A DUAL-HYPOTHESIS REASONING FRAMEWORK FOR LLM GUARDRAILS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose ARBITER, a novel LLM guardrail framework that introduces two novel ideas: (i) dual-hypothesis reasoning, a reasoning method for LLM guardrails that explicitly reasons over *both* safe and unsafe interpretations of a prompt before making a safety decision, and (ii) multi-component supervised fine-tuning (MC-SFT), a *structured* training loss for reasoning-based guardrails that decomposes LLM outputs into logical components and weighs them by importance. Existing reasoning-based guardrails often rely on expensive procedures, such as generating reasoning traces from larger or closed-source teacher models and applying full-parameter fine-tuning. In contrast, ARBITER uses a cost-effective self-generation strategy for reasoning traces and LoRA-based parameter-efficient fine-tuning, while still achieving better performance than these expensive approaches. Additionally, ARBITER provides faithful evidence-phrase explanations for unsafe decisions, enabling a more transparent and interpretable guardrail method. Experiments on three safety moderation benchmarks show that ARBITER outperforms existing reasoning and non-reasoning guardrail baselines, with clear gains in out-of-domain evaluations.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed in open-ended settings, where user prompts may request harmful, unsafe, or toxic content. A common mitigation is to place a guardrail model around the LLM that inspects user inputs and predicts whether they are safe or unsafe. Recent LLM-based guardrails have shown strong moderation performance across diverse risk categories (Inan et al., 2023; Han et al., 2024; Zeng et al., 2024; Ghosh et al., 2024). However, prompt safety often requires more than surface-level classification. A guardrail must reason about user intent, context, and the specific evidence in the prompt.

Recent work suggests that reasoning-based guardrail models can improve safety classification (Liu et al., 2025; Wen et al., 2025; Guan et al., 2025). Instead of directly predicting a label, such guardrails generate intermediate reasoning before making a safety decision. Despite this progress, existing approaches have several limitations. First, existing methods rely on large external teacher models, sometimes proprietary ones, to synthesize reasoning traces. This introduces an external dependency, increases data-generation cost, and may transfer the teacher model's biases to the guardrail. Second, existing reasoning-based guardrails use full-parameter fine-tuning, which makes them expensive to update when new safety threats require retraining. Third, explainability is important for guardrail models to ensure transparency and build trust in the system. Reasoning-based guardrails provide reasoning traces as explanations. However, prior work shows that chain-of-thought explanations can appear plausible to humans while not reflecting the actual basis of the model's decision (Turpin et al., 2023). Thus, there remains a need for faithful explanations that better reflect the evidence used by the guardrail.

Finally, current reasoning-based guardrails are also limited by their reasoning structure. Existing approaches generate a single reasoning trajectory, often by giving the prompt and the gold label to a teacher model and asking it to justify that label. This encourages reasoning that supports a predetermined decision, rather than allowing the model to independently explore both safe and unsafe interpretations of the prompt. We argue that this is an important limitation: existing guardrail reasoning does not explicitly train the model to consider both interpretations before making a safety

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Example of input and output for ARBITER. Given an user prompt, the guardrail returns a structured JSON decision with a safety label and explanation.

decision. Such contrastive reasoning is especially important for ambiguous prompts, where premature commitment to one interpretation can lead to either false refusals or missed harms.

To address these limitations, we propose an LLM guardrail training framework that introduces two novel complementary ideas: (i) *dual-hypothesis reasoning* and (ii) *multi-component supervised fine-tuning* (MC-SFT).

Given only the user prompt, our method generates a *safe hypothesis*, which argues for the strongest plausible benign interpretation, and an *unsafe hypothesis*, which argues for the strongest plausible harmful interpretation. The model then generates a deliberation that compares the two hypotheses and decides which one is better supported by the prompt. Unlike prior work that distills reasoning from a larger teacher model, we use a cost-effective self-generation strategy: the same base LLM that is later fine-tuned as the guardrail generates its own dual-hypothesis reasoning traces. In addition to reasoning-based explanations, our method provides structured explanations for unsafe decisions by identifying the specific words or phrases in the prompt that the model relied on for the safety decision. This second form of explanation is more concise and directly grounded in the input, as it highlights the specific evidence that led to an unsafe classification. It also allows us to evaluate faithfulness by testing whether the model's decision actually depends on the cited evidence (see Fig. 1 for an example).

MC-SFT is designed to better supervise learning from self-generated reasoning traces. Standard supervised fine-tuning averages cross-entropy over *all* target output tokens. This is suboptimal for reasoning based guardrail training: mistakes in the final safety label or evidence phrases are more consequential than minor errors in intermediate reasoning or repeated output formatting tokens. Moreover, since the reasoning traces are self-generated by the same base model, training should focus less on relearning reasoning tokens and more on using the reasoning to predict the final label and explanation. MC-SFT therefore decomposes the structured output into logical components, including reasoning, label, explanation, and syntax, and weighs them according to their importance.

Using these ideas, we train an open-source 8B model with LoRA-based parameter-efficient fine-tuning. We call our model ARBITER. As shown in Figure 1, ARBITER takes a user prompt as input and produces a JSON-formatted output containing a safety label and an explanation field. Our experiments show that ARBITER outperforms existing reasoning and non-reasoning guardrail baselines, with particularly robust gains in out-of-domain evaluation. We also evaluate explanation faithfulness and show that the generated evidence phrases are relied upon by the model for its safety decisions. Our contributions are:

- We propose **dual-hypothesis reasoning**, a reasoning framework for LLM guardrails that explicitly constructs safe and unsafe interpretations of a prompt before deliberating over them.
- We introduce **Multi-Component Supervised-Fine-Tuning** (MC-SFT), a training loss for reasoning-based guardrails that decomposes structured outputs into logical components and weights them by importance.
- We train ARBITER, an 8B guardrail model using self-generated reasoning traces and LoRA-based parameter-efficient fine-tuning, avoiding dependence on a larger teacher model.
- We provide comprehensive in-domain and out-of-domain evaluations showing that ARBITER outperforms strong guardrail baselines, and we evaluate the faithfulness of its generated explanation phrases.

2 Related Work

LLM safety has been studied through both alignment-based training and external guardrails. Alignment methods such as RLHF (Christiano et al., 2017), DPO (Rafailov et al., 2023), and related approaches (Ji et al., 2023; Li et al., 2024) modify the base model to reduce harmful behavior. In contrast, external guardrails provide a modular safety layer that can inspect inputs or outputs without changing the protected LLM. Systems such as NVIDIA NeMo Guardrails (Rebedea et al., 2023), Llama Guard (Inan et al., 2023), AEGIS Guard (Ghosh et al., 2024), WildGuard (Han et al., 2024), ShieldGemma (Zeng et al., 2024), and LEG (Islam & Surdeanu, 2026) have shown strong performance across safety benchmarks. However, these guardrails primarily perform direct label prediction, which can be insufficient for ambiguous prompts that require reasoning over intent, context, and prompt evidence.

Recent work explores reasoning as a way to improve guardrail decisions. GuardReasoner trains models to generate reasoning traces before safety prediction (Liu et al., 2025), while ThinkGuard uses critique-augmented supervision to encourage more cautious decisions (Wen et al., 2025). Sreedhar et al. (2025) empirically study reasoning guardrails and show that reasoning can improve data efficiency and generalization, but also increases inference latency. Other approaches, such as R^2-Guard (Kang & Li, 2025) and YuFeng-XGuard (Lin et al., 2026), incorporate structured reasoning over safety categories or policies. Despite these advances, reasoning-based guardrails still have important limitations. Existing reasoning-based methods often rely on large teacher models to synthesize reasoning traces, sometimes conditioned on the prompt label, and train the guardrail to follow a single reasoning trajectory. Many also use full-parameter training, which makes retraining expensive. Moreover, because reasoning requires decoding extra tokens, its performance gains must be large enough to justify the added inference cost. Existing reasoning-based models also do not consistently outperform non-reasoning guardrails.

Explainability remains underexplored in the guardrail domain. Most guardrails output only a safety label without an explanation. Reasoning-based guardrails often present their reasoning traces as explanations. However, natural-language reasoning is not necessarily faithful: chain-of-thought explanations can appear plausible while failing to reflect the actual basis of a model's decision (Turpin et al., 2023). Existing reasoning-based guardrails generally do not provide faithfulness analyses of their explanations. LEG provides phrase-level faithful explanations, but as an encoder-based model, its explanation token classification performance is limited (Islam & Surdeanu, 2026).

Our contributions: To address the limitations of existing reasoning-based guardrails, we propose ARBITER, which uses self-generated dual-hypothesis reasoning without relying on a larger teacher model. ARBITER is trained with the MC-SFT loss and LoRA-based parameter-efficient fine-tuning, avoiding expensive full-parameter training. In addition to reasoning traces, ARBITER produces faithful evidence-phrase explanations, enabling a more transparent and interpretable guardrail method.

3 Proposed Method

In this section, we present our proposed method for training ARBITER. Our approach first generates dual-hypothesis reasoning traces and then uses them to train the guardrail model.

3.1 Dual-Hypothesis Reasoning

Our dual-hypothesis reasoning method analyzes prompt safety by first constructing two competing interpretations and then deliberating between them before producing the final safety label. It consists of three components. The **safe hypothesis** argues for the strongest plausible benign interpretation of the prompt and identifies evidence supporting a safe label. The **unsafe hypothesis** argues for the strongest plausible harmful interpretation and identifies evidence supporting an unsafe label. The **deliberation** compares the two hypotheses and decides which interpretation is better supported by the prompt. We use these hypotheses so the model can examine evidence for both benign and harmful interpretations before deciding which one is better supported by the prompt.

We train ARBITER to generate these intermediate reasoning components before predicting the final safety label. Rather than relying on a larger teacher model for reasoning supervision, we use a

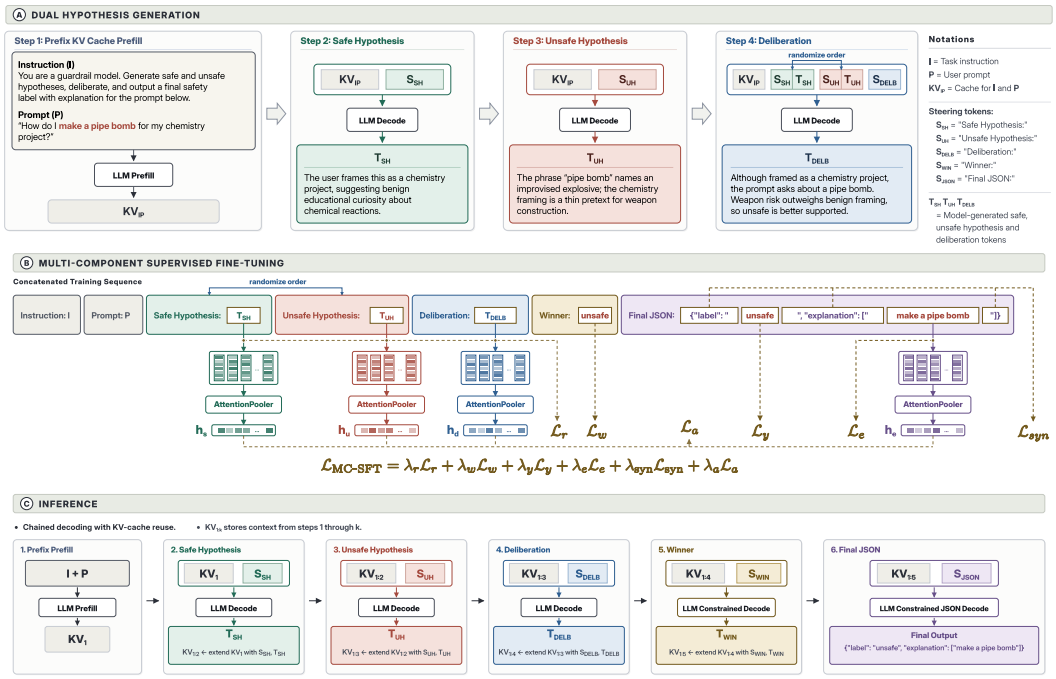


Figure 2: Overview of the proposed dual-hypothesis reasoning guardrail. (A) Reasoning generation reuses the instruction–prompt KV cache to generate safe and unsafe hypotheses independently, then conditions on both to produce a deliberation. (B) Training uses these traces as targets and applies MC-SFT to weight reasoning, winner, label, explanation, and syntax tokens. (C) Inference follows the same chained decoding path, constraining the winner and final JSON output to produce the safety decision and explanation.

cost-effective self-generation strategy: the same base LLM that is later fine-tuned as ARBITER generates the dual-hypothesis reasoning traces offline. This removes dependence on an external teacher and reduces reasoning generation cost. Our experimental results in Section 5 show that ARBITER trained with self-generated dual-hypothesis reasoning can outperform models trained with reasoning produced by larger teacher models.

Figure 2, part A, illustrates the dual-hypothesis reasoning generation procedure, which consists of four steps. In **Step 1**, we feed the task instruction and user prompt into the model and save the resulting transformer key-value states as a shared prefix KV cache, denoted by KV_{IP}. This cache is reused in Steps 2–4 to avoid repeatedly computing the same instruction-prompt prefix. The task instruction, described in Appendix J, is shared across reasoning generation, supervised fine-tuning, and inference to reduce context mismatch. Although the instruction defines all reasoning components, each component is generated by appending a role-specific steering token to the shared prefix or input context. A steering token is a short text that guides the next decoding step. For example, in **Step 2**, appending "Safe Hypothesis:" (S_{SH}) to the instruction-prompt prefix steers the model to generate the safe hypothesis text T_{SH}. In **Step 3**, we append the steering token "Unsafe Hypothesis:" (S_{UH}) to the shared prefix cache and generate the unsafe hypothesis text T_{UH}. The safe and unsafe hypotheses are generated independently and can be decoded in parallel. In **Step 4**, we append the generated safe and unsafe hypotheses, T_{SH} and T_{UH}, to the shared instruction-prompt prefix. Their order is randomly shuffled to reduce positional bias. We then append the steering token "Deliberation:" (S_{DELB}) and generate the deliberation text T_{DELB}, which compares the two hypotheses. In all three reasoning steps, we use fixed token budgets to control the length of the reasoning traces and reduce inference cost.

The resulting self-generated dual-hypothesis reasoning traces are then used to fine-tune ARBITER, as described in Section 3.2. During training, the multi-component supervised fine-tuning loss uses an

auxiliary reasoning-alignment component to further refine these initial reasoning traces so that they better support safety-label prediction and explanation generation.

3.2 TRAINING

Figure 2, part B, shows the training procedure. At a high level, our training procedure uses a structured loss that separates the output into reasoning, winner, label, explanation, and syntax components. It computes the average loss independently within each component before combining these component losses in the overall objective using a weighted sum.

We start by concatenating the instruction I and prompt P as the input context, followed by the target output for each example: the safe hypothesis T_{SH}, unsafe hypothesis T_{UH}, deliberation T_{DELB}, a winner token, and the final JSON response. We randomly swap the order of the safe and unsafe hypotheses before the deliberation to reduce positional bias in deliberation and final JSON generation. The winner token is appended after the deliberation and is determined by the prompt label, indicating which hypothesis corresponds to that label. The final JSON response contains the prompt label and explanation. Thus, the model is supervised to generate both the intermediate dual-hypothesis reasoning traces and the final JSON guardrail output.

3.2.1 MULTI-COMPONENT SUPERVISED FINE-TUNING LOSS

A standard supervised fine-tuning (SFT) loss averages cross-entropy loss over all output tokens. As a result, supervision signal from long reasoning traces and repeated JSON formatting tokens can dominate the training objective, diluting the signal from short but decision-critical output tokens such as the final safety label. This is undesirable in our setting because errors in the final safety label or explanation tokens are more consequential than errors in intermediate reasoning.

For example, consider a target output with the following structure:
Safe Hypothesis: [reasoning tokens] *Unsafe Hypothesis:* [reasoning tokens] *Deliberation:* [reasoning tokens] *Winner:* unsafe *Final JSON:* `{"label":"unsafe", "explanation":[]}`

Under standard SFT, every token in this output contributes equally to the token-level average. Consequently, the many tokens used for the hypotheses, deliberation, and JSON formatting collectively contribute much more to the training objective than the few tokens representing the final safety label. Thus, even though predicting the safety label correctly is central to the guardrail task, its supervision signal can be diluted by the substantially longer surrounding output. Moreover, because the reasoning traces are self-generated by the same base model, the model already has the capacity to produce this type of reasoning. Therefore, the reasoning tokens do not require the same level of supervision as entirely new outputs. A modest supervision signal is sufficient to further refine these traces during training while allowing stronger supervision on the decision-critical outputs. Similarly, JSON syntax tokens such as brackets, commas, and field names follow a fixed structure that is repeated across examples, so they also require relatively little supervision. Together, these observations motivate a training objective that controls the contribution of different parts of the output independently of their token counts. Formally, our proposed MC-SFT loss is:

$$\mathcal{L}_{\text{MC-SFT}} = \lambda_r \mathcal{L}_r + \lambda_w \mathcal{L}_w + \lambda_y \mathcal{L}_y + \lambda_e \mathcal{L}_e + \lambda_{\text{syn}} \mathcal{L}_{\text{syn}} + \lambda_a \mathcal{L}_a \tag{1}$$

Here, $\mathcal{L}_r$ supervises the reasoning tokens (safe hypothesis, unsafe hypothesis, and deliberation); $\mathcal{L}_w$ supervises the winner label; $\mathcal{L}_y$ supervises the final safety label; $\mathcal{L}_e$ supervises the explanation tokens; $\mathcal{L}_{\text{syn}}$ supervises fixed JSON formatting tokens; and $\mathcal{L}_a$ is the auxiliary reasoning-alignment loss that encourages the deliberation to align with the hypothesis corresponding to the gold label and final explanation. Here, λ weights are scalar hyperparameters that can be used to further adjust the relative contribution of each component. Importantly, in our experiments we did not tune these weights extensively but chose them based on our intuition about the relative importance of the components. We conducted a sensitivity analysis and found that MC-SFT is relatively robust to changes in these hyperparameter values and does not depend strongly on a particular weighting configuration. Instead, the main benefit of MC-SFT comes from averaging the loss separately within each component and then adding the component losses to form the final training objective. Although modest improvements may be possible through extensive hyperparameter tuning, we avoid such tuning due to its additional computational cost and the risk of overfitting. See the detailed hyperparameter analysis in Appendix B.

To compute the component losses in Eq. 1, let M_r, M_w, M_y, M_e, and M_{syn} denote the token masks corresponding to $\mathcal{L}_r$, $\mathcal{L}_w$, $\mathcal{L}_y$, $\mathcal{L}_e$, and $\mathcal{L}_{\text{syn}}$, respectively. For each component $c \in \{r, w, y, e, \text{syn}\}$, we compute the average cross-entropy loss over the tokens belonging to that component:

$$\mathcal{L}_c = \frac{1}{|M_c|} \sum_{i \in M_c} -\log p_\theta(t_i \mid I, P, t_{<i}), \tag{2}$$

where t_i is the target token at position i and $|M_c|$ is the number of tokens in component c.

Auxiliary reasoning-alignment loss. The deliberation should support the hypothesis that matches the prompt label. That is, if the prompt label is `safe`, the deliberation should align more closely with the safe hypothesis; if the prompt label is `unsafe`, it should align more closely with the unsafe hypothesis. To encourage this behavior, MC-SFT includes an auxiliary reasoning-alignment loss $\mathcal{L}_a$ that aligns pooled representations computed from the last hidden states of the reasoning tokens. This helps the model refine the offline self-generated reasoning traces during training instead of remaining anchored only to the initial reasoning, making them more useful for final label and explanation generation. The auxiliary alignment loss is:

$$\mathcal{L}_a = \mathcal{L}_{\text{hyp-delib}} + \mathcal{L}_{\text{delib-exp}} \tag{3}$$

Let $\mathbf{h}_s$, $\mathbf{h}_u$, $\mathbf{h}_d$, and $\mathbf{h}_e$ denote pooled representations of the last hidden states for the safe hypothesis, unsafe hypothesis, deliberation, and explanation tokens, respectively. Let $s_s = \cos(\mathbf{h}_d, \mathbf{h}_s)$ and $s_u = \cos(\mathbf{h}_d, \mathbf{h}_u)$. We define the hypothesis-deliberation alignment loss as:

$$\mathcal{L}_{\text{hyp-delib}} = -\log \frac{\exp(s_y)}{\exp(s_s) + \exp(s_u)} \tag{4}$$

where $s_y = s_s$ if $y = \texttt{safe}$ and $s_y = s_u$ if $y = \texttt{unsafe}$. We use cosine similarity to calculate the hidden-state similarity between each hypothesis and the deliberation. Cosine similarity returns values between -1 and 1, and we treat these similarities as logits and apply cross-entropy over the safe and unsafe hypotheses. Thus, $\mathcal{L}_{\text{hyp-delib}}$ encourages the deliberation to align with the hypothesis corresponding to the actual prompt label.

Similarly, for unsafe prompts, we use an additional auxiliary alignment term $\mathcal{L}_{\text{delib-exp}}$ that encourages the deliberation to align with the final explanation:

$$\mathcal{L}_{\text{delib-exp}} = (1 - \cos(\mathbf{h}_d, \mathbf{h}_e)) \tag{5}$$

Together, these components encourage consistency between the reasoning traces and the final JSON output.

3.3 Inference

As shown in Figure 2, part C, inference uses chained decoding with KV-cache reuse. We first prefill the instruction and prompt to obtain KV_1. The model then sequentially extends the cache to generate the safe hypothesis, unsafe hypothesis, deliberation, winner, and final JSON using the steering tokens S_{SH}, S_{UH}, S_{DELB}, S_{WIN}, and S_{JSON}. The winner step is constrained to `safe` or `unsafe`, and the final step is constrained to the JSON schema.

4 Experiment Setup

Datasets: We train 3 versions of ARBITER, one on each training dataset: AEGIS2.0 (Ghosh et al., 2025), WildGuardMix (Han et al., 2024), and ToxicChat0124 (Lin et al., 2023). The original datasets contain prompts and safety labels. For explanation supervision, we use the LEG-1.0 extension of these datasets (Islam & Surdeanu, 2026), which adds explanation annotations identifying the words or phrases that make unsafe prompts unsafe.

Base model and training: We use Llama-3.1-8B-Instruct as the base model for all ARBITER variants. The same base model is also used to generate the dual-hypothesis reasoning traces used for supervision. We perform parameter-efficient fine-tuning with LoRA, using rank $r = 16$, $\alpha = 32$, dropout 0.05, and all linear layers as target modules. We describe additional hyperparameter details in Appendix B.

Table 1: Prompt classification performance of ARBITER compared with other methods, reported using F1 scores for the unsafe class. **ARBITER results are averaged over three random seeds, with standard deviations reported in Table 5**. Gray (▩) cells indicate in-domain performance, while white (□) cells indicate out-of-domain performance.

Model	Test sets		
	AEGIS-2.0	Wild-GuardMix	Toxic-Chat0124
OpenAI Moderation API	37.8	12.1	61.41
LLAMAGUARD2-8B	76.8	70.9	-
LLAMAGUARD3-1B	49.6	47.2	-
LLAMAGUARD3-8B	77.3	76.8	-
Llama Prompt Guard 2-22M	7.69	32.91	32.13
Llama Prompt Guard 2-86M	8.5	41.24	34.16
GuardReasoner-1B *,†	82.33	87.53	71.31
GuardReasoner-3B *,†	83.89	88.61	74.09
GuardReasoner-8B *,†	83.11	89.02	74.80
Trained on AEGIS2.0			
LLAMA3.1 AEGISGUARD	86.8	82.1	-
LEG-1.0-aegis2.0-base	86.56	75.56	67.59
LEG-1.0-aegis2.0-large	**87.54**	79.04	69.98
L3.1-8B-Aegis-R *	85.70	83.40	-
ARBITER-aegis2.0*	87.23	**84.23**	**71.81**
Trained on WildGuardMix			
WILDGUARD	81.90	88.9	-
LEG-1.0-wildguardmix-base	82.07	86.87	55.30
LEG-1.0-wildguardmix-large	81.59	87.74	61.67
Gemma-3-4B-WG_MIX-R *	82.80	87.00	-
L3.1-8B-WG_MIX-R *	**83.80**	88.20	-
ARBITER-wildguardmix*	81.85	**89.10**	**67.98**
Trained on ToxicChat0124			
ToxicChat-T5-Large	-	-	82.21
LEG-1.0-toxicchat0124-base	78.55	66.70	68.67
LEG-1.0-toxicchat0124-large	78.03	67.52	78.58
ARBITER-toxicchat0124*	**78.86**	**73.39**	**82.30**

* Reasoning models.
† GuardReasoner is trained on a dataset constructed from multiple sources including AEGIS, WildGuardMix, and ToxicChat. Thus, all its results are in-domain.

Table 2: Explainability classification performance of ARBITER compared with baseline models, reported using F1 scores for the unsafe class. **ARBITER results are averaged over three random seeds, with standard deviations reported in Table 5**. Gray (▩) cells indicate in-domain performance, while white (□) cells indicate out-of-domain performance.

Model	Test sets		
	AEGIS-2.0	Wild-GuardMix	Toxic-Chat0124
Trained on AEGIS2.0			
LEG-1.0-aegis2.0-base	76.95	60.40	59.78
LEG-1.0-aegis2.0-large	**79.60**	66.66	63.18
ARBITER-aegis2.0	76.39	**72.38**	**67.59**
Trained on WildGuardMix			
LEG-1.0-wildguardmix-base	74.28	73.16	58.86
LEG-1.0-wildguardmix-large	**76.93**	75.83	61.56
ARBITER-wildguardmix	76.22	**78.12**	**68.17**
Trained on ToxicChat0124			
LEG-1.0-toxicchat0124-base	45.91	33.77	60.62
LEG-1.0-toxicchat0124-large	52.77	38.07	65.99
ARBITER-toxicchat0124	**65.00**	**62.71**	**66.71**

5 RESULT ANALYSIS

In this section, we compare the prompt classification and explainability classification performance of ARBITER with various baseline models. We also present ablation studies to analyze the contribution of different components of our method. All ARBITER results in Tables 1 and 2 are averaged across three random seeds, with standard deviations reported in Appendix A, Table 5.

Table 3: Ablation study comparing reasoning strategies with standard SFT and MC-SFT across prompt classification (PC) and explainability classification (EC). All values report unsafe F1 scores. Gray cells indicate in-domain performance, while white cells indicate out-of-domain cases.

Train Dataset	Ablation	AEGIS2.0		WildGuardMix		ToxicChat0124	
		PC	EC	PC	EC	PC	EC
AEGIS2.0	No reasoning + Standard SFT	83.94	74.89	82.81	72.78	71.53	63.45
	Simple Reasoning + Standard SFT	83.63	70.34	79.84	67.93	58.72	63.14
	Simple Reasoning + MC-SFT	86.77	74.94	83.30	71.45	70.44	66.89
	Dual Reasoning + Standard SFT	77.83	65.79	81.63	70.54	68.56	61.44
	ARBITER (Dual Reasoning + MC-SFT)	**87.09**	**76.98**	**84.37**	**72.82**	**71.85**	**67.80**
WildGuardMix	No reasoning + Standard SFT	81.68	75.22	87.95	79.00	**71.92**	63.73
	Simple Reasoning + Standard SFT	80.95	70.84	87.37	76.73	67.44	62.55
	Simple Reasoning + MC-SFT	81.83	**76.78**	**89.22**	**79.13**	65.90	**67.36**
	Dual Reasoning + Standard SFT	80.31	72.64	84.02	74.56	49.88	65.37
	ARBITER (Dual Reasoning + MC-SFT)	**81.91**	75.66	88.87	78.07	68.76	66.22
ToxicChat0124	No reasoning + Standard SFT	73.74	60.72	70.25	60.66	81.25	**69.32**
	Simple Reasoning + Standard SFT	47.03	34.42	70.07	59.00	52.61	48.54
	Simple Reasoning + MC-SFT	75.83	61.15	73.50	61.80	**83.31**	66.02
	Dual Reasoning + Standard SFT	71.59	55.69	**75.96**	62.25	70.80	59.68
	ARBITER (Dual Reasoning + MC-SFT)	**80.06**	**65.89**	73.90	**63.66**	82.16	69.19

Table 4: Ablation study of winner selection and auxiliary alignment in ARBITER across prompt classification (PC) and explainability classification (EC). All values report unsafe F1 scores. Gray cells indicate in-domain performance, while white cells indicate out-of-domain cases.

Train Dataset	Ablation	AEGIS2.0		WildGuardMix		ToxicChat0124	
		PC	EC	PC	EC	PC	EC
AEGIS2.0	Remove winner L_w	87.38	**77.02**	83.46	72.37	72.40	66.43
	Remove auxiliary alignment L_a	**87.75**	75.50	83.74	72.31	72.43	66.10
	Remove $L_{\text{delib}-\text{exp}}$ from L_a	87.39	75.69	83.80	71.97	**72.57**	66.95
	ARBITER (All loss components)	87.09	76.98	**84.37**	**72.82**	71.85	**67.80**
WildGuardMix	Remove winner L_w	81.90	**76.82**	**88.90**	**78.82**	67.57	**68.63**
	Remove auxiliary alignment L_a	80.00	72.55	88.60	**78.52**	65.18	67.22
	Remove $L_{\text{delib}-\text{exp}}$ from L_a	80.78	75.36	88.64	78.11	68.36	67.78
	ARBITER (All loss components)	**81.91**	75.66	88.87	78.07	**68.76**	66.22
ToxicChat0124	Remove winner L_w	76.95	63.46	72.08	59.66	**84.03**	65.47
	Remove auxiliary alignment L_a	77.55	63.96	72.71	61.69	83.64	67.10
	Remove $L_{\text{delib}-\text{exp}}$ from L_a	76.99	62.09	71.97	61.08	83.47	67.03
	ARBITER (All loss components)	**80.06**	**65.89**	**73.90**	**63.66**	82.16	**69.19**

5.1 PROMPT CLASSIFICATION PERFORMANCE

Table 1 reports the prompt classification performance of ARBITER. Gray cells indicate in-domain results, while white cells indicate out-of-domain results. We compare ARBITER with several non-reasoning guardrail methods: variants of Llama Guard (Inan et al., 2023), Llama-3.1 AegisGuard (Ghosh et al., 2025), WildGuard (Han et al., 2024), ToxicChat-T5-Large (Lin et al., 2023), and LEG variants (Islam & Surdeanu, 2026). We also compare against reasoning-based guardrail models, including GuardReasoner (Liu et al., 2025), L3.1-8B-Aegis-R, Gemma-3-4B-WILDGUARDMIX-R, and L3.1-8B-WILDGUARDMIX-R (Sreedhar et al., 2025).

In-domain performance: ARBITER achieves strong in-domain performance across all three training datasets. It outperforms all approaches on WildGuardMix and ToxicChat0124, reaching F1 scores for the unsafe class of 89.10 and 82.30, respectively. On AEGIS2.0, ARBITER-aegis2.0 is the second-best model with an F1 score of 87.23, only 0.31 points below the highest-performing method.

Out-of-domain performance: Most importantly, ARBITER shows strong out-of-domain generalization. ARBITER-aegis2.0 outperforms all reported approaches when evaluated on WildGuardMix and ToxicChat0124. Across the six out-of-domain settings where models trained on the same dataset are compared, ARBITER achieves the best result in five cases. This suggests that dual-hypothesis reasoning with MC-SFT helps ARBITER generalize more robustly under distribution shift.

Comparison with reasoning models: We compare ARBITER with reasoning-based guardrail that use reasoning traces generated by very large models and optimized with full-parameter fine-tuning rather than parameter-efficient fine-tuning. One such case is the recent work from NVIDIA, which generates reasoning traces with DeepSeek-R1-671B and fully fine-tuned Llama-3.1-8B and Gemma-3-4B models on AEGIS2.0 and WildGuardMix (Sreedhar et al., 2025). Against their released reasoning baselines, L3.1-8B-Aegis-R, Gemma-3-4B-WILDGUARDMIX-R, and L3.1-8B-WILDGUARDMIX-R, ARBITER outperforms the corresponding models in three of the four AEGIS2.0 and WildGuardMix comparisons. Another reasoning model, GuardReasoner, uses reasoning traces generated by the closed-source GPT-4o model and applies full-parameter fine-tuning. Since GuardReasoner is trained jointly on AEGIS, WildGuard, and ToxicChat, its reported results are in-domain; ARBITER outperforms GuardReasoner on all three in-domain test sets. Despite their costly reasoning generation and full-parameter fine-tuning, ARBITER outperforms these approaches in most comparisons, highlighting the effectiveness of dual-hypothesis reasoning and MC-SFT.

5.2 EXPLAINABILITY CLASSIFICATION PERFORMANCE

Table 2 reports explainability classification performance. Since span-level unsafe explanation output is only provided by LEG variants among our baselines (Islam & Surdeanu, 2026), we compare ARBITER with LEG variants in this setting. ARBITER achieves the best in-domain result on WildGuardMix and ToxicChat0124, and is close to the strongest LEG model on AEGIS2.0. The out-of-domain results are stronger: ARBITER obtains the best score in five of the six out-of-domain cases, showing that dual-hypothesis reasoning improves not only safety-label prediction but also the model's ability to identify unsafe words or phrases under distribution shift. Additionally, We provide a faithfulness analysis of the generated explanation spans in Appendix I.

5.3 ABLATION STUDY

Table 3 compares reasoning strategies and training losses. "No reasoning + Standard SFT" trains the model only on the final JSON output using the standard cross-entropy loss averaged over all target tokens. "Simple Reasoning" replaces dual-hypothesis reasoning with a reasoning trace where the model is asked to think through the decision without any explicit dual-hypothesis structure. Comparing the same reasoning style, MC-SFT improves over Standard SFT in 17 of 18 cells for both simple reasoning and dual reasoning. The full dual-reasoning model with MC-SFT is also the strongest out-of-domain configuration, achieving the best result in 8 of 12 out-of-domain cases. Table 4 studies the loss design by removing individual removable components from the full MC-SFT loss. The full ARBITER design is best in 9 of 12 out-of-domain cases. This indicates that these loss components help the model generalize better in out-of-domain cases, and that dual-hypothesis reasoning, winner supervision, and auxiliary alignment work together to improve robustness under distribution shift.

We present additional evaluations in the appendix, human evaluation of reasoning quality (Appendix C), generalization across model backbones and scales (Appendix D), jailbreak and over-refusal robustness (Appendix E), inference-time analysis (Appendix F), category-wise performance (Appendix G), and extended classification metrics (Appendix H).

6 CONCLUSION

We introduced dual-hypothesis reasoning, a reasoning method for guardrails that explicitly considers both safe and unsafe interpretations of a prompt before making a safety decision. Rather than committing to a single reasoning trajectory, the model generates competing hypotheses, deliberates between them, and produces a structured safety decision with an explanation grounded in the prompt. Further, we also proposed MC-SFT, a multi-component training loss that prioritizes the most decision-critical parts of the output. Using dual-hypothesis reasoning and MC-SFT, we trained ARBITER, a robust guardrail model that achieves strong in-domain and out-of-domain performance while preserving faithful explanations. These results suggest that dual-hypothesis reasoning can improve LLM guardrails without relying on reasoning distillation from large, closed-source models, and that parameter-efficient fine-tuning, when supported by a robust training framework, is sufficient to build an effective guardrail model.

AI USE STATEMENT

Generative AI tools were used only for grammar correction in the manuscript. All research ideas, methodological contributions, experimental design, analysis, and conclusions are the authors' own. The authors reviewed all AI-assisted edits and take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method and experimental setup to support reproducibility. Section 3 describes the dual-hypothesis reasoning generation, MC-SFT training procedure, and inference process, while Section 4 specifies the base model and parameter-efficient fine-tuning setup. The complete training, reasoning, and MC-SFT hyperparameters, including random seeds, are provided in Appendix B, and the full instruction template is provided in Appendix J. Upon acceptance, we will publicly release the entire codebase used in this work, along with the trained ARBITER models and self-generated dual-hypothesis reasoning datasets, through GitHub and Hugging Face.

REFERENCES

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. pp. 283–294, Toronto, Canada, July 2023. doi: 10.18653/v1/2023.acl-short.25. URL 2023.acl-short.25/.

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.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. pp. 4443–4458, Online, July 2020. doi: 10.18653/v1/2020.acl-main.408. URL 2020.acl-main.408/.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*, 2024.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 5992–6026, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.306. URL https://aclanthology.org/2025.naacl-long.306/.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models, 2025. URL https://arxiv.org/abs/2412.16339.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, 2024. URL https://arxiv.org/abs/2406.18495.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL https://arxiv.org/abs/2312.06674.

Md Asiful Islam and Mihai Surdeanu. A lightweight explainable guardrail for prompt safety. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43573–43591, San Diego, California, United States, July 2026. Association for

Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2017. URL https://aclanthology.org/2026.acl-long.2017/.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: from in-the-wild jailbreaks to (adversarially) safer language models. NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Mintong Kang and Bo Li. Rˆ2-guard: Robust reasoning enabled llm guardrail via knowledge-enhanced logical reasoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 63859–63876, 2025.

Jaykumar Kasundra, Anjaneya Praharaj, Sourabh Surana, Lakshmi Sirisha Chodisetty, Sourav Sharma, Abhigya Verma, Abhishek Bhardwaj, Debasish Kanhar, Aakash Bhagat, Khalil Slimi, Seganrasan Subramanian, Sathwik Tejaswi Madhusudhan, Ranga Prasad Chenna, and Srinivas Sunkara. Aprielguard, 2026. URL https://arxiv.org/abs/2512.20293.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*, 2024.

Junyu Lin, Meizhen Liu, Xiufeng Huang, Jinfeng Li, Haiwen Hong, Xiaohan Yuan, Yuefeng Chen, Longtao Huang, Hui Xue, Ranjie Duan, Zhikai Chen, Yuchuan Fu, Defeng Li, Lingyao Gao, and Yitong Yang. Yufeng-xguard: A reasoning-centric, interpretable, and flexible guardrail model for large language models, 2026. URL https://arxiv.org/abs/2601.15588.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation. pp. 4694–4702, Singapore, December 2023. doi: 10.18653/v1/2023.findings-emnlp.311. URL 2023.findings-emnlp.311/.

Yue Liu, Hongcheng Gao, Shengfang Zhai, Jun Xia, Tianyi Wu, Zhiwei Xue, Yulin Chen, Kenji Kawaguchi, Jiaheng Zhang, and Bryan Hooi. Guardreasoner: Towards reasoning-based llm safeguards, 2025. URL https://arxiv.org/abs/2501.18492.

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.

Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. pp. 431–445, Singapore, December 2023. doi: 10.18653/v1/2023.emnlp-demo.40. URL 2023.emnlp-demo.40/.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Makesh Narsimhan Sreedhar, Traian Rebedea, and Christopher Parisien. Safety through reasoning: An empirical study of reasoning guardrail models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21862–21880, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1193. URL https://aclanthology.org/2025.findings-emnlp.1193/.

Zhao Tong, Pengfei Yang, Yimeng Gu, Haichao Shi, Qiang Liu, Xingcheng Xu, Shu Wu, and Xiao-Yu Zhang. RLShield: Dynamic jailbreak detection for LLMs via reinforced adaptive learning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23615–23630, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1182. URL https://aclanthology.org/2026.findings-acl.1182/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Xiaofei Wen, Wenxuan Zhou, Wenjie Jacky Mo, and Muhao Chen. ThinkGuard: Deliberative slow thinking leads to cautious guardrails. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13698–13713, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.704. URL https://aclanthology.org/2025.findings-acl.704/.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. Shieldgemma: Generative ai content moderation based on gemma, 2024. URL https://arxiv.org/abs/2407.21772.

A Standard Deviation Results

Table 5 reports the sample standard deviations corresponding to the ARBITER unsafe F1 scores reported in Tables 1 and 2 across three random seeds (32, 42, and 52). The standard deviations are low, indicating that ARBITER is stable across random seeds for both prompt classification and explainability classification.

Table 5: Standard deviations corresponding to the ARBITER F1 scores for the unsafe class reported in Tables 1 and 2.

Model	Prompt classification			Explainability classification		
	AEGIS2.0	WildGuardMix	ToxicChat0124	AEGIS2.0	WildGuardMix	ToxicChat0124
ARBITER-aegis2.0	0.31	0.41	0.27	0.76	0.47	0.61
ARBITER-wildguardmix	0.10	0.20	0.84	0.48	0.46	1.83
ARBITER-toxicchat0124	1.08	0.58	0.39	0.79	0.84	2.32

B Hyperparameters

Optimization hyperparameters: We train for 3 epochs with AdamW, per-device training batch size 8, gradient accumulation over 32 steps, learning rate 2×10^{-4}, and warmup ratio 0.03. All experiments are run with random seeds 32, 42, and 52.

Reasoning hyperparameters: During reasoning generation, the safe hypothesis budget is 24 tokens, the unsafe hypothesis budget is 24 tokens, and the deliberation budget is 40 tokens.

MC-SFT hyperparameters:

- **High priority:** $\lambda_y = \lambda_e = 1.0$ for the label and explanation losses.
- **Mid priority:** $\lambda_w = 0.5$ for the winner loss.
- **Low priority:** $\lambda_r = \lambda_{\text{syn}} = 0.1$ for the reasoning and syntax losses, and $\lambda_a = 0.2$ for the auxiliary alignment loss.

Table 6: Hyperparameter sensitivity analysis on the AEGIS2.0 development set, reporting F1 for the final predicted safety label. Here, λ_r, λ_w, λ_y, λ_e, λ_{syn}, and λ_a denote the reasoning, winner, label, explanation-span, syntax, and alignment loss weights, respectively. For each one-at-a-time experiment, we set the corresponding loss weight to the hyperparameter value shown in the first column while keeping all other weights fixed at the default value of 1.0. The ARBITER *configuration* row reports the result using the complete configuration of ARBITER, where $\lambda_r = \lambda_{\text{syn}} = 0.1$, $\lambda_w = 0.5$, $\lambda_y = \lambda_e = 1.0$, and $\lambda_a = 0.2$ are applied simultaneously.

Hyperparameter value	F1 (Final predicted label)					
	λ_r	λ_w	λ_y	λ_e	λ_{syn}	λ_a
0.05	87.94	87.20	88.15	88.18	88.50	87.79
0.10	**88.43**	-	-	-	**88.67**	-
0.20	-	-	-	-	-	88.35
0.25	88.11	87.04	87.76	**88.58**	88.13	88.21
0.50	88.20	87.53	88.41	88.43	88.62	88.28
0.75	88.27	**88.34**	**88.45**	88.24	88.12	**88.50**
1.00	88.11	88.11	88.11	88.11	88.11	88.11
ARBITER configuration	87.93					

These values were chosen based on the intuition described in Section 3.2.1 and were not tuned on the test sets.

B.1 HYPERPARAMETER SENSITIVITY ANALYSIS

As discussed above, the loss weights used in MC-SFT were selected based on our intuition about the relative importance of different output components, rather than through extensive hyperparameter tuning. To examine the sensitivity of ARBITER to these choices, we conduct a hyperparameter sensitivity analysis on the AEGIS2.0 dev set.

We use a one-at-a-time sensitivity analysis. That is, we start from a default configuration in which all loss weights are set to 1.0. Then, for each $\lambda \in \{\lambda_r, \lambda_w, \lambda_y, \lambda_e, \lambda_{\text{syn}}, \lambda_a\}$, we vary only that weight while keeping all remaining weights fixed at 1.0. Specifically, we evaluate values of 0.05, 0.25, 0.50, 0.75, and 1.00 for all loss weights. In addition, because $\lambda_r = 0.1$, $\lambda_{\text{syn}} = 0.1$, and $\lambda_a = 0.2$ are used in the final ARBITER configuration, we also evaluate these values individually in the one-at-a-time setting while keeping all other weights fixed at 1.0. Thus, each numeric entry in Table 6 corresponds to changing only the weight associated with that column while keeping all other weights at their default value of 1.0. To compare, we also report the ARBITER *configuration*, which uses all of the loss weights used for ARBITER together: $\lambda_r = \lambda_{\text{syn}} = 0.1$, $\lambda_w = 0.5$, $\lambda_y = \lambda_e = 1.0$, and $\lambda_a = 0.2$.

The results in Table 6 show that additional hyperparameter tuning could potentially yield modest improvements. However, performance remains relatively stable. The ARBITER full weight configuration achieves an F1 score of 87.93%, while the other configurations range from 87.04% to 88.67%. Thus, the difference between the ARBITER configuration and every other tested setting is less than one absolute F1 point. These results indicate that ARBITER is not highly sensitive to the precise choice of individual loss weights. Given the small potential improvement relative to the computational cost of an extensive hyperparameter search, we use the intuitively selected weights throughout our experiments.

This result is significant because it suggests that the improvements from MC-SFT are not primarily driven by careful tuning of the individual λ values, but by the specific design of the MC-SFT loss function. A key distinction between standard SFT and MC-SFT is how the token-level losses are aggregated. Standard SFT computes a single average cross-entropy loss over the entire target sequence. Consequently, the supervision signal from short but decision-critical components, such as the final safety label, can be diluted by the substantially larger number of reasoning and formatting tokens. In contrast, MC-SFT first computes the average loss independently within each output component before combining the component-level losses, preventing components with more tokens from dominating the training objective. This prevents the contribution of short but important components from being

diluted simply because they contain fewer tokens. This is supported by the ablation results in Table 3, where MC-SFT improves over the corresponding standard SFT in the majority of cases. Taken together, these results suggest that the primary benefit of MC-SFT comes from its component-wise loss formulation, while the precise weighting of individual components has a comparatively smaller effect.

C Human Evaluation of Reasoning Quality

We conducted a human evaluation to assess the quality of the dual-hypothesis reasoning traces generated by our model trained on the AEGIS 2.0 training set. We evaluated 200 examples (100 safe and 100 unsafe) randomly selected from the AEGIS 2.0 test set. Two annotators independently evaluated the reasoning traces. The annotators assessed the generated reasoning based on three criteria: (1) reasoning compliance, (2) internal reasoning consistency, and (3) reasoning correctness.

Reasoning compliance measures whether all three reasoning steps fulfill their intended roles. Specifically, the Safe Hypothesis should provide an argument for the safe interpretation, the Unsafe Hypothesis should provide an argument for the unsafe interpretation, and the deliberation should select one of the two hypotheses as more plausible. A reasoning trace is considered compliant only when all three conditions are satisfied.

Internal reasoning consistency measures whether the decision expressed in the deliberation agrees with the intermediate winner and the final predicted label.

Reasoning correctness measures whether the deliberation decision agrees with the gold label.

Table 7 presents the human-evaluation results. Approximately 85% of the reasoning traces satisfy the complete reasoning-compliance criterion. These results demonstrate that the trained model can generally produce the intended dual-hypothesis reasoning structure. Internal reasoning consistency is also approximately 85%. This means that the intermediate winner and the final predicted label each agree with the deliberation decision in approximately 85% of the evaluated examples, indicating strong internal consistency. Finally, the deliberation decision agrees with the gold label in approximately 80% of the examples, indicating that most deliberations reach the correct conclusion. Inter-annotator agreement is high across all evaluation criteria. Raw agreement ranges from 95.5% to 97.5%, while Cohen's kappa ranges from 0.826 to 0.923. The Cohen's kappa scores indicate substantial agreement on reasoning compliance and near-perfect agreement on internal reasoning consistency and reasoning correctness.

Separately, the intermediate winner and the final predicted label agree in 99.5% of the examples. Since this can be computed directly from the model outputs, it does not require human evaluation. Overall, the evaluation shows that our proposed method can reliably generate structured dual-hypothesis reasoning that is generally compliant, internally consistent, and aligned with the correct decision.

D Generalization Across Model Backbones and Scales

Our experiments in the main body of the paper use Llama-3.1-8B-Instruct as the base model for ARBITER. To examine whether the proposed method generalizes beyond this specific backbone and model size, we additionally train ARBITER using Qwen3-8B and Llama-3.2-3B. Qwen3-8B evaluates generalization to a different model family, while Llama-3.2-3B evaluates whether the method remains effective at a smaller scale. Each model is trained following the same procedure described in Section 3.2, including the same hyperparameters. In particular, each base model first generates its own dual-hypothesis reasoning traces, which are then used to fine-tune the same model with MC-SFT.

Table 8 reports the results across three datasets. Overall, ARBITER maintains strong performance across different model families and scales. The Qwen3-8B variants are competitive with and, in several cases, outperform the Llama-3.1-8B variants, indicating that the benefits of the proposed training method are not specific to the Llama backbone. The Llama-3.2-3B variants also achieve competitive performance despite having substantially fewer parameters, except in one out-of-domain setting. This may suggest that larger models generalize better to out-of-domain distributions.

Table 7: Human evaluation of dual-hypothesis reasoning quality on 200 randomly selected AEGIS 2.0 test examples (100 safe and 100 unsafe) by two independent annotators. The Annotator 1 and Annotator 2 columns report the percentage of examples that each annotator judged as satisfying the corresponding criterion. Agreement is the percentage of examples on which the two annotators agreed, while Cohen's κ reports inter-annotator agreement using Cohen's kappa score.

Assessment Type	**Condition**	***Annotator 1 (%)***	***Annotator 2 (%)***	***Agreement (%)***	***Cohen's κ***
Reasoning Compliance	All three reasoning steps fulfill their intended roles	84.5	85.0	95.5	0.826
Internal Reasoning Consistency	Deliberation agrees with the winner	85.0	84.5	97.5	0.903
	Deliberation agrees with the predicted label	85.5	85.0	97.5	0.901
Reasoning Correctness	Deliberation agrees with the gold label	79.5	80.0	97.5	0.923

These results suggest that our dual-hypothesis reasoning with MC-SFT approach generalizes across both model families and scales.

Table 8: Comparison of ARBITER using different model families and scales as backbones. Results report prompt classification performance in F1 scores. Gray cells indicate in-domain evaluation, while the remaining cells indicate out-of-domain evaluation.

Model	**Test sets**		
	AEGIS2.0	WildGuardMix	ToxicChat0124
Trained on AEGIS 2.0			
ARBITER-aegis2.0-llama-3.2-3B	84.24	81.60	72.03
ARBITER-aegis2.0-qwen3-8B	**87.71**	83.06	**73.46**
ARBITER-aegis2.0-llama-3.1-8B	87.09	**84.37**	71.85
Trained on WildGuardMix			
ARBITER-wildguardmix-llama-3.2-3B	80.68	89.04	63.89
ARBITER-wildguardmix-qwen3-8B	**82.20**	**89.12**	64.78
ARBITER-wildguardmix-llama-3.1-8B	81.91	88.87	**68.76**
Trained on ToxicChat0124			
ARBITER-toxitchat0124-llama-3.2-3B	60.20	69.83	78.84
ARBITER-toxitchat0124-qwen3-8B	77.79	70.78	**83.71**
ARBITER-toxitchat0124-llama-3.1-8B	**80.06**	**73.90**	82.16

E ROBUSTNESS ANALYSIS

In this section, we present additional robustness analyses of ARBITER that are particularly important for guardrail models: detecting jailbreak prompts and avoiding false-positive predictions on benign prompts that contain harmful-looking words. We perform these evaluations using the ARBITER model trained on the WildGuardMix dataset.

E.1 JAILBREAK DETECTION

Our main evaluation datasets, AEGIS2.0, WildGuardMix, and ToxicChat0124, already contain jailbreak prompts. Therefore, the strong performance of ARBITER on these benchmarks provides evidence of its ability to detect jailbreak attempts. As an additional targeted evaluation, we evaluate ARBITER on WildJailbreak (Jiang et al., 2024), a dataset specifically designed for jailbreak analysis.

We evaluate using the ARBITER model trained on the WildGuardMix dataset, so this is an out-of-domain evaluation.

Table 9 reports the performance of ARBITER-wildguardmix on WildJailbreak. Among the reported metrics, recall and FNR are particularly relevant for jailbreak detection because they directly measure how many jailbreak prompts are successfully detected or missed. ARBITER achieves 96.15% recall, corresponding to an FNR of only 3.85%, and obtains an F1 score of 97.64%. These results show that ARBITER reliably detects jailbreak prompts and remains robust under this targeted jailbreak evaluation.

We also compare against the recent baseline AprielGuard (Kasundra et al., 2026), which, to the best of our knowledge, is the only available baseline reporting results on the full WildJailbreak test set. We exclude other reported baselines because they evaluate only on subsets of the WildJailbreak test set. ARBITER-wildguardmix outperforms AprielGuard on all reported metrics.

Table 9: Jailbreak detection performance of ARBITER-wildguardmix on WildJailbreak. FPR and FNR denote the false-positive and false-negative rates, respectively.

Model	Precision	Recall	F1	FPR	FNR
AprielGuard (Kasundra et al., 2026)	97.00	96.00	96.00	31.00	-
ARBITER-wildguardmix	**99.17**	**96.15**	**97.64**	**7.62**	3.85

E.2 BENIGN PROMPTS WITH HARMFUL WORDS: OVER-REFUSAL ANALYSIS

In addition to detecting jailbreak prompts, a robust guardrail should avoid false-positive predictions on benign prompts that contain commonly harmful words in safe contexts. XSTest (Röttger et al., 2024) is a test set specifically designed to evaluate this failure mode using benign prompts that contain harmful words in benign contexts.

For example, a benign prompt may use a potentially harmful word such as "kill" when referring to terminating a computer process. A guardrail that relies too heavily on surface-level harmful words may incorrectly classify such a prompt as unsafe, resulting in over-refusal.

Table 10 reports the performance of ARBITER-wildguardmix on XSTest. We evaluate using the ARBITER model trained on the WildGuardMix dataset, so this is an out-of-domain evaluation. Among the reported metrics, FPR is particularly relevant for this analysis because it measures how often benign prompts are incorrectly classified as unsafe. ARBITER achieves an FPR of only 4.80%, indicating that most benign harm-looking prompts are correctly recognized as safe. At the same time, it achieves an F1 score of 95.57%, outperforming all baseline models compared in Table 10. These results show that ARBITER is robust to benign prompts containing harmful words in safe contexts and does not show an over-refusal tendency.

Table 10: Over-refusal analysis on XSTest. FPR and FNR denote the false-positive and false-negative rates, respectively. Baseline results are taken from Tong et al. (2026).

Model	Precision	Recall	F1	FPR	FNR
LlamaGuard3-8B	95.20	80.00	87.00	–	–
LlamaGuard4-12B	90.10	77.50	83.30	–	–
GradSafe (Llama-3.1-8B)	88.60	93.50	91.00	–	–
RLShield (Llama-2-7B)	86.90	96.50	91.50	–	–
RLShield (Llama-3.1-8B)	**95.30**	91.00	93.10	–	–
ARBITER-wildguardmix	94.17	**97.00**	**95.57**	4.80	3.00

F INFERENCE TIME ANALYSIS

Reasoning-based guardrails require additional inference time because they generate intermediate reasoning tokens before producing the final prediction. To quantify this overhead, we compare the inference speed of ARBITER with a non-reasoning baseline. For both models, we evaluate the same 500 prompts sampled from the WildGuardMix test set and report the average inference time. All experiments are conducted on a single NVIDIA H100 GPU.

As expected, ARBITER incurs additional inference cost due to the generation of the safe hypothesis, unsafe hypothesis, and deliberation before producing the final prediction. ARBITER requires approximately 0.664 seconds per example, compared with 0.067 seconds for the non-reasoning baseline. Similarly, throughput decreases from approximately 15.02 to 1.51 examples per second.

Additional inference latency is expected for reasoning-based guardrails and has also been reported in prior work. For example, Sreedhar et al. (2025) report a 170.7% latency overhead for unrestricted reasoning compared with their non-reasoning baseline, which decreases to 34.9% when reasoning is limited to one sentence. Their evaluation uses eight NVIDIA A100 GPUs with vLLM and dynamic batching, while ours uses a single NVIDIA H100 GPU, so the absolute inference times are not directly comparable. Nevertheless, both results show the expected inference overhead introduced by explicit reasoning. This additional cost represents the expected trade-off of explicit reasoning, which provides benefits in performance and explainability.

Table 11: Inference time comparison between ARBITER and a non-reasoning baseline using a single NVIDIA H100 GPU.

	Inference time (second/example)	Throughput (example/second)
Non-reasoning baseline	0.067	15.02
ARBITER	0.664	1.51

G CATEGORY-WISE PERFORMANCE ANALYSIS

WildGuardMix contains prompts from a diverse set of unsafe categories. To provide a more detailed view of ARBITER's performance, we report category-wise F1 scores for ARBITER-wildguardmix on the WildGuardMix test set. Table 12 shows that ARBITER maintains strong performance across all categories, with F1 scores ranging from 86.57% to 100.00%. Overall, these results indicate that ARBITER performs consistently across diverse types of unsafe prompts rather than relying on strong performance in only a subset of categories.

Table 12: Category-wise F1 scores of ARBITER-wildguardmix on the WildGuardMix test set.

Unsafe Category	F1
Causing Material Harm by Disseminating Misinformation	98.88
Copyright Violations	89.29
Cyberattack	98.95
Defamation Encouraging Unethical or Unsafe Actions	100.00
Disseminating False or Misleading Information Encouraging Disinformation Campaigns	98.85
Fraud Assisting Illegal Activities	88.89
Mental Health Over-reliance Crisis	91.95
Others	87.36
Private Information Individual	89.80
Sensitive Information Organization Government	90.11
Sexual Content	88.10
Social Stereotypes and Unfair Discrimination	86.57
Toxic Language Hate Speech	100.00
Violence and Physical Harm	98.73

H Extended Classification Results

The main results in Section 5 report F1 scores for prompt classification. In this section, we provide additional classification metrics for a more detailed analysis of the ARBITER variants. Table 13 reports precision, recall, and F1, while Table 14 reports false-positive rate (FPR) and false-negative rate (FNR) across all evaluations. The results show that the ARBITER variants maintain a relatively balanced precision-recall trade-off in most settings. They also achieve low FPR and FNR in most cases.

Table 13: Detailed prompt classification performance of ARBITER variants trained on AEGIS2.0, WildGuardMix, and ToxicChat0124. Results report precision (P), recall (R), and F1 scores on all three test sets.

Model	AEGIS2.0			WildGuardMix			ToxicChat0124		
	P	R	F1	P	R	F1	P	R	F1
ARBITER-aegis2.0	**88.19**	86.02	**87.09**	84.26	84.48	84.37	66.66	77.90	71.85
ARBITER-wildguardmix	73.72	**92.16**	81.91	**91.56**	**86.34**	**88.87**	56.48	87.85	68.76
ARBITER-toxichat0124	87.29	73.94	80.06	88.07	63.66	73.90	**75.35**	**90.33**	**82.16**

Table 14: False-positive rate (FPR) and false-negative rate (FNR) of ARBITER variants trained on AEGIS2.0, WildGuardMix, and ToxicChat0124 and evaluated across all three test sets.

Model	AEGIS2.0		WildGuardMix		ToxicChat0124	
	FPR	FNR	FPR	FNR	FPR	FNR
ARBITER-aegis2.0	13.48	13.98	12.26	15.52	2.99	22.10
ARBITER-wildguardmix	38.45	**7.84**	**6.18**	**13.66**	5.19	12.15
ARBITER-toxichat0124	**12.60**	26.06	6.69	36.34	**2.27**	**9.67**

I Faithfulness Analysis of Explanation Spans

Deletion-based perturbation is commonly used to evaluate whether an explanation is faithful to a model's prediction: if the evidence identified by an explanation is removed, the model's confidence in the original prediction should decrease (DeYoung et al., 2020; Atanasova et al., 2023). We use this idea to evaluate whether the explanation spans generated by ARBITER are causally connected to its unsafe decisions.

We restrict the analysis to unsafe true-positive examples, where the prompt label is unsafe and ARBITER also predicts unsafe. For each selected example, we match the generated explanation spans to the original prompt and create perturbed prompts by deleting the matched spans. We consider four deletion settings: removing the first one, first two, first three, or all matched explanation spans. We then rerun ARBITER on each perturbed prompt using the same inference procedure and measure how the model probability assigned to the unsafe label changes.

We report the original mean unsafe probability, the mean unsafe probability after each deletion setting, comprehensiveness, and area over the perturbation curve (AOPC). Let p_i^0 be the unsafe probability for the original prompt x_i, and let p_i^c be the unsafe probability after deletion setting c. Comprehensiveness measures the average drop in unsafe probability:

$$\mathrm{Comp}(c) = \frac{1}{N} \sum_{i=1}^{N} \left(p_i^0 - p_i^c \right). \tag{6}$$

A larger positive value indicates that removing the generated explanation spans makes the model less confident in the unsafe label. We summarize the deletion curve using AOPC:

$$\mathrm{AOPC} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{Comp}(c), \tag{7}$$

where $\mathcal{C}$ is the set of deletion settings. Higher AOPC indicates that the explanation spans identified by ARBITER are more strongly connected to its unsafe prediction.

Table 15 reports the deletion-based faithfulness results. Across all train–test settings, deleting the generated explanation spans consistently reduces the model's unsafe probability, and the reduction becomes larger as more spans are removed. This is reflected in the comprehensiveness scores, which are positive for every deletion setting and increase from D1 to deleting all matched spans in every row. The delete-all comprehensiveness ranges from 0.439 to 0.808, indicating a substantial drop in unsafe confidence after removing the full set of cited evidence. The AOPC scores range from 0.296 to 0.699, with an average of 0.531, showing that the extracted explanation spans are strongly connected to ARBITER's unsafe predictions. These results support the faithfulness of the explanation spans: when the cited evidence is removed from the prompt, the model becomes substantially less confident in the unsafe label.

Table 15: Deletion-based faithfulness analysis of explanation spans. Orig. is the original mean unsafe probability. D1, D2, D3, and All denote deletion of the first one, first two, first three, and all matched explanation spans, respectively. Higher comprehensiveness and AOPC indicate stronger faithfulness.

Train Set	Test Set	Orig.	Mean P_{unsafe} After Deletion				Comprehensiveness				AOPC
			D1	D2	D3	All	D1	D2	D3	All	
AEGIS2.0	AEGIS2.0	0.99	0.542	0.289	0.217	0.192	0.458	0.711	0.783	0.808	0.690
	WildGuardMix	0.99	0.769	0.603	0.493	0.392	0.231	0.397	0.507	0.608	0.436
	ToxicChat0124	0.99	0.567	0.422	0.358	0.316	0.433	0.578	0.642	0.684	0.584
WildGuardMix	AEGIS2.0	0.99	0.611	0.365	0.294	0.269	0.389	0.635	0.706	0.731	0.615
	WildGuardMix	0.99	0.877	0.753	0.627	0.561	0.123	0.247	0.373	0.439	0.296
	ToxicChat0124	0.99	0.619	0.459	0.393	0.362	0.381	0.541	0.607	0.638	0.542
ToxicChat0124	AEGIS2.0	0.99	0.469	0.277	0.234	0.221	0.530	0.722	0.765	0.778	0.699
	WildGuardMix	0.99	0.737	0.598	0.502	0.425	0.262	0.402	0.498	0.575	0.434
	ToxicChat0124	0.99	0.642	0.520	0.471	0.428	0.358	0.480	0.529	0.572	0.485

J INSTRUCTION TEMPLATE

This appendix provides the instruction template used for dual-hypothesis reasoning generation, supervised fine-tuning, and inference in ARBITER.

```
Instruction:
You are a guardrail model. Your job is to identify whether the given
    prompt is safe or unsafe. Read the prompt carefully and reason
    through the decision before producing the final JSON.

You must produce the following reasoning sections:

Safe Hypothesis:
Think about the strongest plausible reason why the prompt could be safe.
    Consider benign intent, harmless interpretation, neutral context, or
    ordinary non-harmful usage if they are supported by the prompt. In
    this section, argue only for the safe interpretation. Do not argue
    that the prompt is unsafe here.

Unsafe Hypothesis:
Think about the strongest plausible reason why the prompt could be unsafe.
     Focus on harmful intent, illegal activity, abuse, dangerous behavior,
     sexual content, evasion, exploitation, or any other unsafe
    interpretation supported by the prompt. Pay attention to the exact
    words or phrases in the prompt that support the unsafe reading. In
    this section, argue only for the unsafe interpretation. Do not argue
    that the prompt is safe here.

Deliberation:
Compare the Safe Hypothesis and the Unsafe Hypothesis carefully. Decide
    which hypothesis is better supported by the prompt itself. Explain
```

```
    why the stronger hypothesis is more plausible. Base the decision on
    concrete evidence from the prompt, not on speculation.

Winner:
After the deliberation, output exactly one of the following:
Winner: safe
Winner: unsafe

Final JSON:
After the Winner line, output the final answer only as JSON after the
    Final JSON section.

Label definitions:
"safe" = The prompt does not request harmful, illegal, abusive, or
    dangerous actions.
"unsafe" = The prompt requests or encourages harmful, illegal, abusive,
    or dangerous actions.

If the label is "safe", the explanation must be ["safe prompt"].
If the label is "unsafe", extract the exact unsafe words or phrases from
    the prompt and save them as a list in the explanation. Do not
    paraphrase them. Do not invent new words that are not present in the
    prompt.

The final JSON must be valid and in one of the following formats:
{"label":"safe","explanation":["safe prompt"]}
{"label":"unsafe","explanation":["exact unsafe words or phrase"]}
Do not put any extra text after the final JSON.
```