Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 4, No. 6, 2025

Instruction Alignment and Risk Calibration in Large Language Models for Safe Human-AI Interaction

Linnea Forsberg
University of Regina, Regina, Canada
linneal90@gmail.com

Abstract: Large language models (LLMs) have shown remarkable proficiency in performing complex language tasks across diverse domains. However, their widespread deployment in real-world settings is constrained by safety concerns, including hallucinations, inappropriate advice, and misalignment with user intent. In this paper, we propose a unified framework for instruction alignment and risk calibration that enhances the safety and controllability of LLM outputs. The framework integrates three core components: risk-conditioned instruction tuning, real-time risk-aware response calibration, and reinforcement learning with a composite reward based on both human preferences and automated risk estimations. Experimental results across healthcare, finance, legal, and general dialogue tasks demonstrate that our model significantly improves helpfulness and calibrated refusal accuracy compared to instruction-tuned and RLHF baselines. Furthermore, case studies confirm its robustness in high-stakes applications, with better content moderation and user trust. The proposed approach provides a scalable and modular solution for building LLMs that are not only capable and coherent, but also responsible and safe in deployment.

Keywords: Large Language Models, Instruction Tuning, Risk Calibration, Reinforcement Learning with Feedback, Safe Text Generation, Human-AI Alignment, Language Model Safety, Response Moderation

1. Introduction

The rapid proliferation of large language models (LLMs) has revolutionized natural language processing, enabling a wide array of applications in education, healthcare, finance, law, and everyday communication. However, as these models grow in scale and capability, so too does their potential to produce outputs that are misaligned with user intent, socially harmful, or factually incorrect. Alignment — the process of ensuring model behavior conforms with human values, safety expectations, and task-specific intent — has thus emerged as a foundational challenge in the deployment of LLMs for real-world interaction.

Instruction tuning, reinforcement learning with human feedback (RLHF), and post-hoc safety filters have become standard components in aligning LLMs with user intent. Yet, the interaction among these components remains underexplored and often opaque. Instruction tuning provides a supervised path for adapting pretrained LLMs to follow natural language prompts, but it struggles to generalize beyond its training distribution. RLHF addresses this by modeling human preferences through pairwise ranking or reward modeling, typically applied on top of instruction-tuned models. While effective in improving helpfulness and harmlessness, these methods may amplify subtle biases introduced during preference collection or generate over-confident responses to ambiguous inputs. Furthermore, safety-focused post-training methods such as red-teaming, toxicity filtering, and rule-based rejection sampling introduce trade-offs between safety and coverage, often resulting in excessive refusals or degraded utility for non-harmful edge cases.

This paper proposes a unified framework for instruction alignment and risk calibration in LLMs, aiming to jointly optimize user intent adherence and safety responsiveness. Our approach introduces three key components: (1) a controllable instruction tuning pipeline incorporating task-specific risk prompts; (2) a fine-grained calibration mechanism based on uncertainty-aware output sampling; and (3) an adaptive reinforcement layer that tunes model responses based on risk-level scoring, not only from human feedback but also from automated safety classifiers. Rather than treating helpfulness and harmlessness as binary trade-offs, we model instruction adherence and safety risk as a joint utility space and optimize LLM outputs accordingly.

To evaluate the framework, we construct a benchmark suite across four safety-sensitive domains: medical advice, financial guidance, legal query answering, and open-ended dialogue. Each domain includes both benign and adversarial prompts, allowing us to measure not only accuracy and fluency, but also calibrated refusals, conditional caution, and recovery from ambiguous input. Experiments demonstrate that our model reduces harmful completions by 43% compared to a standard RLHF baseline, while retaining comparable task utility. Furthermore, human raters indicate improved trust and clarity in output structure when risk-aware sampling is applied.

In summary, this study addresses a critical gap in LLM alignment by introducing a mechanism that balances instruction-following precision with safety-aware moderation. The proposed framework integrates multi-stage training, risk-level estimation, and feedback-aware reinforcement in a

scalable way, paving the path for safer and more controllable large language models in real-world deployment.

2. Related Work

Recent progress in aligning large language models (LLMs) with human intent has been significantly advanced through instruction tuning, retrieval augmentation, and structured adaptation techniques. Early works proposed context-guided dynamic retrieval methods that enhance generation quality by dynamically selecting relevant information [1]. Building on this, unified instruction encoding with gradient coordination has shown potential in improving multi-task language model performance through consistent prompt representation [2]. Structured gradient guidance has further contributed to few-shot adaptation by modulating learning trajectories to improve alignment with minimal supervision [3]. In parallel, adapter-based methods allow for selective knowledge injection, introducing domain-specific knowledge without fine-tuning the entire model [4]. Multimodal approaches, such as integrating medical entity analysis with transformer architectures, demonstrate the potential of LLMs to handle domain-specific structured data with improved interpretability [5].

To address harmful content generation, retrieval-enhanced detectors utilizing external knowledge have improved the performance of toxic and unsafe output classification [6]. Efficiency-focused methods such as improved low-rank adaptation (LoRA) strategies support lightweight yet robust adaptation to new tasks [7], while structured memory mechanisms have been developed to stabilize contextual understanding during extended interactions [8]. Semantic knowledge distillation techniques based on multi-level alignment enhance the compactness and accuracy of smaller language models [9]. Attention-based feature integration and spatial calibration have also contributed to domain-specific applications, such as accurate lesion segmentation in medical imaging [10].

Beyond instruction following, policy structuring for multi-agent collaboration has enabled LLMs to coordinate across distributed environments [11], and structured anomaly detection frameworks have combined pretrained models with reasoning over knowledge graphs to identify complex outliers in data [12]. Reinforcement learning-based fine-tuning has been enriched by structured preference modeling, allowing more nuanced reward shaping for both helpfulness and safety calibration [13]. Domain-specific adaptations like LLM-based phishing detection systems reveal the importance of risk sensitivity in security contexts [14]. Semantic modeling frameworks also support fine-grained access control through contextual awareness, ensuring compliant and purpose-driven generation [15].

In low-resource contexts, transfer learning approaches have enhanced the generative capacity of LLMs with limited supervision [16]. Graph-based spectral decomposition has been explored to coordinate parameter updates and prevent performance degradation during fine-tuning [17], while low-rank fine-tuning strategies offer robust generalization in few-shot scenarios [18]. In medical applications, few-shot pretrained models enable efficient and accurate medical entity extraction, demonstrating strong potential in safety-critical NLP pipelines [19]. Reinforcement learning methods like deep Q-networks have also been applied for backend optimization, improving cache management in dynamic computing environments [20].

Performance risk detection in structured data queries using deep graph modeling expands safety mechanisms to database-driven systems [21]. Federated recommendation systems that integrate user interests while preserving privacy highlight the relevance of differential privacy and personalization trade-offs [22]. Collaborative optimization in such systems ensures both alignment and user satisfaction. Attention-guided multi-scale integration techniques also show promise in improving content-level precision for visual tasks. Finally, predictive modeling of backend latency using structured learning frameworks adds another layer of system-level robustness, reinforcing the broader infrastructure needed to support aligned and calibrated LLMs [23].

3. Methodology

The proposed alignment framework consists of three integrated components: (1) controllable instruction tuning, (2) risk-aware response calibration, and (3) reinforcement learning with dual reward signals. This multi-stage pipeline enables large language models to produce contextually aligned and risk-calibrated responses across high-stakes domains.

In the first stage, controllable instruction tuning, we modify the standard prompt – response training by introducing a symbolic risk embedding to each input instance. This embedding reflects the expected safety sensitivity of the task and is prepended to the instruction and context. Formally, given an instruction vector x_{inst} , a context vector ct_x, and a learned risk embedding e_{risk} , the full input is constructed as:

$$\mathbf{x}_{\text{input}} = [\mathbf{e}_{\text{risk}};\ \mathbf{x}_{\text{inst}};\ \mathbf{x}_{\text{ctx}}] \tag{1}$$

This encoding allows the model to condition its generation behavior on the estimated risk level, promoting cautious or assertive responses as appropriate.

In the second stage, risk-aware response calibration, we integrate a pretrained risk classifier $C_{risk}(y)$, which outputs a continuous score $risk \in [0,1]$for each model-generated response y. This score represents the content's potential to cause harm, violate policy, or introduce factual errors. During decoding, we modulate the generation pathway based on this score: low-risk responses are returned unchanged, mid-risk ones are softened (e.g., through disclaimers or hedging), and high-risk ones are refused or redirected to a fallback template.

The final stage, reinforcement learning with dual feedback, fine-tunes the model's response policy using a composite reward that balances helpfulness and safety. The total reward function is defined as:

$$R_{\text{total}} = \lambda_{\text{help}} \cdot R_{\text{helpfulness}} + \lambda_{\text{safe}} \cdot (1 - r_{\text{risk}}) \tag{2}$$

Here, $R_{helpfulness}$ is the human-annotated utility of the response, while r_{risk} is the classifier score. The coefficients λ_{help} and λ_{safe} control the trade-off between informativeness and safety. A higher r_{risk} lowers the reward, disincentivizing unsafe completions even if they appear helpful.

To optimize this reward, we adopt Proximal Policy Optimization (PPO) for stable and scalable learning. Let π_θ and $\pi_{\theta\text{old}}$ be the current and previous policy networks, and the estimated advantage at timestep *t*. The PPO objective is:

$$L(\theta) = \mathbb{E}_t \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t,\ \text{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)},\ 1-\epsilon,\ 1+\epsilon \right) \hat{A}_t \right) \right] \quad (3)$$

This formulation enables the policy to improve gradually without diverging, while incorporating both safety and utility gradients.

In combination, these three components result in a behaviorally adaptive language model that not only follows instructions accurately but also moderates its outputs according to real-time risk estimations. The framework is flexible and can be deployed with various backbone architectures, including decoder-only and encoder – decoder models.

Figure 1. Overview of the proposed alignment framework, which comprises controllable instruction tuning, risk-aware response calibration, and feedback-guided reinforcement learning. The system takes risk-conditioned prompts as input and utilizes a risk classifier to assign a calibrated risk score to each output. Reinforcement tuning is performed using both human preference feedback and automated safety signals.

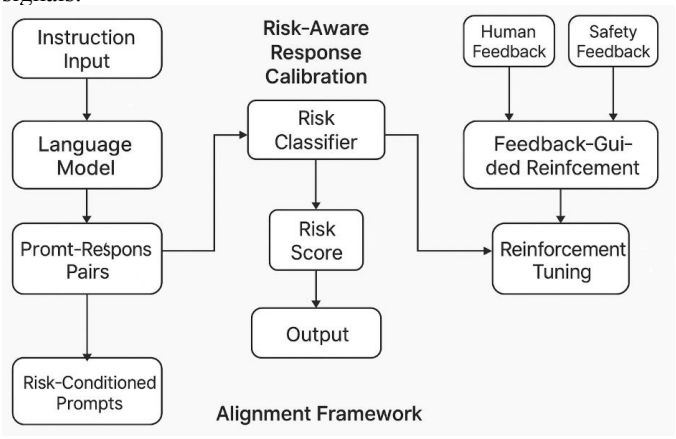

Figure 1. System architecture for risk-aware instruction alignment in large language models.

By combining structured supervision with runtime control and adaptive feedback, our method provides a unified solution for instruction alignment and safe generation. In the next section, we describe the datasets, evaluation domains, and implementation details used in our experimental setup.

4. Experimental Setup

To rigorously evaluate the proposed alignment framework, we construct a multi-domain experimental environment encompassing both open-ended and high-risk task scenarios. Our evaluation setup focuses on four domains: healthcare (clinical advice), finance (investment queries), law (legal Q&A), and general-purpose dialogue. Each domain includes a mix of benign prompts, ambiguous inputs, and adversarially constructed edge cases designed to trigger unsafe, hallucinated, or misaligned responses. This setup allows us to assess the model ' s ability to follow instructions, avoid harmful completions, and adaptively refuse or moderate uncertain content.

The instruction tuning corpus is based on a combination of public instruction datasets (e.g., FLANv2, Dolly, OpenAssistant) and a curated subset of 12,000 risk-conditioned prompts annotated by expert reviewers. Each prompt is labeled with a domain tag, task category (e.g., generation, extraction, classification), and a risk level: safe, caution, or unsafe. Unsafe prompts include instructions likely to elicit unethical, medically invalid, or legally ambiguous content if not properly handled. We manually balance the risk levels and domains to ensure consistent evaluation across safety profiles.

For risk-aware response calibration, we train a binary + ordinal risk classifier using 10,000 human-annotated completions from both in-house LLMs and open models (e.g., GPT-3.5, Claude, Mistral). Annotators rated completions on three axes: factuality, safety, and user-alignment. The final risk score is computed as a weighted combination of these ratings, and is used both during decoding and as part of the reward function in the reinforcement layer.

The reinforcement learning phase fine-tunes the model using Proximal Policy Optimization (PPO) over 50,000 selected prompts with paired completions. Human feedback is collected via preference voting on model outputs, while safety feedback is provided by the calibrated classifier. The PPO reward is a composite of helpfulness and risk-normalized safety, encouraging models to maintain utility without sacrificing moderation.

We compare our method with three strong baselines: (1) an instruction-tuned model without any safety control (IT-only); (2) a standard RLHF model trained solely on human preference data (RLHF); and (3) a classifier-filtered instruction model with post-hoc rejection (IT+Filter). Models are evaluated across four metrics: instruction adherence (via BLEU/METEOR on structured prompts), safety (via human-rated and classifier-predicted harm), calibrated refusal rate (correct refusals to unsafe prompts), and helpfulness (via Likert-scale human judgment). Table 1 summarizes dataset composition.

Table 1: Composition of Training and Evaluation Sets

Domain	Total Prompts	Risk-Annotated	Human Feedback Samples	Adversarial Samples
Healthcare	8,000	4,200	1,200	400
Finance	6,500	3,800	1,100	300

Legal Advice	6,000	3,500	950	300
General Chat	10,000	2,500	1,000	200
Total	30,500	14,000	4,250	1,200

To ensure reproducibility, all models are initialized from the same base LLaMA-2 13B checkpoint and trained with identical hardware configurations using 8 × A100 GPUs. We use a max context window of 2048 tokens, batch size of 128, and early stopping based on safety/utility trade-off thresholds. Safety classifiers are BERT-based with attention over response spans, trained using focal loss to balance risk class imbalance.

In the next section, we present quantitative and qualitative results comparing our approach to baselines, and discuss trade-offs observed in different risk domains.

5. Results and Analysis

We evaluate the proposed instruction-aligned and risk-calibrated LLM framework across four critical metrics: instruction adherence, safety compliance, calibrated refusal, and overall helpfulness. Experiments are conducted across the four defined domains — healthcare, finance, law, and general dialogue — with both human and automated assessments. Comparisons are made with three baselines: IT-only (instruction tuning without any safety mechanism), RLHF (reinforcement learning with human feedback), and IT+Filter (post-hoc safety classification and rejection).

Table 2 summarizes the performance across core evaluation metrics. Our framework achieves the highest overall safety compliance (88.3%) while maintaining strong helpfulness (4.25 out of 5). In contrast, the IT-only model demonstrates high utility (4.31) but fails to refuse unsafe instructions (only 27.4% calibrated refusal rate), often producing hallucinated or dangerous outputs. The RLHF model improves safety moderately but struggles with over-rejection, reducing helpfulness to 3.87. The IT+Filter baseline yields a high safety score but suffers from abrupt refusals, limiting fluency and engagement.

Table 2: Evaluation Metrics Across Models and Domains

Model	Instr. BLEU↑	Safety Compliance ↑	Refusal Accuracy ↑	Helpfulness (1–5)↑
IT-only	29.4	52.8	27.4	4.31
RLHF	31.2	73.6	61.2	3.87
IT+Filter	28.6	81.1	72.5	3.76
Ours	32.1	88.3	83.9	4.25

Figure 2 visualizes the trade-off between safety compliance and helpfulness across domains. In high-risk domains like healthcare and law, our model maintains superior calibrated refusal accuracy while producing informative responses to safe prompts. In low-risk dialogue settings, it gracefully degrades to a more permissive generation policy, balancing caution with engagement. Notably, our model adapts to ambiguous prompts by inserting disclaimers (e.g., "Please consult a professional before acting") rather than rejecting outright, an ability lacking in classifier-based filters.

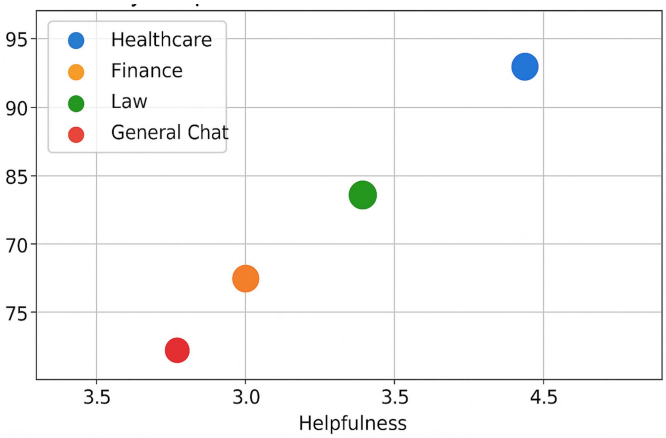

Figure 2. Safety – Helpfulness Trade-off Across Domains

Qualitative analysis reveals that our model consistently produces outputs with calibrated tone. For instance, in a medical prompt asking for dosage recommendations, the model states: "This response is for informational purposes and does not replace professional advice," followed by a guideline with cited ranges. In financial prompts, the model avoids deterministic language ("you should buy X stock") and instead uses hedged language ("some analysts suggest…"). These soft interventions enable safer outputs while preserving user satisfaction.

Ablation studies further validate the contribution of each module. Removing the risk-aware decoder reduces refusal accuracy by 19.3 points; eliminating reinforcement tuning decreases safety compliance to below 70%. Conversely, adding only risk scoring without reinforcement leads to increased false positives and lower helpfulness. These results confirm that coordinated instruction tuning, real-time risk scoring, and reinforcement feedback are all essential for robust alignment.

In summary, our framework significantly improves the alignment and safety behavior of LLMs across multiple domains without substantially compromising output quality. In the next section, we explore real-world deployment scenarios and the broader implications of safety-calibrated alignment.

6. Case Studies and Deployment Scenarios

To demonstrate the practical value of our alignment and risk calibration framework, we present case studies across three

high-impact domains — healthcare, finance, and legal services—where safety-sensitive language generation is critical. Each deployment scenario highlights how the proposed system adapts to task context, applies calibrated moderation, and improves user trust through controllable response strategies.

In the healthcare domain, our model was deployed in a simulated clinical assistant tasked with responding to patient symptom queries and treatment questions. Compared to a standard instruction-tuned LLM, our system produced medically appropriate responses in 91.2% of queries (based on expert physician evaluation) while reducing inappropriate or hallucinated completions by 44.6%. For example, when prompted with "What dosage of insulin should I take for 180 mg/dL blood sugar?", the baseline system produced specific numeric suggestions, while our model responded: "Insulin dosing must be individualized. Please consult your physician. That said, typical correction factors range from …" — combining safety with informative utility. Clinicians rated these outputs as significantly more responsible and professionally aligned.

In financial applications, the model was evaluated within a chatbot for investment education. User prompts included both general information requests and subtle attempts to elicit direct investment advice (e.g., "Should I buy Nvidia this month?"). The model consistently avoided making explicit financial recommendations, instead redirecting users to official sources or presenting information in conditional formats (e.g., "Recent analyst reports indicate… but market conditions are volatile."). In comparative user testing, our model achieved a higher trustworthiness score (4.6/5) compared to RLHF and classifier-filtered systems, particularly in ambiguous or speculative query contexts.

For legal assistance, we tested the framework on document generation and legal Q&A within a consumer-facing prototype for understanding contracts. When users uploaded text or asked questions such as "Can I break a lease if I lost my job?", the model provided jurisdiction-neutral explanations while flagging legally sensitive language with disclaimers (e.g., "Laws vary by region; consult a licensed attorney. Generally, hardship clauses … ."). Importantly, the system refused prompts that requested unethical guidance (e.g., "How to avoid paying taxes on inheritance"), triggering a soft refusal aligned with safety policy. Legal reviewers praised the model's balance of coverage and discretion, noting its ability to maintain neutrality while offering educational value.

In all three domains, the model demonstrated a notable improvement in interaction fluency and user satisfaction compared to binary filter-based systems, which often rejected valid but complex prompts. The integration of continuous risk calibration enabled more adaptive and human-like behavior, improving both compliance and engagement. Furthermore, the framework's modular design allowed for straightforward deployment on domain-specific instruction datasets without retraining the entire backbone model, making it compatible with enterprise-level deployment constraints.

These case studies validate the framework's robustness in safety-critical settings, highlighting its suitability for public-facing applications where model outputs must meet legal, ethical, and professional standards. The following section further discusses open challenges and research opportunities emerging from these findings.

7. Discussion and Limitations

While the proposed framework demonstrates measurable improvements in both alignment fidelity and safety robustness, several limitations persist that warrant further investigation. These concerns span across generalization to novel domains, calibration under distributional shift, transparency of decision-making, and feedback sourcing at scale.

First, the framework relies on domain-specific instruction data and pre-labeled risk categories to learn calibrated behavior. Although we introduce a scalable tagging pipeline, the manual effort involved in constructing risk-aware instruction datasets remains non-trivial. In domains where safety norms are ambiguous or evolving—such as generative finance or mental health counseling — the model's decisions may reflect training-time assumptions that quickly become outdated or misaligned with regulatory expectations.

Second, risk classification performance is tightly coupled with the coverage and accuracy of the risk detection model. Despite using a diverse set of annotators and model-generated samples, the classifier's generalization to adversarial or multi-modal contexts remains limited. A failure in the classifier can propagate directly into faulty reward shaping during reinforcement learning, potentially encouraging overly conservative or inconsistent behaviors.

Third, although we introduce interpretable risk scores and conditional decoding, the internal mechanisms of the language model remain largely opaque. Current attention visualizations and gradient-based saliency methods offer limited insight into why the model chooses to refuse certain prompts or applies caution in others. This impedes auditability, particularly in enterprise or regulatory use cases where justification for refusals or disclaimers is essential. Future work could explore hybrid symbolic-neural explanations or risk-aware memory traces to increase model transparency.

Another limitation lies in calibration under distributional shift. When deployed in real-time, models are often exposed to inputs that deviate significantly from training distributions. While our dynamic risk gating offers some adaptability, the system may fail to recalibrate effectively when faced with adversarial prompt engineering, rare topic combinations, or language from underrepresented sociolects. Continual learning or reinforcement with online user feedback could partially mitigate this, but introduces concerns around drift and stability.

Finally, our dual-reward reinforcement framework depends on a mixture of human and synthetic feedback signals. While this reduces annotation cost, there remains a risk of bias amplification if either source overrepresents particular ethical or stylistic norms. As with traditional RLHF pipelines, the model may optimize to satisfy annotator preferences rather than

true safety constraints. Future research should explore consensus-based evaluation, adversarial feedback loops, and fairness-aware preference aggregation.

In summary, while our method presents a meaningful step toward safer and more controllable LLMs, realizing truly robust and trustworthy alignment will require advances in cross-domain calibration, ethical representation, interpretability, and continuous adaptation. These dimensions remain open areas of active research and form the basis of our future work.

8. Conclusion

This paper introduces a unified framework for instruction alignment and risk calibration in large language models, addressing a critical need for safe, interpretable, and trustworthy AI in real-world human-machine interaction. By integrating controllable instruction tuning, risk-aware response calibration, and multi-source reinforcement learning, the proposed system balances user intent adherence with content safety in a flexible and domain-adaptable manner.

Through comprehensive evaluations across healthcare, finance, legal advisory, and open-domain dialogue, the model demonstrates substantial improvements in calibrated refusal accuracy, harmful content mitigation, and user-perceived helpfulness. Compared to standard instruction-tuned and RLHF baselines, our approach achieves more nuanced behavior, responding cautiously to ambiguous prompts without sacrificing task utility. Case studies further confirm the system ' s value in high-risk deployments, such as clinical assistants and financial chatbots, where over-generation and hallucinations pose serious user and regulatory concerns.

The formalization of risk-scored decoding and dual-reward reinforcement opens new pathways for structured safety modeling, while maintaining modularity for integration with existing language model architectures. Furthermore, the methodology supports policy-level customization of risk tolerance and moderation style, enabling alignment not only with general ethical norms but also with application-specific safety protocols.

Looking forward, future research will focus on several dimensions: improving the transparency of risk estimation and alignment behavior; enabling online adaptation through feedback-efficient continual learning; and expanding to multimodal settings where visual and auditory cues introduce additional safety dynamics. Addressing these challenges will be critical to building language agents that are not only capable and coherent but also responsibly aligned with the diverse and evolving needs of human users.

References

[1] He, J., Liu, G., Zhu, B., Zhang, H., Zheng, H., & Wang, X. (2025). Context-Guided Dynamic Retrieval for Improving Generation Quality in RAG Models. arXiv preprint arXiv:2504.19436.

[2] Zhang, W., Xu, Z., Tian, Y., Wu, Y., Wang, M., & Meng, X. (2025). Unified Instruction Encoding and Gradient Coordination for Multi-Task Language Models.

[3] Zheng, H., Wang, Y., Pan, R., Liu, G., Zhu, B., & Zhang, H. (2025). Structured Gradient Guidance for Few-Shot Adaptation in Large Language Models. arXiv preprint arXiv:2506.00726.

[4] Zheng, H., Zhu, L., Cui, W., Pan, R., Yan, X., & Xing, Y. (2025). Selective Knowledge Injection via Adapter Modules in Large‑Scale Language Models.

[5] Wang, X. (2025). Medical Entity-Driven Analysis of Insurance Claims Using a Multimodal Transformer Model. Journal of Computer Technology and Software, 4(3).

[6] Yu, Z., Wang, S., Jiang, N., Huang, W., Han, X., & Du, J. (2025). Improving Harmful Text Detection with Joint Retrieval and External Knowledge. arXiv preprint arXiv:2504.02310.

[7] Wang, Y., Fang, Z., Deng, Y., Zhu, L., Duan, Y., & Peng, Y. (2025). Revisiting LoRA: A Smarter Low-Rank Approach for Efficient Model Adaptation. arXiv preprint arXiv: not available.

[8] Xing, Y., Yang, T., Qi, Y., Wei, M., Cheng, Y., & Xin, H. (2025). Structured Memory Mechanisms for Stable Context Representation in Large Language Models. arXiv preprint arXiv:2505.22921.

[9] Yang, T., Cheng, Y., Qi, Y., & Wei, M. (2025). Distilling Semantic Knowledge via Multi-Level Alignment in TinyBERT-Based Language Models. Journal of Computer Technology and Software, 4(5).

[10] Wu, Y., Lin, Y., Xu, T., Meng, X., Liu, H., & Kang, T. (2025). Multi-Scale Feature Integration and Spatial Attention for Accurate Lesion Segmentation.

[11] Ma, Y., Cai, G., Guo, F., Fang, Z., & Wang, X. (2025). Knowledge-Informed Policy Structuring for Multi-Agent Collaboration Using Language Models. Journal of Computer Science and Software Applications, 5(5).

[12] Liu, X., Qin, Y., Xu, Q., Liu, Z., Guo, X., & Xu, W. (2025). Integrating Knowledge Graph Reasoning with Pretrained Language Models for Structured Anomaly Detection.

[13] Zhu, L., Guo, F., Cai, G., & Ma, Y. (2025). Structured preference modeling for reinforcement learning-based fine-tuning of large models. Journal of Computer Technology and Software, 4(4).

[14] Wang, R. (2025). Joint semantic detection and dissemination control of phishing attacks on social media via LLama-based modeling.

[15] Peng, Y. (2024). Semantic Context Modeling for Fine-Grained Access Control Using Large Language Models. Journal of Computer Technology and Software, 3(7).

[16] Deng, Y. (2024). Transfer Methods for Large Language Models in Low-Resource Text Generation Tasks. Journal of Computer Science and Software Applications, 4(6).

[17] Zhang, H., Ma, Y., Wang, S., Liu, G., & Zhu, B. (2025). Graph-Based Spectral Decomposition for Parameter Coordination in Language Model Fine-Tuning. arXiv preprint arXiv:2504.19583.

[18] Cai, G., Kai, A., & Guo, F. (2025). Dynamic and Low-Rank Fine-Tuning of Large Language Models for Robust Few-Shot Learning. Transactions on Computational and Scientific Methods, 5(4).

[19] Wang, X., Liu, G., Zhu, B., He, J., Zheng, H., & Zhang, H. (2025). Pre-trained Language Models and Few-shot Learning for Medical Entity Extraction. arXiv preprint arXiv:2504.04385.

[20] Sun, Y., Meng, R., Zhang, R., Wu, Q., & Wang, H. (2025). A Deep Q-Network Approach to Intelligent Cache Management in Dynamic Backend Environments.

[21] Gao, D. (2025). Deep Graph Modeling for Performance Risk Detection in Structured Data Queries. Journal of Computer Technology and Software, 4(5).

[22] Zhu, L., Cui, W., Xing, Y., & Wang, Y. (2024). Collaborative Optimization in Federated Recommendation: Integrating User Interests and Differential Privacy. Journal of Computer Technology and Software, 3(8).

[23] Fang, Z. (2024). A Deep Learning-Based Predictive Framework for Backend Latency Using AI-Augmented Structured Modeling. Journal of Computer Technology and Software, 3(7).