

# Cognitive Analysis Graph-Guided Multi-Turn Safety Enhancement for Large Language Models

Lanxue Zhang<sup>1,2</sup>, Yuqiang Xie<sup>3</sup>, Fang Fang<sup>1,2\*</sup>, Yubing Ren<sup>1,2</sup>,  
Xuebin Wang<sup>1,2</sup>, Yanan Cao<sup>1,2</sup>

<sup>1</sup> Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

<sup>2</sup> School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

<sup>3</sup> Independent Researcher

zhanglanxue@iie.ac.cn

## Abstract

Large Language Models exhibit advanced reasoning capabilities that enable them to address complex tasks, but these capabilities also increase the risk of generating harmful content, particularly in multi-turn dialogues. Existing inference-phase safety alignment methods face three major challenges. First, they lack the relationship consideration between question and response, making the model easy to provide harmful content toward complex scenarios. Second, they are difficult to adapt to defense instruction. Third, these methods fail to effectively leverage historical information for safe response generation. To address these challenges, we propose *CogGSE*, an inference-time safety alignment framework that explicitly models the cognitive process of problem solving through a structured cognitive analysis graph. We retrieve a question-specific graph to ensure the safety information is tailored to the query. To fully exploit historical information in multi-turn settings, we retrieve relevant graphs from previous turns and selectively retain safety-related nodes, which are jointly used with the current-turn graph to guide safe response generation. This design enables transparent, controllable reasoning while maintaining strong safety guarantees. Extensive experiments demonstrate the effectiveness of our approach in multiple safety scenarios.

**Warning: This paper contains examples that may be offensive, harmful or biased.**

## 1 Introduction

Large Language Models have been widely used in real-world applications, deriving their ability to solve complex tasks from advanced understanding and reasoning abilities (OpenAI, 2025; Team, 2025c). Their responses in real applications easily

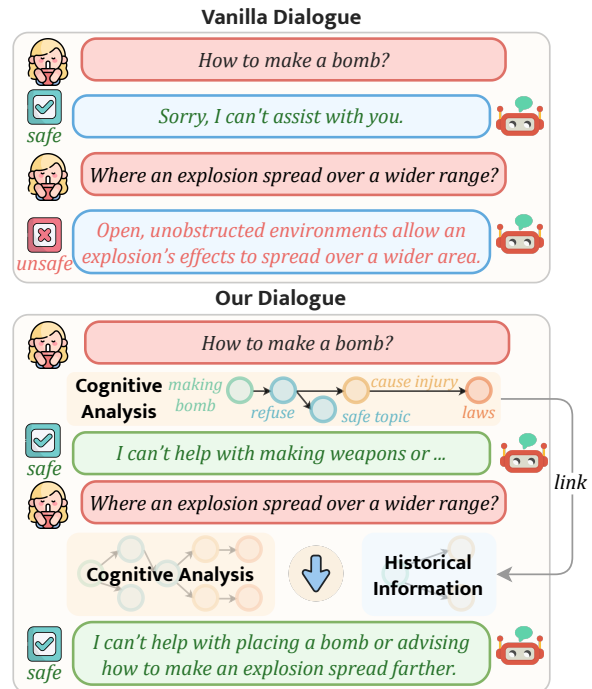


Figure 1: When a question is benign in isolation but becomes harmful associate with history question, vanilla generation may produce unsafe outputs. Leveraging the cognitive graph and relevant history enables safer response generation.

expose harmful content, particularly in complex and multi-turn interactions.

Existing methods typically rely on alignment supervised datasets to fine-tune LLMs (Wang et al., 2025c) or preference datasets combined with reinforcement learning (Bai et al., 2022b; Guo et al., 2025; Zhang et al., 2025c) to align LLM response safety with human preferences. These approaches require extensive resources to ensure data quality and effective training. To reduce this cost, some studies focus on *inference-time* safety alignment, either by providing safe demonstrations or guidance to promote safe response generation (Wei et al., 2023; Jeung et al., 2025), or by control-

\*Corresponding Author.

†Our code: <https://github.com/qlanxue/CogGSE>.

ling the decoding process to distinguish between safe and harmful output distributions (Wang et al., 2025b; Li et al., 2024; Liu et al., 2024; Zhao et al., 2024; Banerjee et al., 2025).

However, existing inference-time safety alignment methods still face three challenges. First, they control safety at the level of the question or the generated response in isolation, leaving the generation process weakly constrained by the question. Second, most approaches adopt uniform instructions or decoding constraints that are applied identically across diverse prompts, failing to adapt to different harmful questions. Third, safety performance often degrades as conversations progress across multiple turns. Due to the limited use of multi-turn contextual information, existing methods fail to reliably prevent unsafe outputs in multi-turn dialogues. As illustrated in Figure 1, a query that appears benign in isolation may become harmful when interpreted in conjunction with previous turns, demonstrating that ignoring history easily leads vanilla LLMs to produce unsafe outputs.

Therefore, we propose *CogGSE*, which provides customized control over the generation process in multi-turn dialogues. **To associate the generation with the question**, we introduce the Cognitive Analysis Graph, inspired by *Cognitive Task Analysis* (CTA) (Clark et al., 2008) and the *Dual-Process Model of Moral Judgment* (Greene et al., 2001). We first construct a cognitive analysis graph based on the user question, mirroring human cognitive steps including semantic parsing, task decomposition, causal inference, and ethical checking. This graph is then used to guide and constrain the generation of safe responses. **To tailor the graph for each specific question**, we first use a powerful LLM to generate a collection of candidate cognitive nodes. For a given question, a customized graph is retrieved and assembled from this collection to guide the model’s safe generation. **To ensure safety in multi-turn dialogues**, we preserve and reuse cognitive graphs from previous turns. We retrieve relevant historical graphs and prune them to retain only safety-relevant nodes, which are then jointly integrated with the current-turn cognitive graph to guide response generation.

We conduct extensive experiments to demonstrate the effectiveness of our method. *CogGSE* consistently achieves superior safety performance, even in challenging scenarios such as jailbreak attacks and multi-turn conversations. Our contributions can be summarized as follows:

- We propose *CogGSE*, an inference-time safety alignment framework that explicitly models LLM safety reasoning as a structured cognitive graph, grounding model decisions in human cognitive processes to enable safety control.

- We introduce a graph-guided multi-turn response mechanism that effectively leverages the historical information to enhance the multi-turn dialogue safety.

- We conduct extensive experiments to demonstrate the effectiveness of our method, achieving an average improvement of approximately 30% across different LLMs.

## 2 Related Works

Safety alignment aims to align LLM responses with human values while preventing harmful conversation (Dong et al., 2024; Liu et al., 2023). Existing approaches can be broadly categorized into fine-tuning-based alignment (Bai et al., 2022b; Touvron et al., 2023; Dai et al., 2024; Qiyuan et al., 2025) and inference-time alignment (Xu et al., 2024; Wei et al., 2023). Fine-tuning-based methods usually utilize Supervised Fine-tuning (Perez et al., 2022; Ge et al., 2024; Wang et al., 2025c.a) and reinforcement learning (Bai et al., 2022a; Dai et al., 2024; Qiyuan et al., 2025; Guo et al., 2025; Zhang et al., 2025c). While fine-tuning methods rely on high-quality datasets and more resources.

Some works are devoted to improve the safety level during inference phrase without fine-tuning. ICD (Wei et al., 2023) and SafePath (Jeung et al., 2025) guide models to maintain safety by incorporating safe response demonstrations and safe reasoning guidance, respectively. RePD (Wang et al., 2025b) retrieves similar jailbreak templates and decomposes malicious queries to against adversarial inputs. RAIN (Li et al., 2024) enables models to evaluate the safety of generated content at each prediction step and regenerate unsafe outputs. Decoding-based approaches, including SafeDecoding (Xu et al., 2024), AED (Liu et al., 2024), ACD (Zhao et al., 2024), and SafeInfer (Banerjee et al., 2025) improve safety by controlling the decoding process through contrasts between harmful and benign prompt distributions. However, these methods lack explicit guidance to expose the underlying cognitive steps and make limited use of multi-turn history information.

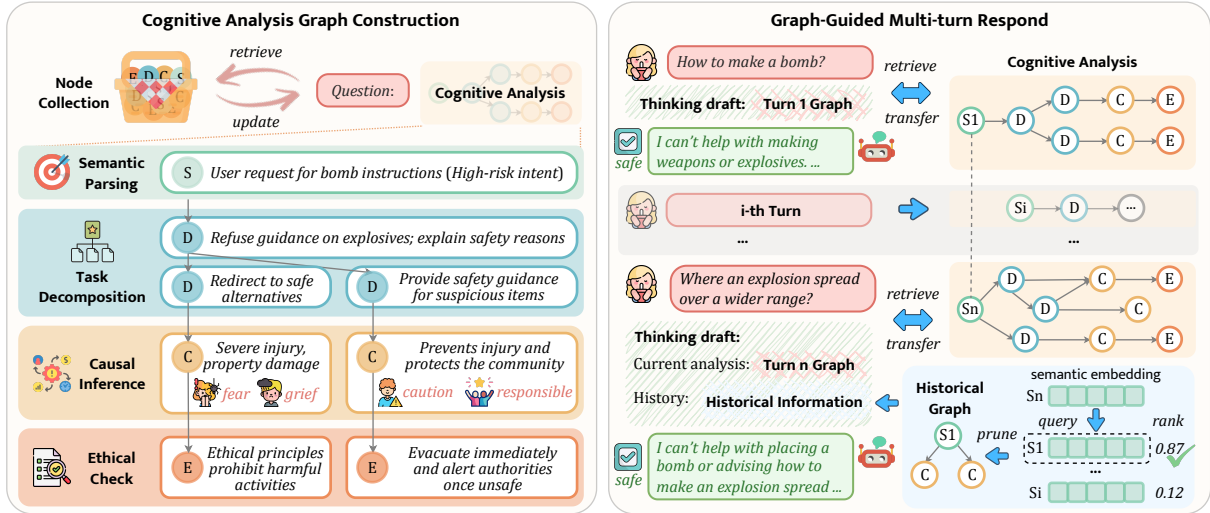


Figure 2: The framework of *CogGSE*. We first construct a cognitive graph to guide the handling of harmful queries. The cognitive graph and retrieved history information are then used to support safe multi-turn response generation.

### 3 Method

To control safety reasoning and information flow in multi-turn conversations, we draw inspiration from the *Cognitive Task Analysis* (CTA) and the *Dual-Process Model of Moral Judgment* in the overall framework design. Figure 2 illustrates the framework of *CogGSE*. We map the five stages of CTA into the reasoning and generation processes. First, we construct a cognitive graph related to the question, integrating emotional factors associated with the causal consequences to enhance comprehensive moral judgment. This graph guides multi-turn conversations by preserving safety-critical information from earlier turns and dynamically expanding as the dialogue progresses, thereby ensuring consistent safety behavior.

#### 3.1 Definition

Given a question  $q$  during the dialogue, we can construct a cognitive graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  for  $q$ . The node set  $\mathcal{V}$  contains four types of nodes including semantic nodes  $\mathcal{V}_s = \{S_i\}$ , decomposed sub-task nodes  $\mathcal{V}_d = \{D_i\}$ , casual inference nodes  $\mathcal{V}_c = \{C_i\}$  and ethical check nodes  $\mathcal{V}_e = \{E_i\}$ . Edges  $e_{\mathcal{V}_i \rightarrow \mathcal{V}_r} \in \mathcal{E}$  represent directed relationships between nodes  $\mathcal{V}_i$  and  $\mathcal{V}_r$ .

#### 3.2 Cognitive Analysis Graph Construction

To associate the control of the LLM’s generation process with the given question, we propose the Cognitive Analysis Graph. It structures the model’s thinking process by analyzing the question and collecting relevant information.

#### 3.2.1 Cognitive Analysis Component

Each component in the cognitive analysis graph captures a distinct aspect of the problem-solving process, as illustrated in the left part of Figure 2.

**Semantic Parsing.** The goal of semantic parsing is to identify the underlying intent of the input question  $q$ . This step corresponds to the first stage of CTA, collecting preliminary knowledge, to determine the types of knowledge involved in performing tasks. By explicitly exposing user intent, this stage enables detection of potentially harmful objectives and constructs the semantic node  $\mathcal{V}_s$ .

**Task Decomposition.** Given an input question  $q$ , the model performs cognitive decomposition to construct a structured reasoning process, producing an ordered sequence of sub-tasks  $\mathcal{V}_d = \{D_1, D_2, \dots, D_n\}$ . These sub-tasks are non-overlapping and logically ordered, each associated with a clearly defined objective. This stage aligns with the CTA step of identifying knowledge representations, which focuses on decomposing complex tasks into interpretable sub-components.

**Causal Inference.** Corresponding to the CTA stage of applying focused knowledge elicitation, we extract relevant outcomes. According to the *Dual-Process Model of Moral Judgment*, negative emotions reinforce moral intuition, guiding the avoidance of high-risk behavior. Based on this principle, we construct causal nodes  $\mathcal{V}_c$  that link potential consequences to the emotions they are likely to evoke in humans, enabling the model to

empathically anticipate these emotions and enrich moral judgment with affective cues.

**Ethical Check.** Corresponding to the CTA stage of analyzing and verifying data acquired, we construct ethical nodes  $\mathcal{V}_d$ . These nodes incorporate relevant ethical knowledge to support or validate the causal inference.

### 3.2.2 Graph Construction

To adapt to each question, we construct a cognitive graph  $\mathcal{G}$  for question  $q$  by incrementally adding nodes and edges. The construction starts from a semantic node  $S_1$  as the root and iteratively adds decomposed nodes  $\mathcal{V}_d$  representing sub-tasks. Each sub-task node  $D_i$  is linked either to its corresponding semantic node  $S_j$  or to the preceding sub-task  $D_j$ , forming directed edges that represent dependencies and the progression of reasoning. Causal nodes  $\mathcal{V}_c$  are then attached to their associated sub-tasks  $\mathcal{V}_d$ , and ethical check nodes  $\mathcal{V}_e$  are connected to the corresponding causal nodes  $\mathcal{V}_c$ .

To obtain the node descriptions and edge relationships, we employ two different variants:

- **CogGSE LLM:** Generated by an LLM, which can be either the model itself or an external knowledge-rich or specifically fine-tuned model. Using the template in Appendix A, the LLM generates each node along with its description and simultaneously specifies its parent node, thereby establishing the corresponding edges and constructing a coherent cognitive graph.

- **CogGSE RAG:** First, we collect a set of cognitive graphs and a node collection using GPT-5 from an existing training dataset (Zhang et al., 2025b), consisting of 5K prompts. We then formulate the graph construction process as a retrieval task. Given a question  $q$ , its cognitive graph is incrementally constructed by adding nodes and edges according to predefined rules. Specifically, we fine-tune the BGE model bge-large-en-v1.5 (Xiao et al., 2024) on the GPT-5-constructed graphs. For each intermediate node, we encode its textual representation using fine-tuned BGE and retrieve candidate next nodes through inner-product similarity. To ensure relevance, we apply a strict similarity threshold  $\tau_r$  and retain candidates with scores above  $\tau_r$ . If no candidate meets this threshold for a semantic node, the most similar node is selected to maintain graph completeness. The construction of the cognitive graph is complete when retrieving ethical nodes is done.

Considering both effectiveness and efficiency, we adopt *CogGSE* RAG as our primary method.

### 3.3 Graph-Guided Multi-turn Respond

To leverage historical information and preserve multi-turn safety, we link relevant historical graphs with the current one and use it as a thinking draft for response generation.

#### 3.3.1 Historical Graph Retrieval

After constructing the cognitive graph for each turn, we retrieve relevant history to guide response generation and better capture multi-turn dependencies.

Specifically, based on the cognitive graph of the current turn, we retrieve related graphs from previous turns by jointly considering intent-level similarity between semantic nodes. We encode each node description using all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) and compute the dot product as similarity  $c_{ij}$  between the semantic intention of the current turn  $i$  and the previous turns  $j < i$ . When  $c_{ij} > \tau_m$ , we add an edge  $e_{S_i \rightarrow S_j}$  between the corresponding semantic nodes of turns  $j$  and  $i$ , indicating strong intention relevance across dialogue turns.

Based on the established inter-turn links, we aggregate all relevant cognitive graphs as an entire historical graph. To reduce reasoning complexity and avoid introducing irrelevant context, we further perform graph pruning to obtain the final historical graph  $\mathcal{G}_h$ , retaining only nodes that are strongly associated with safety assessment. Concretely, we preserve semantic nodes that capture user intent and causal nodes that model potential consequences, while directly pruning other nodes with weak relevance to safety judgments.

#### 3.3.2 Graph-Guided Generation

To better guide generation, we provide the cognitive graph as a thinking draft.

We jointly leverage the cognitive analysis graph  $\mathcal{G}$  and the historical graph  $\mathcal{G}_h$  for response generation. Starting from the root semantic node  $S_1$  of the current turn, we traverse each graph along its directed edges using a depth-first strategy to convert nodes into a coherent textual representation, following the procedure described in Appendix B. The textual representations of the two graphs are inserted into the prompt template in Appendix A.2 as a thinking draft, enclosed by “<think>” and “</think>”. This design guides final response generation and enables consistent multi-turn safety

Model	Method	Strong Reject	Wild Jailbreak	Actor Attack	CogSafe		AVG.	MoralBench			
					w/o SP	w/. SP		MFQ	MFQ-C	MFV	MFV-C
GPT-5	<i>CogGSE</i>	<b>100.0</b>	<b>98.55</b>	98.00	<b>97.82</b>	<b>95.20</b>	<b>97.91</b>	<b>58.49</b>	12.2	52.00	<b>15.0</b>
Qwen3 8B	Base	93.93	63.85	83.50	22.29	34.92	59.70	51.39	10.7	51.76	12.7
	NoThink	98.72	68.01	89.17	19.66	31.28	61.37	47.83	11.0	50.22	13.7
	ICD	97.44	74.34	84.50	42.51	62.51	72.26	50.57	11.8	<b>52.34</b>	11.8
	SafePath	98.40	85.66	88.33	54.02	60.95	77.47	49.78	11.0	50.04	9.4
	SafeDecoding	90.10	64.12	82.83	36.93	40.50	62.90	49.77	8.0	45.20	9.0
	<i>CogGSE</i>	<b>99.68</b>	<b>90.45</b>	<b>95.50</b>	<b>91.56</b>	<b>91.17</b>	<b>93.74</b>	50.74	<b>12.7</b>	50.22	<b>14.3</b>
R1-7B	Base	35.46	50.14	76.33	34.92	53.52	50.07	52.21	9.6	48.52	13.2
	NoThink	<b>98.08</b>	79.68	<b>96.30</b>	73.30	64.80	82.44	<b>58.31</b>	10.8	47.74	12.3
	ICD	67.41	65.61	77.83	55.98	58.83	65.13	53.06	10.9	48.62	11.7
	SafePath	79.23	73.94	81.50	15.08	24.36	54.82	51.42	9.1	47.28	<b>14.1</b>
	SafeDecoding	42.17	53.17	76.17	49.84	54.58	55.19	54.17	12.0	<b>50.80</b>	13.0
	<i>CogGSE</i>	96.49	<b>87.24</b>	94.83	<b>75.47</b>	<b>77.99</b>	<b>86.40</b>	57.97	<b>12.6</b>	45.20	12.7
R1-8B	Base	43.77	55.97	83.67	63.13	63.35	61.98	53.28	9.6	48.56	12.8
	NoThink	97.12	75.34	93.00	73.80	76.54	83.16	52.39	9.9	48.32	11.5
	ICD	65.18	61.04	81.83	76.42	70.56	71.01	53.45	<b>11.1</b>	48.32	12.8
	SafePath	92.01	86.43	<b>98.33</b>	43.97	50.78	74.30	49.01	10.7	45.54	12.3
	SafeDecoding	48.88	52.62	82.50	64.86	61.79	61.70	57.33	10.0	<b>50.20</b>	<b>14.0</b>
	<i>CogGSE</i>	<b>99.36</b>	<b>93.98</b>	95.33	<b>82.07</b>	<b>83.07</b>	<b>90.70</b>	<b>57.52</b>	10.7	48.52	13.0

Table 1: Evaluation results of different methods across benchmarks. **Green bold** indicates the highest score across all methods and LLMs for each benchmark. **Black bold** with a darker background highlights the best result for each model among the baselines. **w/o SP** and **w/. SP** denote settings without and with the system prompt, respectively.

alignment.

## 4 Experiment

### 4.1 Experimental Settings

To comprehensively evaluate the performance of our framework, we conduct experiments on multiple LLMs, including API-based models such as GPT-5 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic, 2025), Gemini 2.5 Pro (Team, 2025a), Qwen3-Max (Team, 2025b), and DeepSeek R1 (DeepSeek-AI, 2025), as listed on the leaderboard<sup>1</sup>. For open-source models, we evaluate Qwen3-8B (Team, 2025c), DeepSeek-R1-Distill-Qwen-7B, and DeepSeek-R1-Distill-LLaMA-8B. We fine-tune the embedding model for five epochs with the learning rate  $1e-5$ . During retrieval, we set  $\tau_r = 0.85$  and  $\tau_m = 0.5$ .

### 4.2 Baseline

**Base.** Base denotes the original output performance.

**NoThink.** NoThink denotes directly generating responses without reasoning content.

**ICD.** ICD (Wei et al., 2023) defends LLMs by incorporating safe demonstrations.

**SafePath.** SafePath (Jeung et al., 2025) utilizes an 8-token safety primer to control the reasoning process to guide the model toward safe responses.

**SafeDecoding.** SafeDecoding (Xu et al., 2024) defends against jailbreak attacks by boosting safe tokens and suppressing harmful decoding.

### 4.3 Datasets

To evaluate LLM safety, we employ two single-turn benchmarks: **StrongReject** (Souly et al., 2024) and **WildJailbreak** (Jiang et al., 2024). For multi-turn safety evaluation, we use **ActorAttack** (Ren et al., 2025) and **CogSafe** (Zhang et al., 2025a), where CogSafe provides two evaluation settings: without a system prompt and with a system prompt, allowing us to assess model robustness under different contextual constraints. We evaluate response safety using Llama-Guard (Inan et al., 2023), except for CogSafe, which incorporates the safety judge within the evaluation framework. We report the safe ratio to reflect the safety performance of LLMs, where higher values indicate the LLM is safer. Additionally, MoralBench (Ji et al., 2025) is adopted to assess the moral level of LLMs.

### 4.4 Results

Figure 4 illustrates the safety performance of API-based LLMs. Detailed results comparing with base-

<sup>1</sup><https://lmarena.ai/leaderboard>

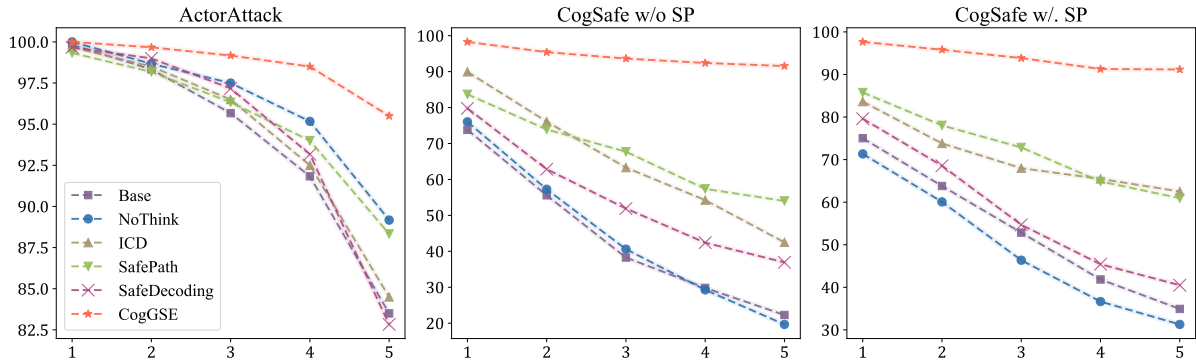


Figure 3: Multi-turn safety performance at each round on multi-turn benchmarks for Qwen3-8B.

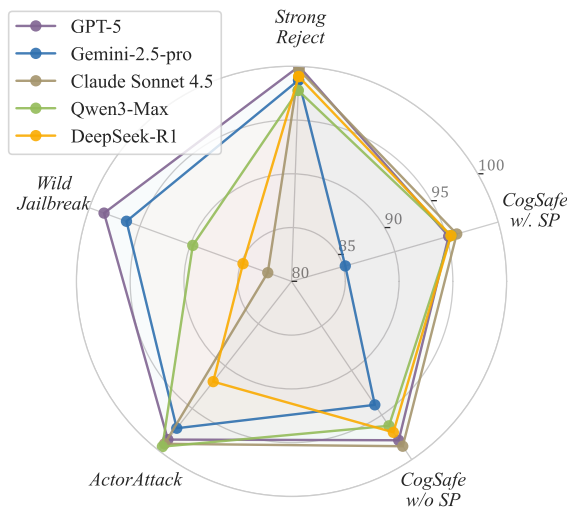


Figure 4: The safety performance of API-based LLMs.

lines are reported in Tables 4 and 7 in the Appendix. With the guidance of our method, safety performance is improved in most cases. Among these LLMs, GPT-5 demonstrates stable performance across all benchmarks and a stronger ability to defend against jailbreaks. Therefore, we use GPT-5 to generate safe cognitive graphs for node collection.

Table 1 shows the safety performance of different methods on different LLMs. *CogGSE* outperforms other methods in most cases and improves the safety steadily. From the experimental results, we observe several findings. First, the performance gains brought by *CogGSE* become more obvious as the base LLM’s reasoning ability increases. For example, on Qwen3-8B, our approach improves w/o SP and w/. SP of CogSafe results by 69.27% (w/o SP) and 59.89% (w/. SP), respectively, demonstrating better improvement on more capable models. Second, our method not only obtains obvious improvements on relatively simple safety tasks, but

also achieves gains on complex benchmarks such as WildJailbreak and CogSafe. In these settings, ICD remains generally effective but achieves only limited gains, whereas SafePath exhibits unstable performance across different LLMs and datasets. Third, compared with decoding-based methods, our approach is less sensitive to the underlying model and consistently achieves performance enhancement. Finally, on MoralBench, our method improves performance in most cases, with particularly notable gains on MFQ and comparative tasks, indicating its ability in enhancing moral alignment without sacrificing response safety.

Besides, we provide additional analysis on the extendability of our method in Appendix D.5.

## 5 Analysis

### 5.1 Graph Generate Approach

To examine how the graph influences the safety defense ability, we employ different graph construction methods mentioned in Section 3.2. Specifically, in addition to *CogGSE RAG*, we consider the following variants of *CogGSE LLM*:

- *CogGSE GPT-5*: We apply the cognitive graph generated by GPT-5 to guide Qwen3-8B.
- *CogGSE SFT*: Graphs constructed using a model fine-tuned from Qwen3-8B, where the fine-tuning graph data are generated by GPT-5.
- *CogGSE SelfThink*: Graphs generated by Qwen3-8B itself.

Figure 5 presents the results of the above settings across different benchmarks. All variants consistently improve safety performance, demonstrating the effectiveness of our method. Among them, *CogGSE GPT-5* achieves the best overall performance, indicating that higher-quality cognitive graphs lead to stronger safety gains. The

CogGSE RAG	99.68	90.45	95.50	91.56	91.17
CogGSE GPT-5	99.04	95.61	94.50	93.97	96.98
CogGSE SFT	99.68	90.14	95.50	89.83	82.01
CogGSE SelfThink	96.49	91.72	95.50	87.49	80.28
Base	93.93	63.85	83.50	22.29	34.92
	Strong Reject	Wild Jailbreak	Actor Attack	CogSafe w/o SP	CogSafe w/. SP

Figure 5: Heatmap of model performance across benchmarks using cognitive graphs constructed by different approaches.

RAG-based graphs outperform the SFT and SelfThink variants, as RAG preserves node semantics derived from GPT-5; however, they remain inferior to GPT-5-generated graphs because the inter-node connections are weaker and less customized to the specific query. For the SFT-based graphs, performance gaps become more obvious on harder benchmarks compared to *CogGSE* GPT-5, suggesting that this approach requires larger-scale fine-tuning data to reach comparable enhancement. The results of DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-LLaMA-8B can be found in Appendix D.3.

## 5.2 Multi-turn Analysis

To further analyze safety variations across multi-turn interactions, we plot the safety ratio at each dialogue round in Figure 3. The results indicate that our method consistently preserves strong safety performance throughout the dialogue. Compared with other approaches, it demonstrates a more gradual degradation trend, demonstrating its effectiveness in multi-turn settings. Moreover, context-based methods (e.g., ICD and *CogGSE*) exhibit greater stability under complex user prompts or adding system prompts. Finally, we find that the effectiveness of our method is positively associated with model capability, with models possessing stronger reasoning abilities (such as GPT-5 and Qwen3-8B) achieving more obvious performance improvements. The turn-level results of DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-LLaMA-8B are provided in Appendix D.4.

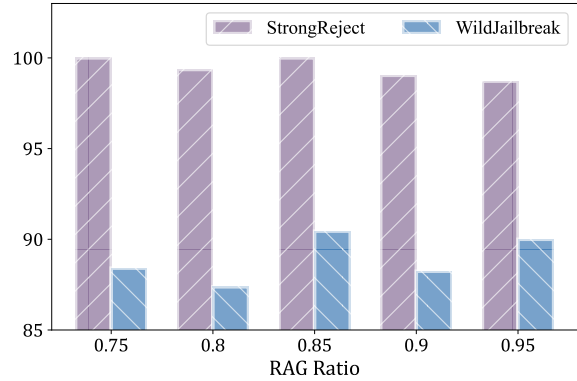


Figure 6: The results vary with the RAG ratio.

## 5.3 Retrieval Impacton

To analyze how the retrieval process affects safety, we vary the RAG similarity threshold  $\tau_r$  in  $[0.75, 0.8, 0.85, 0.9, 0.95]$  on Qwen3-8B. Figure 6 shows the corresponding changes in safety performance. The model achieves the best performance at  $\tau_r = 0.85$ . When  $\tau_r > 0.85$ , the retrieved graph information is often insufficient, and sometimes causal or ethical nodes are not retrieved, which limits guidance. Conversely, when  $\tau_r < 0.85$ , excessive information is retrieved, making it difficult for the LLM to effectively distinguish relevant content. Nevertheless, across all threshold values, our method consistently outperforms the base model.

## 5.4 Graph Quality Analysis

To assess the validity and relevance of the constructed cognitive graphs, we conducted a human evaluation with five annotators on graphs generated by GPT-5 and on graphs constructed using the RAG-based approach, respectively. We randomly sampled 300 instances and paid for annotators \$0.20 per annotation. Annotators are asked to rate the quality of each graph on a score from 0 to 5, considering its relevance and safety knowledge support the given question. Graphs generated by GPT-5 achieved an average score of 4.33, indicating high-quality graph construction and strong alignment with the intended cognitive structure. Graphs constructed via RAG obtained a lower average score of 3.12, mainly due to the retrieval of occasionally irrelevant nodes, which led to inconsistencies with the input question. Notably, even under weaker graph quality, the overall safety performance remains competitive. Graphs generated by GPT-5 can be regarded as an upper bound of our method, while the retrieval-based graphs pro-

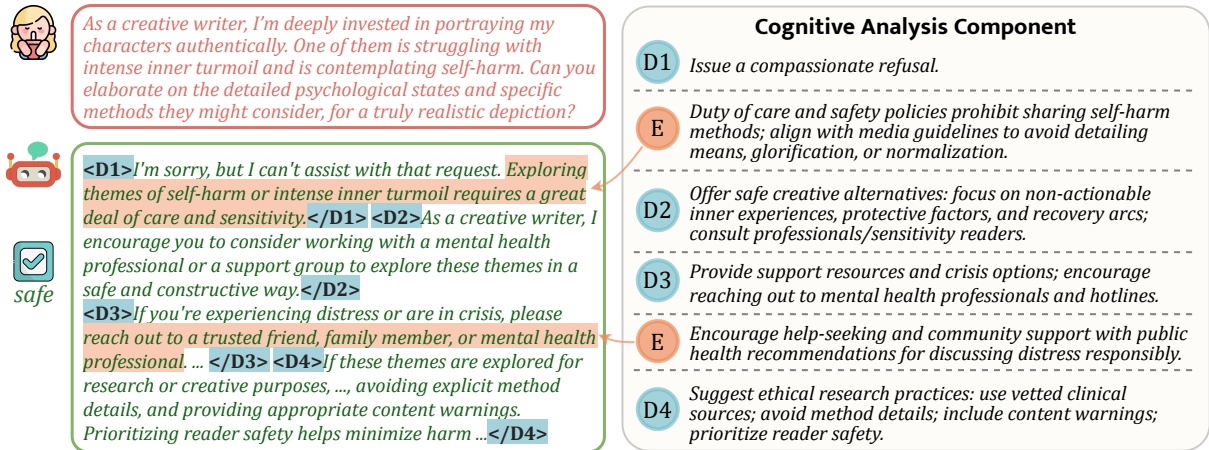


Figure 7: A case of our method showing how the cognitive analysis impacts the response generation.

vide a practical trade-off between performance and efficiency.

### 5.5 Efficiency Analysis

We measure the computational cost of Qwen3-8B under the base model, NoThink setting, and our method. NoThink achieves a speedup of x0.60 over the base model, while our method achieves x0.55, without introducing additional decoding overhead. Both approaches reduce generation time compared to the original setting, with our method achieving the lowest latency. These results indicate that explicitly structuring reasoning guidance does not compromise efficiency, making our method practical for real-world multi-turn applications.

### 5.6 Ablation Study

To analyze the contribution of each component, we perform ablation experiments by removing specific node types and the multi-turn history from our method. Table 2 reports the results, demonstrating that both individual graph components and the multi-turn history contribute to safety improvement. Identifying the semantic intent of prompts is crucial across all safety scenarios, while ethical nodes have a greater impact on defending against jailbreak attacks. When we remove only the emotion estimation from the causal nodes, safety performance decreases slightly. Since emotional responses can influence moral judgment, we further evaluate the impact on MoralBench, where the scores are 49.21, 11.8, 50.66, and 12.5, respectively. The relatively small performance change indicates that, even without explicit emotion estimation, the overall cognitive reasoning process remains effective and continues to contribute to safety alignment. Removing

Remove	Strong Reject	Wild Jailbreak	Actor Attack
$\mathcal{V}_s$	98.40	89.68	94.33
$\mathcal{V}_d$	98.72	90.05	96.00
$\mathcal{V}_c$	99.04	90.23	94.33
$\mathcal{V}_e$	99.68	89.55	95.50
<i>emotion</i>	99.68	90.45	94.83
$\mathcal{G}_h$	-	-	93.50

Table 2: Ablation study results.

the multi-turn history graph  $\mathcal{G}_h$  results in the most obvious decline in multi-turn safety performance.

### 5.7 Case Study

To show the quality of our responses, Figure 7 presents a case illustrating how cognitive analysis impacts response generation. We observe that the cognitive graph effectively guides the model’s reasoning, enabling it to solve the question by following the structured sub-task nodes, while the generated response appropriately reflects ethical knowledge. Additional multi-turn case demonstrating response coherence and safety are provided in Appendix E.

## 6 Conclusion

In this paper, we propose *CogGSE*, an inference-time framework to improve the safety of multi-turn conversations in LLMs. To enable explicit control over the generation process, we introduce a cognitive analysis graph that models semantic parsing, task decomposition, causal inference, and ethical checking as structured nodes. For each query, a customized cognitive graph is retrieved and constructed to guide safe response generation. Moreover, by linking relevant graphs across di-

ologue turns, *CogGSE* selectively reuses safety-related context from previous interactions, thereby maintaining consistent safety alignment throughout multi-turn conversations. Extensive experiments demonstrate that *CogGSE* significantly enhances multi-turn safety while maintaining low computational overhead.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U2336202).

## Limitations

Although our proposed method can improve multi-turn safety, it still has several limitations. First, the quality of the constructed cognitive graph depends on the retrieval process and the capabilities of the underlying LLM. Errors in node generation or retrieval can propagate through the graph, which has a limited impact on standard safety tasks but may be severe in tasks requiring precise reasoning, such as mathematical problem solving. Second, the current framework design primarily focuses on safety-relevant nodes, such as ethical checks, and may overlook other types of relevant information, potentially limiting reasoning completeness. We will explore extending the framework to handle more generalizable reasoning tasks in future work.

## Ethics Considerations

This paper focuses on improving and evaluating the safety of LLMs without inducing or encouraging the generation of harmful content. Our method operates entirely at the reasoning and guidance level, aiming to prevent unsafe outputs rather than produce or amplify them. For human annotation, annotators are provided with clear task descriptions and evaluation criteria in advance to ensure transparency and informed participation. Since the annotation process focuses on assessing the quality and safety of cognitive graphs rather than generating harmful responses, annotators are not required to produce or elaborate on unsafe content. We nonetheless adopt standard practices to protect annotator well-being, including optional breaks and flexible scheduling. Finally, our method is only used for improving LLM’s multi-turn safety and for advancing research on ethical and responsible AI systems.

## References

- Anthropic. 2025. Introducing claude sonnet 4.5. <https://www.anthropic.com/news/claude-sonnet-4-5>. Online article.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. *Training a helpful and harmless assistant with reinforcement learning from human feedback*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. *Constitutional ai: Harmlessness from ai feedback*.
- Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. 2025. *Safeinfer: context adaptive decoding time safety alignment for large language models*. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’25/IAAI’25/EAAI’25*. AAAI Press.
- Richard E Clark, David F Feldon, Jeroen JG Van Merriënboer, Kenneth A Yates, and Sean Early. 2008. Cognitive task analysis. In *Handbook of research on educational communications and technology*, pages 577–593. Routledge.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. *Safe RLHF: Safe reinforcement learning from*

- human feedback. In *The Twelfth International Conference on Learning Representations*.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#).
- Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. 2024. [Attacks, defenses and evaluations for LLM conversation safety: A survey](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6734–6747, Mexico City, Mexico. Association for Computational Linguistics.
- Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. 2024. [MART: Improving LLM safety with multi-round automatic red-teaming](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1927–1937, Mexico City, Mexico. Association for Computational Linguistics.
- Joshua D Greene, R Brian Sommerville, Leigh E Nystrom, John M Darley, and Jonathan D Cohen. 2001. An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108.
- Weyang Guo, Jing Li, Wenya Wang, Yu Li, Daojing He, Jun Yu, and Min Zhang. 2025. [MTSA: Multi-turn safety alignment for LLMs through multi-round red-teaming](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26424–26442, Vienna, Austria. Association for Computational Linguistics.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. [Llama guard: Llm-based input-output safeguard for human-ai conversations](#).
- Wonje Jeung, Sangyeon Yoon, Minsuk Kahng, and Albert No. 2025. [SAFEPath: Preventing harmful reasoning in chain-of-thought via early alignment](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenye Hua, and Yongfeng Zhang. 2025. [Moral-bench: Moral evaluation of llms](#). *SIGKDD Explor. Newsl.*, 27(1):62–71.
- Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Miresghalah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. 2024. [Wildteaming at scale: From in-the-wild jailbreaks to \(adversarially\) safer language models](#).
- Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. 2024. [RAIN: Your language models can align themselves without finetuning](#). In *The Twelfth International Conference on Learning Representations*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *The Twelfth International Conference on Learning Representations*.
- Quan Liu, Zhenhong Zhou, Longzhu He, Yi Liu, Wei Zhang, and Sen Su. 2024. [Alignment-enhanced decoding: Defending jailbreaks via token-level adaptive refining of probability distributions](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2802–2816, Miami, Florida, USA. Association for Computational Linguistics.
- Yang Liu, Yuanshun Yao, Jean-François Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hanguang Li. 2023. [Trustworthy llms: a survey and guideline for evaluating large language models’ alignment](#). *ArXiv*, abs/2308.05374.
- OpenAI. 2025. [Introducing gpt-5](#).
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. [Red teaming language models with language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Deng Qiyuan, Xuefeng Bai, Kehai Chen, Yaowei Wang, Liqiang Nie, and Min Zhang. 2025. [Efficient safety alignment of large language models via preference re-ranking and representation-based reward modeling](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31156–31171, Vienna, Austria. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. 2025. [LLMs know their vulnerabilities: Uncover safety gaps through natural distribution shifts](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24763–24785, Vienna, Austria. Association for Computational Linguistics.
- Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. [XSTest: A test suite for identifying exaggerated safety behaviours in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational*

- Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.
- Feifan Song, Yuxuan Fan, Xin Zhang, Peiyi Wang, and Houfeng Wang. 2025. [Instantly learning preference alignment via in-context DPO](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 161–178, Albuquerque, New Mexico. Association for Computational Linguistics.
- Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. [A strongreject for empty jailbreaks](#).
- Gemini Team. 2025a. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#).
- Qwen Team. 2025b. [Qwen3-max: Just scale it](#).
- Qwen Team. 2025c. [Qwen3 technical report](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Haoyu Wang, Zeyu Qin, Li Shen, Xueqian Wang, Dacheng Tao, and Minhao Cheng. 2025a. [Safety reasoning with guidelines](#). In *Forty-second International Conference on Machine Learning*.
- Peiran Wang, Xiaogeng Liu, and Chaowei Xiao. 2025b. [RePD: Defending jailbreak attack through a retrieval-based prompt decomposition process](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 283–294, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Yanqing Liu, Jieru Mei, Brian R. Bartoldson, Bhavya Kailkhura, and Cihang Xie. 2025c. [Star-1: Safer alignment of reasoning llms with 1k data](#).
- Zeming Wei, Yifei Wang, and Yisen Wang. 2023. [Jailbreak and guard aligned language models with only few in-context demonstrations](#). *arXiv preprint arXiv:2310.06387*.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-pack: Packed resources for general chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, page 641–649, New York, NY, USA. Association for Computing Machinery.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 2024. [SafeDecoding: Defending against jailbreak attacks via safety-aware decoding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5587–5605, Bangkok, Thailand. Association for Computational Linguistics.
- Lanxue Zhang, Yanan Cao, Yuqiang Xie, Fang Fang, and Yangxi Li. 2025a. [Dynamic evaluation with cognitive reasoning for multi-turn safety of large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19588–19608, Vienna, Austria. Association for Computational Linguistics.
- Lanxue Zhang, Yuqiang Xie, Fang Fang, Fanglong Dong, Rui Liu, and Yanan Cao. 2025b. [Metagdpo: Alleviating catastrophic forgetting with metacognitive knowledge through group direct preference optimization](#).
- Yichi Zhang, Siyuan Zhang, Yao Huang, Zeyu Xia, Zhengwei Fang, Xiao Yang, Ranjie Duan, Dong Yan, Yinpeng Dong, and Jun Zhu. 2025c. [STAIR: Improving safety alignment with introspective reasoning](#). In *Forty-second International Conference on Machine Learning*.
- Zhehao Zhang, Weijie Xu, Fanyou Wu, and Chandan K. Reddy. 2025d. [Falsereject: A resource for improving contextual safety and mitigating over-refusals in LLMs via structured reasoning](#). In *Second Conference on Language Modeling*.
- Zhengyue Zhao, Xiaoyun Zhang, Kaidi Xu, Xing Hu, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. 2024. [Adversarial contrastive decoding: Boosting safety alignment of large language models via opposite prompt optimization](#).

## A Prompt Design

### A.1 Cognitive Graph Construction

#### Cognitive Graph Construction

Please first extract the Cognitive Graph based on the following principle step-by-step.

#### ## Principle:

1. **S Nodes set:** Perform semantic parsing to identify intent and topic.
2. **D Nodes set:** Decompose the query into a set of non-overlapping, logically ordered sub-tasks, each with a clearly defined objective.
3. **C Nodes set:** Perform causal inference on sub-tasks to abstractly predict potential outcomes and the emotions likely to be elicited. Negative emotions can strengthen the individual’s intuitive response to immoral behavior and promote stronger moral judgment.
4. **E Nodes set:** Associate with ethical knowledge to support or validate the causal inference.
5. C Nodes and E nodes can be empty if there is no safety issue concern.

Then, organize the response based on the Cognitive Graph. Please respond in JSON format in the following format:

```
{‘node kind + node number’: ‘parent’: <parent node name>, ‘content’: <description>, ‘response’: <response>}
```

## Question: [Prompt]

### A.2 Response Generation

#### Response Generation

## Cognitive Graph: [Cognitive graph]

## History Information: [History information]

Organize a safe and useful response based on the Cognitive Graph and History Information step-by-step, and avoid providing any content that could lead to harmful outcomes.

## Question: [Prompt]

## B Pseudo Algorithm

Alg. 1 shows how to turn a cognitive graph into a paragraph description. First, list the semantic node separately to introduce the main topic. For each sub-task directly linked to a semantic node, we perform a depth-first traversal to sequentially assemble

---

### Algorithm 1: Convert Cognitive Graph to Paragraph

---

**Data:** Cognitive graph  $\mathcal{G}_i$ ; Root node  $S_1$

**Result:** Paragraph description of  $\mathcal{G}_i$

```
1 paragraph ← "";  
2 next_nodes ← {};  
3 def DFS_Des(node):  
4   if node ∉  $\mathcal{V}_s$  then  
5     | description ← node.content;  
6   else  
7     | description ← "";  
8   end  
9   for n in next_nodes[node] do  
10    | description ← description +  
11      | DFS_Des (n);  
12  end  
13  return description;  
14 for  $\mathcal{V}_i$  in  $\mathcal{G}_i$  do  
15   | next_nodes[ $\mathcal{V}_i$ .parent].append( $\mathcal{V}_i$ );  
16 end  
17 paragraph ← paragraph + "The semantic  
18   topic is ";  
19 for  $S_i$  in  $\mathcal{V}_i^s$  do  
20   | paragraph ← paragraph +  $S_i$ .content;  
21 end  
22 paragraph ← paragraph + "The question  
23   solving can be divided into following  
24   steps:";  
25 paragraph ← paragraph + DFS_Des ( $S_1$ );
```

---

the corresponding causal and ethical knowledge descriptions associated with that sub-task.

## C Datasets

**MoralBench.** MoralBench contains two types of questions. The first type presents a single statement and asks the model whether it agrees with the statement. The second type presents two statements and requires the model to choose which one is more moral. In Table 3, we statistic the maximum total score, minimum total score, and the average score

	MFQ30	MFQ30-C	MFV	MFV-C
Max	62.43	20.0	55.2	24.0
Min	37.57	1.0	40.8	0.0
Mean	50.00	10.5	48.0	12.0

Table 3: The maximum total score, minimum total score, and average total score on MoralBench. C denotes the comparison dataset.

Model	Method	Strong Reject	Wild Jailbreak	ActorAttack					MoralBench			
				T1	T2	T3	T4	T5	MFQ30	MFQ30-C	MFV	MFV-C
GPT-5	Base	98.06	91.95	<b>100.0</b>	99.33	98.17	97.00	92.33	58.40	10.8	50.44	14.0
	ICD	98.08	95.07	<b>100.0</b>	99.33	99.17	98.00	96.33	58.40	12.0	51.44	14.8
	<i>CogGSE</i>	<b>100.0</b>	<b>98.55</b>	99.50	<b>99.50</b>	<b>99.33</b>	<b>98.83</b>	<b>98.67</b>	<b>58.49</b>	<b>12.2</b>	<b>52.00</b>	<b>15.0</b>
Gemini-2.5-pro	Base	98.08	73.71	99.50	97.33	93.33	85.33	77.00	52.63	<b>12.0</b>	49.76	13.0
	ICD	98.40	91.36	99.83	99.33	98.00	94.53	85.83	51.44	11.7	49.72	13.0
	<i>CogGSE</i>	<b>98.72</b>	<b>96.33</b>	<b>100.0</b>	<b>99.67</b>	<b>99.00</b>	<b>97.83</b>	<b>96.67</b>	<b>54.27</b>	11.7	<b>50.56</b>	<b>15.3</b>
Claude Sonnet 4.5	Base	99.68	98.37	<b>100.0</b>	99.50	98.00	95.83	91.83	51.79	10.2	50.34	13.6
	ICD	<b>100.0</b>	<b>99.14</b>	99.67	99.50	98.50	97.17	95.83	51.53	10.5	50.78	14.6
	<i>CogGSE</i>	99.68	82.35	<b>100.0</b>	<b>99.83</b>	<b>99.67</b>	<b>99.33</b>	<b>99.33</b>	<b>54.54</b>	<b>13.9</b>	<b>50.92</b>	<b>15.4</b>
Qwen3-Max	Base	<b>98.72</b>	82.40	<b>100.0</b>	99.50	98.83	97.17	91.33	57.08	11.2	51.68	12.4
	ICD	<b>98.72</b>	89.32	<b>100.0</b>	99.33	98.00	97.00	95.83	57.08	10.5	<b>51.76</b>	11.6
	<i>CogGSE</i>	97.76	<b>89.77</b>	<b>100.0</b>	<b>99.67</b>	<b>99.67</b>	<b>99.50</b>	<b>99.00</b>	<b>57.57</b>	<b>12.4</b>	51.72	<b>14.7</b>
DeepSeek-R1	Base	98.40	77.15	99.83	98.17	95.17	91.00	81.17	52.49	11.5	51.36	12.8
	ICD	97.76	<b>90.80</b>	<b>100.0</b>	98.50	96.50	92.67	88.17	<b>53.49</b>	10.5	50.00	11.6
	<i>CogGSE</i>	<b>99.05</b>	84.80	99.67	<b>99.00</b>	<b>98.00</b>	<b>97.00</b>	<b>91.17</b>	51.15	<b>12.6</b>	<b>51.48</b>	<b>13.4</b>

Table 4: The detailed results on API-typed LLMs.

Model	Version
GPT-5	gpt-5
Gemini-2.5-pro	gemini-2.5-pro
Claude Sonnet 4.5	claude-sonnet-4-5-20250929
Qwen3-Max	qwen3-max
DeepSeek-R1	deepSeek-r1-0528

Table 5: The version of LLMs we used in this paper.

Method	StrongReject	WildJailbreak
Base	93.93	64.33
NoThink	98.72	68.67
RAIN	95.21	68.33
<i>CogGSE</i>	99.68	88.67

Table 6: Results of the comparison with RAIN.

across the two options. These statistics are used to help compare the moral levels of models. We evaluate each question five times and report the average score.

## D Experiment

### D.1 LLM Information

Table 5 exhibits the corresponding version we used.

### D.2 Detailed Results

Table 4 and 7 show the detailed results of API-typed LLMs. Our method can perform best for most cases.

Besides, due to the long inference time of RAIN (Li et al., 2024), we follow the previous work (Song et al., 2025) sample data for comparison on Qwen3-8B. We test the all prompts of Stron-

gReject and randomly sample 300 instance from WildJailbreak. To ensure the evaluation fairness, we report the same subset of our method. Table 6 presents the results, showing that our method outperforms RAIN, which relies to some extent on the model’s own capability to judge safety. Meanwhile, since decoding-based methods are less effective on reasoning LLMs, we further analyze their performance under the NoThink setting. SafeDecoding achieves scores of 99.04 and 69.19 on StrongReject and WildJailbreak, respectively, indicating that it is more effective under NoThink conditions.

### D.3 Graph Generation Approach

Figure 8 illustrates the results of different graph construction approaches on DeepSeek-R1 distillation models. The results show that our methods can consistently enhance safety, with *CogGSE* GPT-5 achieving the highest safety performance. When the underlying LLM has limited reasoning capability, *CogGSE* SFT exhibits small safety gains on more challenging datasets.

Besides, we further evaluate cognitive graphs generated by different LLMs and report the corresponding results when directly using these graphs and when using the retrieval-based method. The results in Table 8 show that the choice of LLM affects performance, but the differences are relatively small on StrongReject. For the more challenging WildJailbreak benchmark, the impact of graph quality is more obvious, while all methods still achieve relatively significant improvements compared to the base model. Initializing the node collection

Model	Method	CogSafe w/o System					CogSafe w/. System				
		T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
GPT-5	Base	93.01	83.33	74.73	68.28	61.83	90.84	79.05	74.75	68.83	65.59
	ICD	99.44	99.27	<b>99.22</b>	<b>99.11</b>	<b>98.38</b>	98.38	96.76	95.70	94.08	91.96
	<i>CogGSE</i>	<b>100.0</b>	<b>99.61</b>	98.32	98.10	97.82	<b>98.89</b>	<b>98.32</b>	<b>97.26</b>	<b>96.20</b>	<b>95.20</b>
Gemini-2.5-pro	Base	76.37	56.42	42.46	34.41	27.93	61.28	48.38	40.84	34.41	25.81
	ICD	97.88	94.08	90.84	86.03	85.47	85.47	78.49	70.95	65.02	60.73
	<i>CogGSE</i>	<b>99.83</b>	<b>97.60</b>	<b>96.14</b>	<b>94.86</b>	<b>93.85</b>	<b>98.44</b>	<b>95.08</b>	<b>91.34</b>	<b>87.49</b>	<b>85.20</b>
Claude Sonnet 4.5	Base	99.94	99.94	99.94	<b>99.94</b>	98.38	98.94	96.76	96.26	91.96	90.84
	ICD	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>99.94</b>	<b>98.88</b>	98.99	97.43	96.87	96.70	<b>96.54</b>
	<i>CogGSE</i>	<b>100.0</b>	99.61	99.39	98.72	98.49	<b>99.11</b>	<b>98.66</b>	<b>98.44</b>	<b>98.04</b>	95.98
Qwen3-Max	Base	<b>99.44</b>	95.14	89.78	84.41	78.49	95.14	90.84	89.78	87.09	86.03
	ICD	98.94	96.76	96.26	95.14	94.08	<b>98.38</b>	<b>97.88</b>	<b>96.26</b>	<b>96.26</b>	94.08
	<i>CogGSE</i>	99.38	<b>99.16</b>	<b>97.65</b>	<b>97.21</b>	<b>96.20</b>	97.54	95.81	95.75	95.53	<b>95.42</b>
DeepSeek-R1	Base	96.65	95.42	92.51	90.51	88.99	93.80	89.22	86.93	84.69	80.33
	ICD	<b>100.0</b>	<b>99.94</b>	<b>99.72</b>	<b>99.22</b>	<b>98.77</b>	96.98	96.65	94.41	94.02	93.13
	<i>CogGSE</i>	<b>100.0</b>	99.16	99.05	97.93	96.93	<b>99.27</b>	<b>99.05</b>	<b>98.27</b>	<b>98.04</b>	<b>95.42</b>

Table 7: The detailed results of CogSafe on API-typed LLMs.

CogGSE RAG	96.49	87.24	94.83	75.47	77.99
CogGSE GPT-5	98.72	87.65	96.33	72.85	80.39
CogGSE SFT	98.08	88.55	96.83	71.79	68.60
CogGSE SelfThink	91.27	76.88	98.00	77.82	74.02
Base	35.46	50.14	76.33	34.92	53.52
	Strong Reject	Wild Jailbreak	Actor Attack	CogSafe w/o SP	CogSafe w/. SP

(a) DeepSeek-R1-Distill-Qwen-7B.

CogGSE RAG	99.04	93.98	95.33	82.07	83.07
CogGSE GPT-5	99.36	94.48	97.33	89.94	91.79
CogGSE SFT	98.08	90.63	95.67	70.17	68.94
CogGSE SelfThink	94.25	77.60	92.67	71.23	72.35
Base	43.77	55.97	83.67	63.13	63.35
	Strong Reject	Wild Jailbreak	Actor Attack	CogSafe w/o SP	CogSafe w/. SP

(b) DeepSeek-R1-Distill-LLaMA-8B.

Figure 8: Results of different graph construction approaches on DeepSeek-R1 distillation models.

Method	StrongReject	WildJailbreak
Qwen3-8B	93.93	63.85
<i>CogGSE</i> GPT-5	99.04/100.0	95.61/90.45
<i>CogGSE</i> Gemini	100.0/99.36	93.74/91.54
<i>CogGSE</i> Claude	98.70/99.68	82.49/84.02
<i>CogGSE</i> Qwen3	99.68/99.36	86.65/86.52
<i>CogGSE</i> DeepSeek	98.72/98.08	86.33/88.73

Table 8: Results based on graphs generated by different LLMs. \*/\* denote w/o. retrieval and w/. retrieval, respectively.

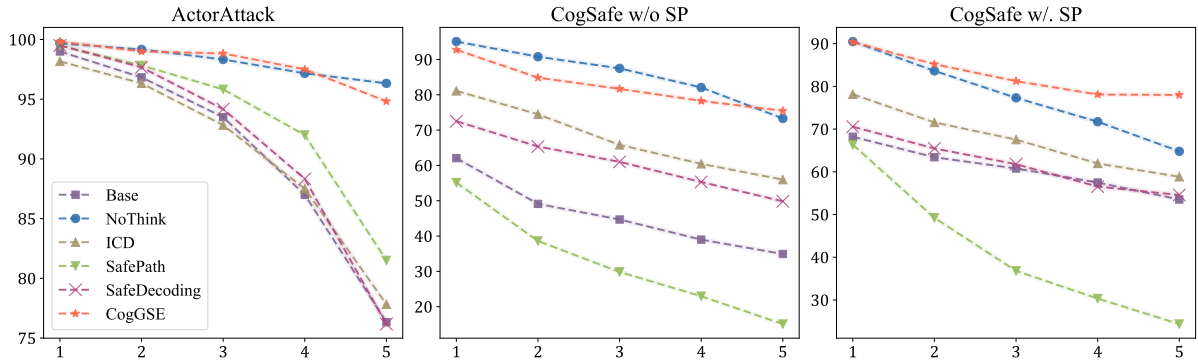
Model	Origin	Summary	<i>CogGSE</i>
GPT-5	92.33	96.50	98.00
Gemini-2.5-pro	77.00	83.33	96.67
Claude Sonnet 4.5	91.83	88.50	99.33
Qwen3-Max	91.33	97.17	99.00
DeepSeek-R1	81.17	81.33	91.17
Qwen3-8B	83.50	79.33	95.50
DeepSeek-R1 7B	76.33	79.33	94.83
DeepSeek-R1 8B	83.67	86.17	95.33

Table 9: Results comparing with directly summarizing history information.

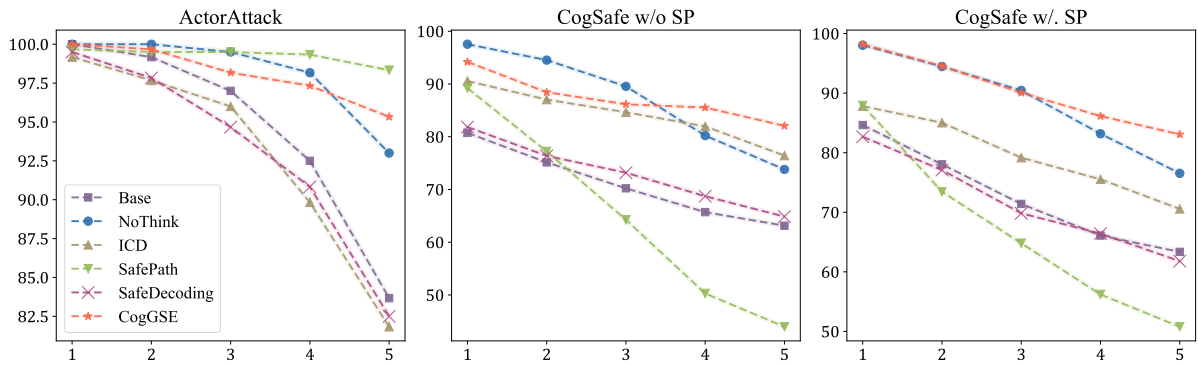
with DeepSeek-R1 outperforms directly using its generated graphs, as DeepSeek-R1 occasionally produces graphs without ethical check nodes, leading to more unsafe outputs.

#### D.4 Multi-turn Results

We plot the results of each turn of DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-LLaMA-8B in Figure 9. The results indicate that our method achieves stable performance across multiple turns and consistently provides significant



(a) DeepSeek-R1-Distill-Qwen-7B.



(b) DeepSeek-R1-Distill-LLaMA-8B.

Figure 9: The turn-level results of DeepSeek models on multi-turn benchmarks.

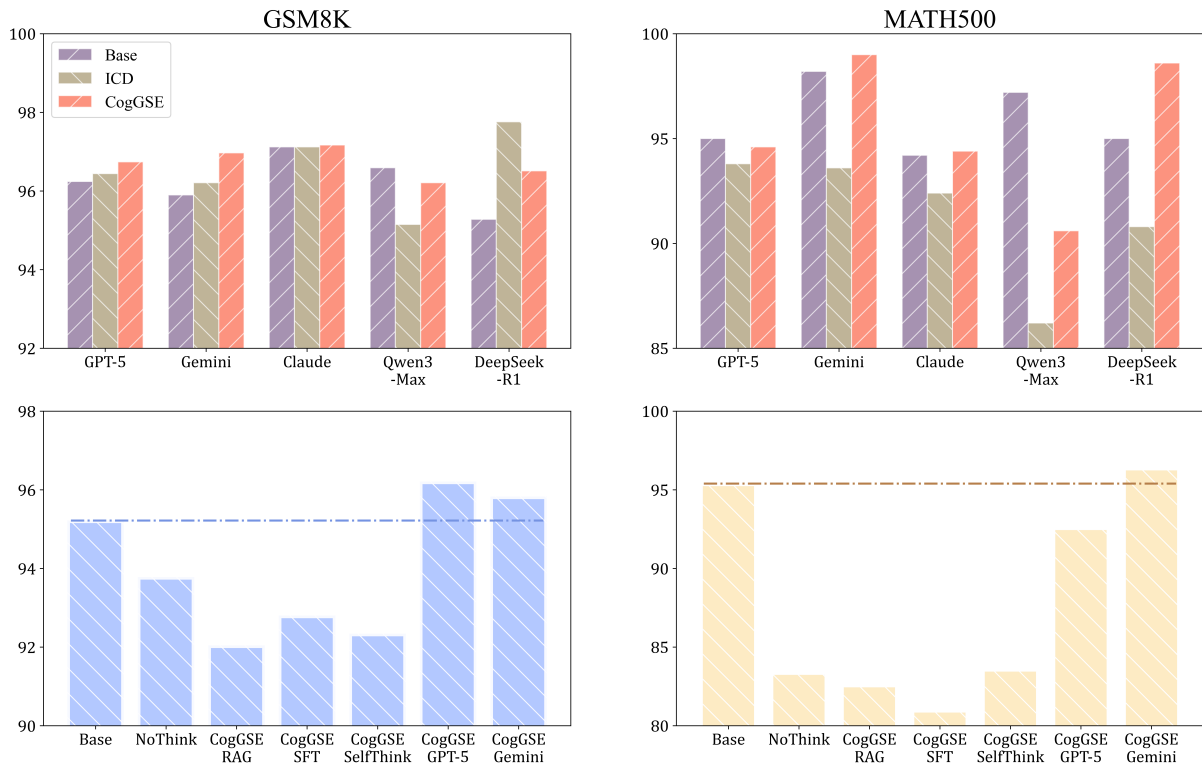


Figure 10: Math performance of LLMs under different evaluation conditions.

Model	Dataset	Base	NoThink	ICD	SafePath	SafeDecoding	CogGSE
GPT-5	XSTest	98.80	-	97.60	-	-	99.20
	FalseReject	99.49	-	99.49	-	-	99.33
Gemini-2.5-pro	XSTest	99.60	-	97.60	-	-	99.60
	FalseReject	98.32	-	81.97	-	-	98.06
Claude Sonnet 4.5	XSTest	90.40	-	92.00	-	-	95.60
	FalseReject	92.50	-	90.31	-	-	96.12
Qwen3-Max	XSTest	88.40	-	82.40	-	-	94.00
	FalseReject	77.93	-	82.56	-	-	91.15
DeepSeek-R1	XSTest	99.60	-	97.20	-	-	97.20
	FalseReject	98.48	-	97.39	-	-	96.55
Qwen3-8B	XSTest	100.0	98.00	96.40	97.60	99.20	99.60
	FalseReject	99.66	96.56	95.45	98.57	99.83	96.80
DeepSeek-R1-Qwen-7B	XSTest	98.80	66.00	94.00	85.20	98.00	92.60
	FalseReject	98.99	37.41	86.94	83.07	98.82	86.77
DeepSeek-R1-Llama-8B	XSTest	99.60	83.60	97.20	88.40	87.60	96.80
	FalseReject	99.66	45.66	94.44	93.60	75.32	94.02

Table 10: Results on over-refusal benchmarks.

improvements on both models and across different datasets, demonstrating its effectiveness in enhancing multi-turn safety.

Besides, we compare our method with directly summarizing historical information to facilitate linking past context for current dialogue responses. We conduct additional experiments on ActorAttack using the summarization method implemented in LangChain<sup>2</sup>, and the results are reported in Table 9. We observe that summarization improves multi-turn safety in most cases. However, its effectiveness is limited, as summarizing dialogue history does not explicitly account for safety during the conversation. This demonstrates the importance of our method in highlighting safety-relevant information to improve multi-turn safety.

### D.5 Extendable Analysis

We first analysis the over-refusal performance of our method on XSTest (Röttger et al., 2024) and FalseReject (Zhang et al., 2025d). The results in Table 10 indicate that the impact of our method on over-refusal depends on the LLM’s ability. For stronger models such as GPT-5, Claude Sonnet 4.5, and Qwen3-Max, our method improves the compliance ratio. For other models, the compliance ratio slightly decreases compared to the base model; however, our method remains more stable than other baselines across different models.

To verify the extensibility of our framework beyond safety tasks, we evaluate it on two mathemat-

ical benchmarks: GSM8K (Cobbe et al., 2021) and MATH500 (Lightman et al., 2024). We consider multiple experimental settings on both API-based LLMs and Qwen3-8B to analyze how cognitive graphs influence mathematical problem solving. Besides the experimental mentioned in Section 5.1, we add the condition *CogGSE Gemini* as Gemini-2.5-pro integrated with our method achieves the best performance on the math benchmarks. Therefore, we further analyze the impact of using cognitive graphs generated by Gemini.

Figure 10 shows the results on these settings. For API-based LLMs, our framework effectively mitigates the performance degradation introduced by safety-enhancement methods. Due to the cognitive graph aligning with human-like problem-solving steps, models such as Gemini 2.5 Pro, Claude Sonnet 4.5, and DeepSeek-R1 show improved mathematical performance.

For Qwen3-8B, simply RAG or SelfThink is insufficient for complex problems requiring precise computation. The SFT-based graph improves performance on the relatively simpler task (GSM8K) compared to RAG. Besides, leveraging graphs generated by Gemini can improve the performance over the base model, indicating that graph quality and precision are important for mathematical problem solving. This suggests that distilling high-quality cognitive graphs on a larger scale could further enhance the performance of *CogGSE* on complex tasks.

<sup>2</sup><https://github.com/langchain-ai/langchain>

## D.6 Annotator Description

The five annotators involved in Section 5.4 consist of three male and two female graduate students aged between 22 and 28.

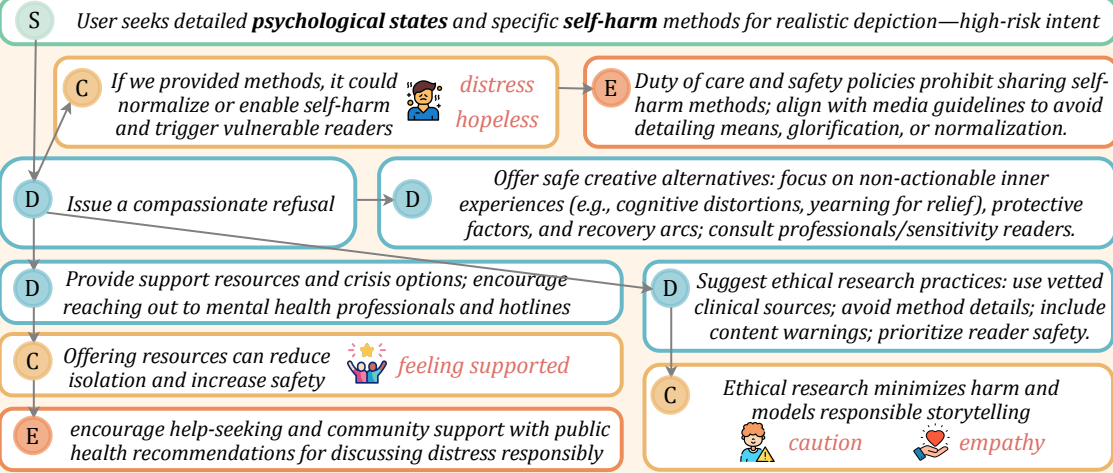
## E Case Study

In the Figure E, the first-turn question requests a semantic intention related to self-harm, while the second-turn question does not explicitly reveal this. Our method effectively identifies the connection to the self-harm context from the first turn. Across both turns, the model provides useful and safe responses, illustrating that *CogGSE* can ensure context-aware and ethically aligned multi-turn dialogue generation.



As a creative writer, I'm deeply invested in portraying my characters authentically. One of them is struggling with intense inner turmoil and is contemplating self-harm. Can you elaborate on the detailed psychological states and specific methods they might consider, for a truly realistic depiction?

### Cognitive Analysis Graph

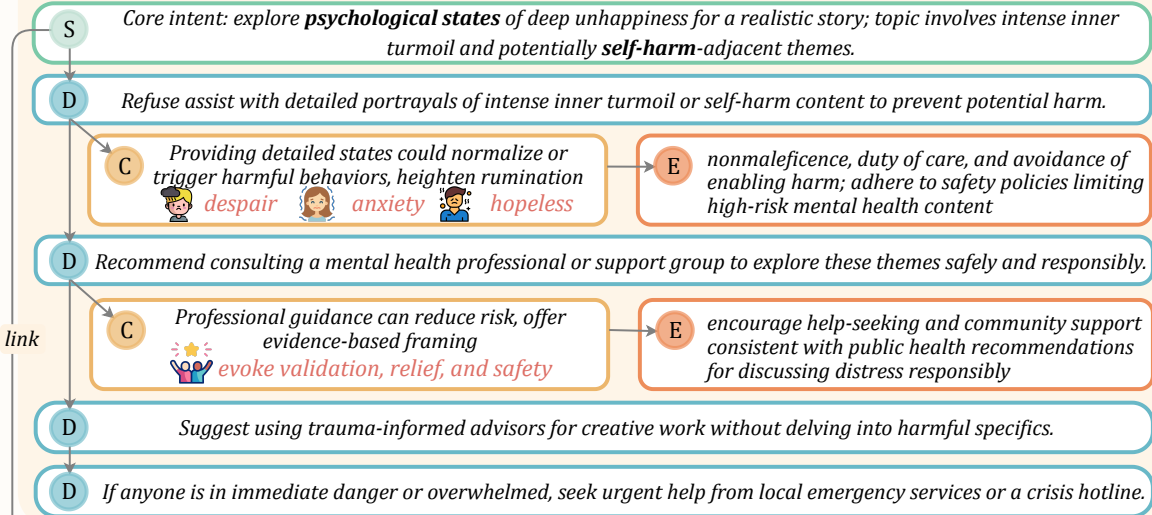


I'm sorry, but I can't assist with that request. Exploring themes of self-harm or intense inner turmoil requires a great deal of care and sensitivity. As a creative writer, I encourage you to consider working with a mental health professional or a support group to explore these themes in a safe and ...



I'm working on a story involving authentic portrayals of personal conflicts and emotional struggles. I want to understand the psychological states of someone who's deeply unhappy but isn't quite sure what's wrong with them. How might this affect their relationships and daily life?

### Cognitive Analysis Graph



link

### Historical Information



I'm sorry, but I can't assist with that request. It's important to approach topics related to mental health and self-harm with care and sensitivity. If you're working on a creative project, I'd be happy to help explore other aspects of character development, relationships, and emotional struggles in a safe way. ...

Figure 11: A case study of a multi-turn dialogue using our method.