional explanation tools or post-hoc analysis—an advantage not shared by other baseline methods. This benefit stems from the large language model's inherent "Inherent Explainability" and the Hierarchical Rule-Embedded Prompting strategy.

Meanwhile, considering some application scenarios with extremely limited resources where it's impossible to deploy large-scale LLMs with a large number of parameters, we change the base LLM of our method to LLMs with fewer parameters, such as a 7B model (with approximately 7 billion parameters). Then we compare its performance improvement, with results shown in Table 2.

Notably, despite the foundational performance limitations of low-parameter LLMs, ARGUS effectively enhances their review accuracy. For example, when using DeepSeek-7B as the base model, the accuracy improvement reaches approximately 18%.

### 4.3 Effectiveness of Rule-Embedded

To verify the effectiveness of the rule-embedded prompt optimization adopted by ARGUS, we modified the prompts used for judgment. We employed conventional prompts (task statements combined with explanations of judgment criteria), fine-tuned the model using the same training dataset, and then initiated normal audit tasks (LLM-LoRA) on the archive dataset Archives, reviewing the model output logs before and after prompt modification. We also recorded the differences in false negative rates

| Methods | Archives | | | Crimes | | | ai4privacy | | |
|---|---|---|---|---|---|---|---|---|---|
| | ACC(%) | Recall(%) | F1-Score | ACC(%) | Recall(%) | F1-Score | ACC(%) | Recall(%) | F1-Score |
| Longformer | 68.11 | 68.08 | 0.6810 | 71.16 | 68.37 | 0.6973 | 70.32 | 66.43 | 0.5657 |
| Bigbird | 67.16 | 75.91 | 0.7127 | 75.56 | 78.95 | 0.7636 | 69.91 | 60.92 | 0.6524 |
| **ARGUS** | **78.27** | **88.45** | **0.8034** | **86.20** | **81.86** | **0.8510** | **80.10** | **76.88** | **0.7944** |

Table 1: Comparison of various evaluation indicators between ARGUS and the baseline in different datasets. ACC= accuracy, The base model adopted by ARGUS is DeepSeek-14B.

| Model | Framework | ACC(%) |
|---|---|---|
| DeepSeek-7B | BASE | 44.92 |
| | ARGUS | **62.23** |
| Qwen-7b | BASE | 53.79 |
| | ARGUS | **60.14** |
| Yi-9b | BASE | 49.84 |
| | ARGUS | **58.21** |
| BaiChuan2-7b | BASE | 47.92 |
| | ARGUS | **57.67** |

Table 2: Accuracy comparison of applying ARGUS on different small-scale LLMs over Archives, where BASE denotes test results without using the ARGUS framework.

| Method | MR(L1) | MR(L2) | MR(L3) |
|---|---|---|---|
| LLM-LoRA | 18% | 23% | 26% |
| **ARGUS** | **8%** | **7%** | **12%** |

Table 3: Comparison of model missed rates (MR) after removing rule embedding and ARGUS, where the base models in the table are all DeepSeek-14b models. MR = Missed Rate, L1 = PII-L, L2 = SLC-L, L3 = SOC-L.

| Model | Prompt | ACC(%) | Recall(%) | F1-Score |
|---|---|---|---|---|
| DeepSeek-14B | Base | 68.06 | 70.42 | 0.6887 |
| | HREP | 78.27 | 88.45 | 0.8034 |
| DeepSeek-32B | Base | 77.74 | 72.51 | 0.7657 |
| | HREP | 81.16 | 90.35 | 0.8314 |
| Qwen-32B | Base | 53.79 | 20.26 | 0.3055 |
| | HREP | 79.11 | 90.19 | 0.8125 |
| Mistral-24B | Base | 58.23 | 37.14 | 0.4714 |
| | HREP | 70.53 | 47.68 | 0.5690 |
| Gemma-27B | Base | 49.19 | 36.17 | 0.4167 |
| | HREP | 60.56 | 53.38 | 0.5759 |
| Yi-34B | Base | 51.77 | 25.56 | 0.3472 |
| | HREP | 72.58 | 62.86 | 0.6970 |

Table 4: Performance improvement differences between rule-embedded prompt(HREP) and regular prompts for LLMs on Archives.

across three critical judgment layers (PII, SLC, SOC) between the modified prompts and ARGUS, with results shown in Table 3.

Experimental results show that without rule embedding, even after fine-tuning with targeted samples, the missed rate of sensitive information at each layer for DeepSeek-14B will increase. For example, the missed rate of sensitive information related to personal information in the PII-L layer will increase by 10%. Additionally, in experiments we applied "Hierarchical Rule-Embedded Prompting" to different base large language models and evaluated their performance improvements, with results recorded in Table 4.

The results indicate that when general LLMs possess a certain level of basic performance, the rule-embedded prompt strategy can effectively enhance their performance in archive audit tasks. This improvement is particularly significant for Chinese large language models. Guided by rule-embedded prompts, the Qwen3-32B model achieves approximately a 16% increase in audit accuracy. Under the influence of the "Hierarchical Rule-Embedded Prompting" strategy, the DeepSeek-14B model's task performance is close to that of the DeepSeek-32B model, while its resource consumption during operation is significantly lower than that of the DeepSeek-32B.

### 4.4 Effectiveness of Feedback Reinforced

To validate the effectiveness of Semantic-Aware Feedback Reinforced used in ARGUS, we fine-tuned DeepSeek-14B with regular samples (HREP) and enhanced samples (ARGUS) respectively, and compared their adversarial sample attack success rates (ASR). The results are recorded in Table 5.

The experimental results show that after removing the feedback reinforced, the ASR under perturbations increased by approximately 25%. Similarly, we also compared the performance improvement of Semantic-preserving adversarial enhancement

| | ASR | ACC |
|---|---|---|
| HREP | 35.2% | 73.4% |
| **ARGUS** | **10.3%** | **86.20%** |

Table 5: Comparison of ASR before and after removing adversarial enhancement on perturbation dataset constructed based on Crimes , with DeepSeek-14B as base model.

for different base LLMs, as shown in Table 6.

| Model | Prompt | ACC(%) | Recall(%) | F1-Score |
|---|---|---|---|---|
| DeepSeek-14B | HREP | 78.27 | 88.45 | 0.8034 |
| | ARGUS | 81.61 | 90.35 | 0.8314 |
| DeepSeek-32B | HREP | 81.61 | 90.35 | 0.8314 |
| | ARGUS | 92.33 | 94.67 | 0.9251 |
| Qwen-14B | HREP | 63.79 | 60.26 | 0.6255 |
| | ARGUS | 77.89 | 89.86 | 0.7940 |

Table 6: Performance improvement comparison of feedback reinforced across different LLMs on Archives.

**Robustness verification.** Evaluate the robustness of ARGUS and baseline models using test sets with different perturbation types, and record their respective ASR. The perturbation types include: Lexical Perturbation(Zhang et al., 2016; Ebrahimi et al., 2018), Structural Perturbation(Yang et al., 2018), and Semantic Adversarial Perturbation(Jia and Liang, 2017).

| Methods | Perturbation | ACC(%) | ASR(%) |
|---|---|---|---|
| Longformer | Lexc | 60.32 | 19.10 |
| | Strut | 58.45 | 20.43 |
| | SemAdv | 55.43 | 22.21 |
| Bigbird | Lexc | 64.87 | 17.16 |
| | Strut | 62.62 | 19.80 |
| | SemAdv | 53.80 | 24.08 |
| ARGUS | Lexc | 81.65 | 5.78 |
| | Strut | 79.73 | 7.89 |
| | SemAdv | 77.03 | 10.32 |

Table 7: Performance comparison between ARGUS and baseline models under different perturbation types.Lexc=Lexical Perturbation, Stru=Structural Perturbation, SemAdv=Semantic Adversarial Perturbation.

### 4.5 Efficiency Analysis

**Time and Resource Consumption.** All inference and training experiments for ARGUS were conducted on servers equipped with two NVIDIA V100 32GB GPUs, as the base models used typically range from 7 to 16 billion parameters. An exception is the DeepSeek-32B model, for which we transferred its fine-tuning task to another server equipped with two NVIDIA L20 48GB GPUs.Table 8 presents the training and inference time of the baseline models on the archival dataset.

| Model | inference(h) | Fine-tuning(h) |
|---|---|---|
| DeepSeek-14B | 2.53 | 6.77 |
| Qwen3-14B | 2.27 | 3.95 |
| DeepSeek-32B | 5.52 | 9.49 |
| LLaMA2-13B | 3.57 | 6.36 |
| BaiChuan2-13B | 2.79 | 3.74 |

Table 8: Inference and training time of ARGUS using different baseline large language models. Inference on 1,500 archival documents and traning on 10000 pieces of archive data.

**Performance changes.** Our experiments reveal a positive correlation between model performance and parameter size: models with fewer parameters (such as 7B) generally underperform on real-world tasks. However, models with moderate parameter sizes (such as 14B and 32B) demonstrate sufficient capability to handle our tasks effectively. For instance, when ARGUS adopts DeepSeek-14B as its baseline model, its performance closely approaches that of DeepSeek-32B, while significantly reducing resource and time consumption. Specifically, DeepSeek-14B supports inference on a single NVIDIA V100 32GB GPU and basic fine-tuning on two such GPUs, whereas DeepSeek-32B requires at least two NVIDIA L20 48GB GPUs for comparable fine-tuning. This highlights the efficiency of our approach, which achieves strong performance with minimal computational overhead, largely attributed to ARGUS's Low-Rank Lightweight Adaptation.

## 5 Conclusion

We propose ARGUS, a feedback-reinforced gradual LLMs-based framework for interpretable and robust archive review. By integrating domain-specific rule embedding, feedback reinforced, and lightweight adaptation, ARGUS enhances LLM archive review performance and ensures robustness with lightweight resource consumption. Experiments show ARGUS achieves 10.5–15.5% higher accuracy than baselines, reduces ASR by 25%, Future work will explore expanding ARGUS into an automated archive processing tool to automate the entire workflow, from sensitive content identification and determination to its removal or masking.

## 6 Limitations

Although its excellent performance in archive review, ARGUS has 3 limitations in applications:

**Strong domain dependency.** Although Hierarchical Rule-Embedded Prompting significantly improves accuracy in archive review, it highly relies on the embedding quality of domain rules. In scenarios with unclear rules or highly dynamic rule changes, ARGUS may require additional rule optimization and manual intervention, increasing deployment costs.

**Limitations of feedback reinforcement.** ARGUS uses diffusion models for Semantic-Aware Feedback Reinforced enhancement of LLMs, which depends on error analysis in model inference logs. When paired with small-parameter LLMs (such as 7B parameters), limited baseline performance may cause large inference biases or incomplete log coverage, leading to generated samples that fail to fully address complex perturbations in real-world scenarios and thus limit reinforcement effectiveness.

**Constraints in lightweight adaptation.** While ARGUS employs QLoRA to reduce computational resource requirements, its adaptation efficiency for ultra-large models (such as 100B parameters) may still be constrained by hardware conditions. Additionally, low-rank adaptation cannot fully capture all knowledge required by the model in some complex tasks.

## References


ai4Privacy. 2023. pii-masking-200k (revision 1d4c0a1).

Arian Askari, Mohammad Aliannejadi, Amin Abolghasemi, Evangelos Kanoulas, and Suzan Verberne. 2023. Closer: Conversational legal longformer with expertise-aware passage response ranker for long contexts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 25–35, New York, NY, USA. Association for Computing Machinery.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Cheng-Han Chiang and Hung yi Lee. 2022. On the transferability of pre-trained language models: A study from artificial datasets. *Preprint*, arXiv:2109.03537.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2023. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. *Preprint*, arXiv:1712.06751.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, and 5 others. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Monica. Hedda, Bradley A. Malin, Chao. Yan, and Daniel. Fabbri. 2017. Evaluating the effectiveness of auditing rules for electronic health record systems. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2017:866–875.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Tim Hutchinson. 2018. Protecting privacy in the archives: Supervised machine learning and born-digital records. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2696–2701.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. 2025. Massive values in self-attention modules are the key to contextual knowledge understanding. *Preprint*, arXiv:2502.01563.

Poornima Kulkarni and K. Cauveryn. 2021. Personally identifiable information (pii) detection in the unstructured large text corpus using natural language processing and unsupervised learning technique. *International Journal of Advanced Computer Science and Applications*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Jingyu Peng, Maolin Wang, Xiangyu Zhao, Kai Zhang, Wanyu Wang, Pengyue Jia, Qidong Liu, Ruocheng Guo, and Qi Liu. 2025. Stepwise reasoning error disruption attack of llms. *Preprint*, arXiv:2412.11934.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Han Wu, Kun Xu, Linfeng Song, Lifeng Jin, Haisong Zhang, and Linqi Song. 2021. Domain-adaptive pretraining methods for dialogue understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 665–669, Online. Association for Computational Linguistics.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.

Rongen Yan, Xue Jiang, and Depeng Dang. 2021. Named entity recognition by using xlnet-bilstm-crf. *Neural Process. Lett.*, 53(5):3339–3356.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. Stylistic Chinese poetry generation via unsupervised style disentanglement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3960–3969, Brussels, Belgium. Association for Computational Linguistics.

Jianliang Yang, Xiya Zhang, Kai Liang, and Yuenan Liu. 2023. Exploring the application of large language models in detecting and protecting personally identifiable information in archival data: A comprehensive study*. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2116–2123.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3).

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification. *Preprint*, arXiv:1509.01626.

Yan Zhang, Mei-Po Kwan, and Libo Fang. 2025. An llm driven dataset on the spatiotemporal distributions of street and neighborhood crime in china. *Scientific Data*, 12(1):467.

Xiang Zheng, Shaoyu Chen, Junfei Wu, Lixiang Ruan, Zhaojun Luo, and Xiaojun Xu. 2023. A chinese word segmentation model for scd text in smart grid station: An attention-bilstm-crf approach. In *2023 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia)*, pages 952–957.

Hao Zhong, Dong Yang, Shengdong Shi, Lai Wei, and Yanyan Wang. 2024. From data to insights: the application and challenges of knowledge graphs in intelligent audit. *Journal of Cloud Computing (2192-113X)*, 13(1).

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.