

NeurIPS’25 - Reinforcement Learning with Backtracking Feedback - Supplemental

A. Additional Experimental Details

Training Setup: All experiments were conducted using A100 GPUs with 80GB memory. Models were trained using mixed precision (bf16), AdamW optimizer, with a linear warmup over 100 steps followed by cosine decay. The learning rate for fine-tuning was set to 1e-5 for SFT and 5e-6 during RL. For RL, we used Group Relative Policy Optimization (GRPO) and balanced the loss with $\lambda_{\text{SFT}} = 0.1$.

Dataset Details: We curated high-quality prompt-response pairs from instruction-tuned LLMs, and used BSAFE+ for injecting coherent violations. Violation types include *Toxicity*, *Hate Speech*, *Health*, and *Finance*. A total of 300k augmented examples were created for SFT.

Critic Details: A single LLM-based critic evaluated model generations. It outputs violation spans, categories, and post-edit evaluations, which informed the reward function. The critic was calibrated on a 2k manually annotated set to ensure sensitivity and specificity to safety categories.

B. Detailed Ablations

We performed an ablation to assess the importance of the backtracking mechanism in different positions within generation. Results are summarized below:

Model Variant	ASR on LMSYS (%)
Full RLBF	1
RLBF w/o Backtracking	18
RLBF w/o Middle Backtracking	7

These results confirm the central importance of dynamic backtracking throughout the generation process.

C. Safety Category Breakdown

We report the prevention rate per category on LMSYS-MF. RLBF achieves over 0.96 in most categories, demonstrating strong generalizability across adversarial content types.

Category	Prevention Rate
Hate Speech	0.98
Toxicity	0.96
Politics	0.98
Health	0.98
Dangerous Content	0.96
Sexual Content	0.94
Public Safety	0.96
Illicit Drugs	0.94

D. Resource Usage and Reproducibility

Total compute usage was approximately 800 GPU hours per model. SFT took 250 GPU hours and RL 550 GPU hours for training all given models. All prompts, reward functions, and critic guidelines are provided in the main paper and available upon request post-acceptance.

E. License and Terms of Use for Assets

We used several publicly available models and datasets under their respective licenses:

- **LLaMA 3 (Meta)** – Used under the LLaMA Community License Agreement. No modifications were made.
- **Gemma 2 (Google)** – Used under the Gemma license (Creative Commons Attribution-NonCommercial 4.0 International).
- **LMSYS Datasets** – Used under CC BY-SA 4.0, where applicable. These include open instruction-following benchmark datasets.

F. Limitations and Societal Considerations

While RLBF performs well on safety metrics, its dependency on a critic LLM means performance may degrade if the critic fails to detect nuanced violations. Future work may consider critic ensembles or retrieval-augmented evaluators for robustness.

We acknowledge that RLBF mitigates but does not eliminate misuse risk in open-ended generation. As such, deployment should be paired with continuous monitoring and user-side safety layers.