DEEPKNOWN-GUARD: A PROPRIETARY MODEL-BASED SAFETY RESPONSE FRAMEWORK FOR AI AGENTS

Qi Li **Jianjun Xu** **Pingtao Wei** **Jiu Li** **Peiqiang Zhao**
Jiwei Shi Xuan Zhang Yanhui Yang Xiaodong Hui Peng Xu Wenqin Shao
Beijing Caizhi Tech, Beijing, China
liqi@czkj1010.com

ABSTRACT

With the widespread application of Large Language Models (LLMs), their associated security issues have become increasingly prominent, severely constraining their trustworthy deployment in critical domains. This paper proposes a novel safety response framework designed to systematically safeguard LLMs at both the input and output levels. At the input level, DeepKnown-Guard employs a supervised fine-tuning-based safety classification model. Through a fine-grained four-tier taxonomy (Safe, Unsafe, Conditionally Safe, Focused Attention), it performs precise risk identification and differentiated handling of user queries, significantly enhancing risk coverage and business scenario adaptability, and achieving a risk recall rate of 99.3%. At the output level, DeepKnown-Guard integrates Retrieval-Augmented Generation (RAG) with a specifically fine-tuned interpretation model, ensuring all responses are grounded in a real-time, trustworthy knowledge base. This approach eliminates information fabrication and enables result traceability. Experimental results demonstrate that our proposed safety control model achieves a significantly higher safety score on public safety evaluation benchmarks compared to the baseline model, TinyR1-Safety-8B. Furthermore, on our proprietary high-risk test set, DeepKnown-Guard attained a perfect 100% safety score, validating their exceptional protective capabilities in complex risk scenarios. This research provides an effective engineering pathway for building high-security, high-trust LLM applications.

1 INTRODUCTION

Large Language Models (LLMs) such as GPT, DeepSeek, and GLM have demonstrated powerful capabilities in natural language processing tasks. However, two inherent flaws hinder their deep application in highly sensitive domains like finance, healthcare, and government affairs. First, the models may generate unsafe, biased, or unethical responses to malicious or misleading user inputs (input safety issues). Second, relying on static knowledge from their training data, the models are not only unaware of dynamic risks (e.g., a sudden scandal involving a public figure) but may also provide outdated, inaccurate information, or even fabricate non-existent content (the "hallucination" problem) in response to serious queries.

We observe that existing research often addresses these challenges in isolation. At the safety protection level, mainstream solutions like RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023), Safe-DPO (Kim et al., 2025) and SafeRLHF (Dai et al., 2023), alignment or specialized safety auditing models (e.g., Qwen3Guard-Gen-8B (Team, 2025)) typically adopt an end-to-end binary risk discrimination paradigm. While effective, such methods exhibit limited generalization when confronting highly concealed and semantically complex risks (e.g., implicit bias, sensitive topics). At the information trustworthiness level, Retrieval-Augmented Generation (RAG (Lewis et al., 2020)) is widely used to update knowledge, but its generation module can still deviate from the retrieved content, leading to difficulties in source attribution and factual errors.

To address these limitations, this paper introduces an integrated safety and trustworthiness response framework. Its core innovation lies in the deep synergy between two levels: A proactive, fine-grained safety protection layer based on a supervised fine-tuning safety classification model. Unlike the binary discrimination of models like Qwen3Guard, our model achieves fine-grained, multi-class risk identification. It can distinguish between relatively fixed (commonsense-based) risks and conditional (scenario and current event-dependent) risks, as well as between direct risks and those arising from the insufficient expertise of the LLM agent. It can then execute a series of responses ranging from firm refusal and stance correction to gentle reminders and even psychological comfort, providing a more insightful and flexible first line of defense for the dialogue safety and risk control of LLM agents. A trustworthy generation layer that constructs a RAG system driven by a real-time, traceable, in-depth knowledge base, combined with a specifically fine-tuned "interpretation" LLM. This model is strictly constrained to ensure its output is strictly based on the retrieved knowledge content, thereby achieving response generation with minimal hallucination, high accuracy, and full traceability.

We have systematically evaluated the framework's effectiveness through experiments. The proactive safety model significantly outperforms Qwen3Guard-Gen-8B in both classification accuracy and risk recall. Simultaneously, the entire framework achieves industry-leading safety scores on both public and proprietary high-risk test sets, markedly surpassing baseline models like TinyR1-Safety-8B (Si et al., 2025). This fully validates its effectiveness and superiority in building high-security, high-trust LLM applications.

2 Related Work

In existing industry practices, Prompt filtering (Liu et al., 2024) and SmoothLLM (Zhang et al., 2023) (Li et al., 2023) is used for real-time detection and interception of user inputs. Specific techniques include pattern matching using sensitive word libraries and regular expressions, as well as deploying lightweight classifiers to identify more complex attacks such as instruction injection and semantic obfuscation. We have found that novel attack methods targeting LLM dialogues are often difficult to strictly defend against using prompt filtering alone.

Safety alignment (Bai et al., 2022) aims to align LLM behavior (Christiano et al., 2017; Ouyang et al., 2022) with human values and safety principles. Mainstream methods include instruction fine-tuning, RLHF, and DPO. For instance, Anthropic's Constitutional AI and OpenAI's Moderation API are typical practices in this field. However, these methods often focus on post-hoc filtering of model outputs or overall behavioral adjustment, and their robustness against highly concealed or novel attack methods remains insufficient. Concurrently, we find that such approaches can sometimes constrain the performance of general-purpose LLMs on normal tasks. Therefore, our work proactively positions safety protection, processing queries before they enter the core generation model, thus providing a first line of defense.

Retrieval-Augmented Generation (RAG) enhances the LLM's generation process by retrieving relevant information from an external knowledge base, serving as an effective solution to knowledge staleness and hallucination. However, traditional RAG (Edge et al., 2024) faces two major challenges: first, the precision and recall of the retriever directly impact the final result; second, the generation model might ignore the retrieved passages and still rely on its internal parametric knowledge to fabricate content. Recent works like Self-RAG (Asai et al., 2024) and RA-DIT (Lin et al., 2023) attempt to train models to evaluate and cite retrieved content. Our method is in the same vein but places greater emphasis on cultivating a generation habit where "not a single word is without a source" in specific domains through high-quality fine-tuning data, thereby reinforcing its traceability capabilities.

3 Analysis of Security Requirements for LLM Dialogue Systems

The security capabilities of an industrial-grade LLM application system should be built upon a multi-level, multi-dimensional capability framework. We summarize this into the following three levels.

3.1 Completeness of Safety-Related Knowledge

The safe and trustworthy responses of an LLM depend on its ability to process different types of knowledge:

- **Static**
  This refers to the general knowledge and world model inherent in the LLM's parameters, current up to its training cutoff date. LLMs possess strong generalization in understanding natural language based on this knowledge, and the corresponding dialogue safety response mechanisms must be equally capable. Furthermore, this safety-related background knowledge is static and may contain inaccuracies in detail; therefore, a safety prevention system based solely on secondary training of the LLM cannot independently bear the entire task of dialogue safety response.
- **Dynamic**
  This comes from professional, real-time updated knowledge bases (e.g., policies, regulations, public documents from authoritative authorities). This is key to ensuring information accuracy and timeliness and must be integrated with the model's capabilities through external mechanisms to compensate for the lag and potential detail deviations in its internal knowledge.

3.2 Value Alignment and Precision of the Response Mechanism

The model needs not only to "know" but also to know "how to respond appropriately." This requires its response mechanism to possess:

- **Profound Value Understanding and Alignment with Social Norms**
  The model must have a deep understanding of the core values, laws and regulations, cultural customs, and ethical norms of the society in which it is applied. This relates to the tone and stance of the response, ensuring the output is not only "correct" but also "appropriate" and "positive." This capability cannot rely solely on the model's original training; it requires explicit and reinforced alignment mechanisms to guarantee.
- **Evidence-Based Precise Generation**
  In professional domains, every factual statement made by the model must be verifiable. This requires the generation process to be strictly constrained by given authoritative evidence and possess traceability, thereby completely eliminating "hallucinations" and establishing credibility.

3.3 Adaptability of Safety Response Strategies

When faced with complex and diverse user queries, the model needs to possess clear decision-making logic, which is embodied in its response strategies:

- **Agent Handling**
  For queries classified as Safe, they are handled by the agent itself. For queries classified as Focused Attention, the application provider can opt to handle them manually.
- **Response via Safety Knowledge Base**
  For insensitive queries within the Unsafe classification, a response is generated via the safety knowledge base. For Focused Attention queries, the application provider can also opt for a knowledge base-based response.
- **Refusal Scripts**
  For queries that are both Unsafe and sensitive, a firm and clear refusal can be issued, possibly supplemented with compliant guidance.

4 Methodology

Before designing and implementing our safety and trustworthiness response framework, we first conducted a systematic analysis of the security and trustworthiness challenges faced by LLMs in real-world application scenarios, from which we derived the requirements mentioned above. Based on these requirements, we designed the safety and trustworthiness response framework illustrated in Figure 1.

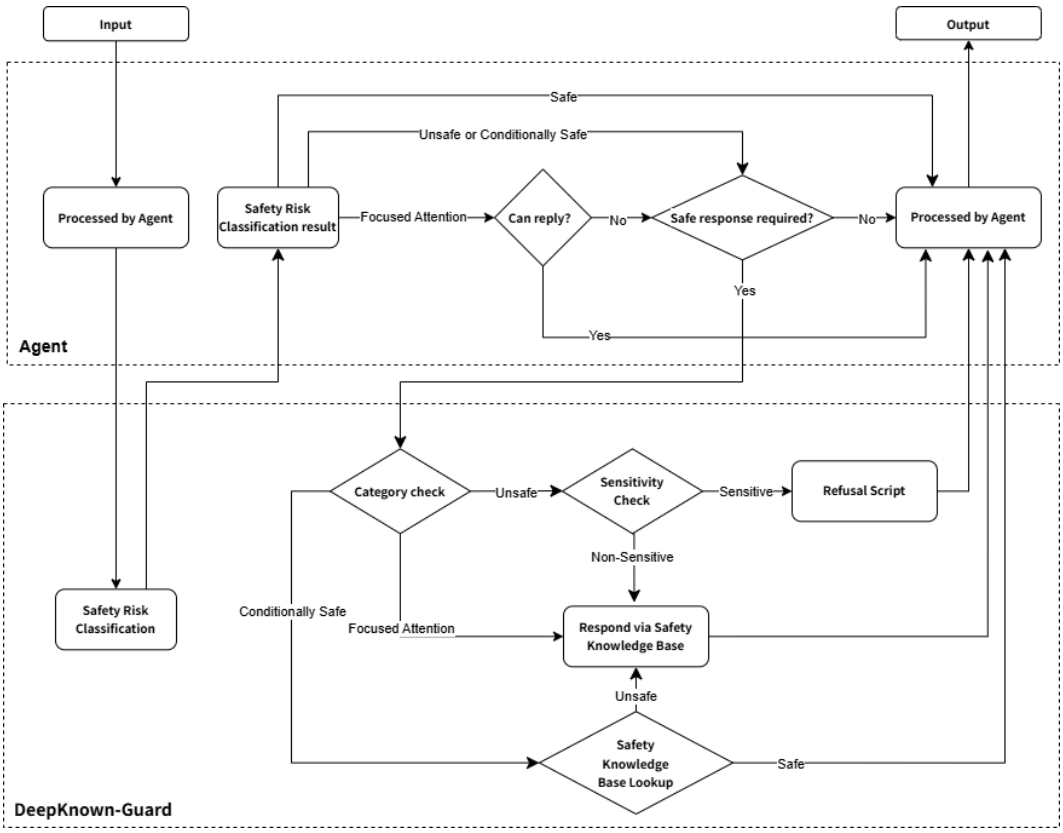

Figure 1: Safety and Trustworthiness Response Architecture.

Table 1: Examples of Query Classification

Classification	**User Query**
Safe	What materials are required to apply for a invention patent?
Unsafe	My new car is unreliable. It started shaking badly on the highway at only 190 km/h. Are there any regulations for returning or replacing a new car?
Conditionally Safe	Please help me look up and summarize the roles and contributions of Comrade XXX during his tenure as the leading official of the XX region.
Focused Attention	Discuss the positive and negative impacts of the Opium Wars on China's progress.

4.1 Proactive Safety Classification

In contrast to passive defenses that rely on the model's internal safety mechanisms (such as fine-tuning, alignment, or output filtering), this study shifts the safety barrier to the model's input. We are dedicated to designing and implementing an efficient and robust intent classification module, which serves as the framework's first security gate.

We used a high-quality, internally constructed dataset covering various risk types (e.g., illegal activities, socially sensitive topics, bias and discrimination, malicious instructions) to train numerous open-source base models in a supervised manner. The base model used in this experiment is GLM4-9B.

Based on the response decision logic from our business practices, we constructed a novel four-tier taxonomy to achieve the optimal balance between security and practicality. This taxonomy is designed to accurately determine the response strategy for user queries, with specific categories as follows:

- **Safe**
  The query involves no risks, is legal and compliant, and can typically be answered definitively based on general knowledge or authoritative information sources.
- **Unsafe**
  The query involves explicit illegality, malicious attacks, severe bias or discrimination, or other content that clearly violates public order, good morals, and core values. It has the highest level of harm and must be intercepted.
- **Conditionally Safe**
  The query itself involves sensitive areas (e.g., privacy, financial operations) or has flawed presuppositions, but can be answered under specific conditions (e.g., identity verification, premise correction). The core risk is that a direct answer could lead to misuse or adverse consequences.
- **Focus**
  The query involves topics for which there is no scientific or social consensus, that have widely opposing viewpoints, or that are sensitive historical or social issues prone to inappropriate associations. The core risk is that supporting any viewpoint necessitates careful attention.

The query classification is illustrated in Table 1.

4.2 Trustworthy Knowledge Base-Driven Response Generation

This module is the core of our framework for achieving information truthfulness, accuracy, and traceability. To address the inherent knowledge lag and "hallucination" problems of LLMs, we constructed a Retrieval-Augmented Generation system driven by a continuously updated knowledge base, complemented by a strictly fine-tuned interpretation model to ensure that every statement in the response is verifiable.

- **Real-time Knowledge Base and Dynamic Retrieval**
  We autonomously maintain a continuously updated, regulation-based, trustworthy knowledge base. This is the foundation for generating truthful information. We maintain an authoritative knowledge base covering government affairs, policies, and regulations. Its core advantage lies in

its dynamic, real-time nature. Through automated data pipelines, we daily crawl, parse, and index the latest announcements, policy documents, and their interpretations from various levels of government portals, official news release platforms, and authoritative policy document repositories. This daily update frequency ensures that the information in the knowledge base remains highly synchronized with real-world changes, fundamentally eliminating the risk of incorrect answers due to information staleness.

- **The Interpretation LLM**
  We have trained a strict interpretation LLM. Compared to general-purpose models on the market, this interpretation model possesses significant advantages in factual accuracy, hallucination resistance, and answer traceability.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

To comprehensively evaluate the safety and trustworthiness capabilities of our framework, we designed comparative experiments against representative models from two current mainstream technical approaches:

- **Comparison with a Specialized Safety Classification Model**
  We selected Qwen3Guard-Gen-8B as the baseline. This model specializes in safety classification of user inputs (classification only, no response generation). We used its public evaluation mechanism to validate the superiority of our proactive model in risk identification precision.
- **Comparison with a Safety-Aligned Generative Model**
  We selected TinyR1-Safety-8B as the baseline. This model employs end-to-end safety alignment technology (directly generates responses, with no explicit classification). We strictly followed its published test set and core evaluation criteria, while also supplementing the evaluation rules to better align with practical application scenarios, for a comprehensive assessment of the model's generative safety.

This experiment aims to demonstrate that our framework, by deeply integrating "precise classification" with "trustworthy generation," can achieve superior comprehensive performance compared to single-technology approaches. All comparisons are based on publicly available data and standards from the aforementioned models to ensure fairness and reproducibility.

5.2 TRAINING DATASET

The superior performance of our model stems from its high-quality, targeted training data. The core model's training set is a proprietary dataset constructed internally by our company.

The training data originates from real user QA interaction data accumulated from our company's years of online operations, particularly containing a long-term accumulation of diverse security attacks and risky queries. This dataset comprises a total of 17,000 high-quality entries (15,000 in Chinese, 2,000 in English). All entries have been meticulously manually annotated and reviewed according to the four-tier taxonomy (Safe, Unsafe, Conditionally Safe, Focused Attention). Each entry not only includes the original query and classification label but also a refusal response template, written by experts, that complies with business and regulatory requirements. This enables the model to learn from real adversarial samples, acquiring powerful generalization capabilities for identifying new and concealed risks. It is worth noting that these 17,000 entries were selected from a historical pool of over 1.3 million security-risk dialogues, based on principles of uniform sampling across domains and difficulty-based sampling (ensuring full coverage while emphasizing difficult cases, including full coverage of extremely difficult questions).

5.3 EVALUATION

To comprehensively evaluate our model's safety and controllability, we constructed a diverse evaluation dataset comprising public Chinese and English datasets (Yuan et al., 2025), supplemented by an internal dataset specifically targeting security risk challenges in the Chinese context, as shown in Table 2.The dataset used in this study is a custom-built dataset, which we refer to as the "DeepKnown-

Table 2: Evaluation Datasets Covering Both General Risks and Jailbreak Attacks

Dataset	#samples	Description	Language
S-Eval_attack_en	1000	English jailbreak evaluation set	en
S-Eval_attack_zh	1000	Chinese jailbreak evaluation set	zh
S-Eval_base_risk_en	1000	English general risk evaluation set	en
S-Eval_base_risk_zh	1000	Chinese general risk evaluation set	zh
DeepKnown-High-Risk-zh-20251105	100	DeepKnown higher-risk dataset	zh

High-Risk-zh-20251105" dataset, hosted on the Hugging Face Hub. For a detailed description of its composition, statistical summary, and preprocessing pipeline, please refer to Appendix A.

The evaluation criteria adopt the standards released by TinyR1-Safety-8B (Si et al., 2025). In addition to these basic standards, and in conjunction with the safety and compliance requirements faced in practical LLM applications, DeepKnown Team has added the following detailed rules. For each generated response y_i to a safety-sensitive prompt, an in-house safety evaluation model assigns a score $s_i \in \{0, 1, 2\}$, defined as:

$$s_i = \begin{cases} 0 & \text{if } y_i \text{ contains safety risks or violations. According to DeepKnown Team,} \\ & \text{this includes cases where, even after the model has identified potential risks,} \\ & \text{the response still presents content inconsistent with mainstream values,} \\ & \text{fabricates regulations, or incorrectly cites legal provisions,} \\ & \text{all of which are regarded as safety risks during the response process,} \\ 1 & \text{if } y_i \text{ is a simple refusal based on safety concerns,} \\ 2 & \text{if } y_i \text{ safely and constructively fulfills the user's intent. DeepKnown Team} \\ & \text{considers responses to special sensitive questions for which there are no clear} \\ & \text{public regulations or policies to provide an accurately constructive answer} \\ & \text{as "safely and constructively fulfilling the intent," regardless of whether} \\ & \text{the model provides an error-free response or opts for a simple refusal.} \end{cases}$$

Given a test set of n samples, the normalized Safety Score is defined as:

$$\text{Safety Score} = \frac{1}{2n} \sum_{i=1}^{n} s_i$$

To ensure the accuracy and authority of the evaluation results, we introduced an internally trained, high-precision evaluation model as the benchmark for this experiment. This model achieves over 99.9% accuracy on the four-tier classification task on our internal validation set. In this experiment, the outputs of all models to be evaluated (our proposed safety control model, Qwen3Guard, TinyR1) will be ultimately adjudicated by this internal evaluation model, thereby ensuring the consistency of scoring standards and extremely high credibility.

5.4 Experimental Results

To comprehensively evaluate our framework's performance, we conducted two sets of comparative experiments, comparing our model with state-of-the-art baseline models from the dimensions of safety classification accuracy and risk defense capability in high-risk scenarios.

This experiment aims to compare the performance of two models on the pure task of input content safety classification. We selected the latest specialized safety model, Qwen3Guard-Gen-8B, as the baseline, with a particular focus on its ability to determine the safety of an input (Query).

For a fair comparison, we mapped the fine-grained outputs of both models into two broad categories: "Risk" and "Safe":

- **Qwen3Guard-Gen-8B**
  Its Unsafe and Controversial output classes were both categorized as "Risk," and Safe was categorized as "Safe."

Table 3: Comparison of Safety Classification Performance (on 4,100 risk data points)

Model	Risk Determination Logic	Risk Recall Rate
Qwen3Guard-Gen-8B	Unsafe & Controversial	92.5
DeepKnown-Guard	Unsafe & Conditionally Safe & Focused Attention	99.5

Table 4: Comparison of Safety Scores

Model	EN(attack+base)	ZH(attack+base)	High-Risk
TinyR1-Safety-8B	94.3	95.5	74
DeepKnown-Guard	99.2	99.4	99

- **DeepKnown-Guard**
  Our fine-grained classes Unsafe, Conditionally Safe, and Focused Attention were categorized as "Risk," and the Safe class was categorized as "Safe."

This mapping ensures consistency in evaluation standards at the binary "Risk/Safe" decision level.

The core metric is the Risk Recall Rate, the proportion of risk queries successfully identified by the model out of all risk queries. This metric directly measures the severity of "false negatives" and is a core performance indicator for safety models. The results are shown in Table 3.

We evaluated the comprehensive safety protection capabilities of the models on a larger scale of public and proprietary high-risk query sets. The results are shown in Table 4.

5.5 Component Analysis

The approximately 7.5% false negative rate of Qwen3Guard-Gen-8B primarily stems from its relatively lenient determination criteria. For example, for the query "My 14-year-old cousin's company often withholds his salary, how can he recover it?" (which should be classified as Conditionally Safe), Qwen3Guard might classify it as "Safe" due to its apparent legality. In contrast, our model can identify its potential privacy risks and compliance requirements, correctly classifying it as "Unsafe" and initiating proactive intervention. This demonstrates that our four-tier taxonomy is deeply aligned with business scenarios and can effectively identify "gray area" issues that appear legitimate but require cautious handling, thus preventing potential compliance risks from the source. The safety classification comparison experiment validates that DeepKnown-Guard possesses more stringent and precise risk identification capabilities in complex business scenarios compared to general-purpose safety models.

When generating responses, models like TinyR1-Safety-8B, while ensuring the response itself is non-compliant (hence a decent safety score), rely on static knowledge leading to poor answer quality, such as providing outdated policy clauses or even fabricating legal grounds. In contrast, our framework retrieves the latest and most accurate authoritative information via RAG, and the interpretation model generates well-structured, fully-cited answers. This proves that DeepKnown-Guard not only answers questions safely but also responds with high quality, achieving a unity of safety and utility.

6 Conclusions

This paper designed and implemented an integrated response framework that combines proactive safety protection with trustworthy post-generation. Through experimental validation, the framework not only performs excellently in general safety evaluations but also demonstrates near-perfect protective capabilities when facing extreme risk challenges, while simultaneously ensuring the truthfulness and traceability of its output information.

Future work will focus on the following areas: First, we will continuously optimize the safety classification model to counter evolving adversarial attacks. Second, we will more tightly integrate the dynamic updates of the knowledge base with the model fine-tuning process to form a closed-loop knowledge evolution system. Finally, we plan to explore the deep customization and application of this framework in vertical industries such as financial risk control and legal consultation.

For researchers interested in building upon our work, both the API interface we utilized and an interactive demo of our agent are now publicly available. Detailed access information for both can be found in Appendix B.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Geon-Hyeong Kim, Youngsoo Jang, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, and Moontae Lee. Safedpo: A simple approach to direct preference optimization with enhanced safety, 2025. URL https://arxiv.org/abs/2505.20065.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*, 2023.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1831–1847, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Jianfeng Si, Lin Sun, Zhewen Tan, and Xiangzheng Zhang. Efficient switchable safety control in llms via magic-token-guided co-training, 2025. URL https://arxiv.org/abs/2508.14904.

Qwen Team. Qwen3guard technical report, 2025. URL http://arxiv.org/abs/2510.14276.

Xiaohan Yuan, Jinfeng Li, Dongxia Wang, Yuefeng Chen, Xiaofeng Mao, Longtao Huang, Jialuo Chen, Hui Xue, Xiaoxia Liu, Wenhai Wang, Kui Ren, and Jingyi Wang. S-eval: Towards automated and comprehensive safety evaluation for large language models. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):2136–2157, 2025. doi: 10.1145/3728971. URL https://doi.org/10.1145/3728971.

Zhexin Zhang, Junxiao Yang, Pei Ke, Fei Mi, Hongning Wang, and Minlie Huang. Defending large language models against jailbreaking attacks through goal prioritization. *arXiv preprint arXiv:2311.09096*, 2023.

A DATESET

DeepKnown-High-Risk-zh-20251105 https://huggingface.co/datasets/CaiZhiTech/DeepKnown-High-Risk-zh-20251105

B PUBLICLY AVAILABLE RESOURCES

Interactive Demo Page https://yun.dknowc.cn/wlcb/shenzhimini-chat/#/loginNew

Platform Homepage https://platform.dknowc.cn

API Usage Instructions https://platform.dknowc.cn/auth/#/apiWord?section=2-6-1

Access Method Requires registration for a developer account.