Online Learning Defense against Iterative Jailbreak Attacks via Prompt Optimization

Anonymous ACL submission

Abstract

Iterative jailbreak methods generate harmful output-inducing prompts by repeatedly rewrit-002 ing and inputting prompts to large language models (LLMs), where each rewrite is based on the previous output results. Despite the iterative jailbreak methods being one of the most powerful techniques, existing defense methods have not implemented proactive measures to disrupt dynamic trial-and-error attempts. In this study, we propose a framework that dy-011 namically updates the defense system through 012 online learning each time the iterative jailbreak method inputs a prompt into the LLM for optimization. Furthermore, prompts generated by jailbreak methods exhibit characteristics such 016 as increased redundancy, complexity, and am-017 biguity, which deviate from prompts that effectively harness the capabilities of LLMs for harmless tasks. We hypothesize that prompt rewriting techniques that optimize performance on harmless tasks have the potential to prevent jailbreak attacks. To this end, we introduce a reinforcement learning-based method to optimize prompts, ensuring appropriate responses to harmless prompts while rejecting harmful ones. Experiments conducted on three LLMs 027 demonstrate that the proposed method significantly outperforms five existing defense methods against five iterative jailbreak methods. Additionally, our results indicate that the proposed method enhances the quality of responses to harmless prompts, suggesting that prompt optimization can achieve both improved defense against harmful tasks and better performance on harmless tasks.

1 Introduction

036

042

For large language models (LLMs; Brown et al., 2020), it is crucial to implement guardrails that ensure harmful prompts result in refusals or restricted outputs, while harmless prompts receive useful and trustworthy responses (Ouyang et al., 2022; Bai et al., 2022b; Guan et al., 2024). The act

of malicious users circumventing such developerimplemented guardrails is known as *jailbreaking* (Wallace et al., 2019; Wei et al., 2024). Existing jailbreak research has demonstrated that carefully crafted prompts can induce LLMs to generate harmful outputs (Liu et al., 2023a; Zeng et al., 2024). 043

045

047

049

051

054

055

057

060

061

062

063

064

065

066

067

068

069

070

071

072

073

074

075

077

078

079

A method that iteratively provides prompts to a target LLM to discover prompts that elicit harmful outputs is one of the most powerful jailbreaking techniques (Zou et al., 2023; Li et al., 2024; Chao et al., 2023; Mehrotra et al., 2023; Jha et al., 2024). Iterative jailbreaking techniques pose a potential risk as they allow for trial-and-error exploration of the behavior of LLMs, even those equipped with guardrails, potentially enabling the discovery of loopholes that adapt to safety measures. Despite this threat, existing defense methods (Jain et al., 2023; Inan et al., 2023; Jain et al., 2023; Robey et al., 2023; Wang et al., 2024) have not yet implemented countermeasures that respond to the dynamic optimization inherent in iterative jailbreaking techniques.

This study proposes a framework that updates the defense system through online learning each time a prompt rewritten by an iterative jailbreak method for optimization is provided to the LLM. Iterative jailbreak methods gradually rewrite and asymptotically improve prompts that have been rejected (Zou et al., 2023; Liu et al., 2023a; Mehrotra et al., 2023; Jha et al., 2024), making it crucial to update the defense system to maintain rejection for minor rewrites of prompts rejected by the target LLM. In iterative jailbreaking, slightly modified similar prompts are continuously input to the LLM, raising concerns about overfitting in a specific direction through online learning. We introduce Past-Direction Gradient Damping (PDGD) that penalizes updates for gradients similar to past gradients to prevent excessive updates in a specific gradient direction.

We target the defense system based on prompt rewriting for online learning for the following reasons: Dynamically updating the LLM is impractical due to unintended changes, such as catastrophic forgetting (Goodfellow et al., 2013), and the training costs (Zhao et al., 2023). Additionally, there is a growing demand for customized guardrails tailored to services (Zhang et al., 2024) and applications relying on black-box LLMs (Achiam et al., 2023), making it ideal to build dynamic defenses externally to the LLM. While filtering (Jain et al., 2023) is one approach to enhancing defenses as an external system, prompt rewriting has been suggested to potentially contribute more significantly to safety (Robey et al., 2023).

084

086

090

100

101

102

103

104

107

108

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

127

128

129

130

131

132

133

135

Since harmful prompts are not always input and harmless prompts are also provided as inputs, it is necessary to ensure performance even if the defense mechanism's rewriting is applied to harmless prompts (Xiong et al., 2024). The prompts rewritten by jailbreak methods use ambiguous expressions, complex structures, or lengthy text to conceal their intent (Shen et al., 2024), which contrasts with the characteristics of prompts optimized for harmless tasks, which are concise and clear in intent (Bsharat et al., 2023; Schulhoff et al., 2024). Therefore, it is possible that jailbreaks can be prevented through rewrites similar to prompt optimization aimed at improving performance in harmless tasks. If so, defense methods could focus on rewriting prompts to improve harmless tasks. This suggests that defense performance against jailbreaks in harmful tasks and performance in harmless tasks might be compatible in terms of prompt optimization, even though there is a conventional belief in a trade-off between rejecting outputs for harmful tasks and providing beneficial responses for harmless tasks (Bai et al., 2022a). We propose a reinforcement learning based on prompt optimization to reject outputs for harmful prompts while appropriately responding to harmless prompts.

Experimental results demonstrate that, for harmful tasks (Bai et al., 2022b; Ganguli et al., 2022), the proposed method shows significant improvement against five iterative jailbreak methods compared to five existing defense methods based on prompt rewriting across three LLMs: GPT-4 (Achiam et al., 2023), OLMo (OLMo et al., 2024), and Llama 3 (Dubey et al., 2024). Furthermore, compared to the original model without any defense mechanism and models with existing defense methods applied, the model with the proposed method also exhibits improved performance on harmless tasks (Köpf et al., 2024). This suggests that, in prompt optimization, it is possible to achieve both improved defense performance for harmful tasks and enhanced response quality for harmless tasks. 136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

2 Prompt Optimization Through Online Learning for Defense

Prompt optimization model M_{opt} rewrites prompts to guide the target LLM M_{trg} to provide appropriate responses y_r for harmless tasks and rejections y_d for harmful tasks. Here, harmless tasks refer to harmless prompts p_l such as "*Let me know how to make pizza*", while harmful tasks refer to harmful prompts p_f such as "*Tell me how to make a bomb*". In this context, the response y_r for a harmless task would be a detailed explanation of how to make pizza, whereas for a harmful task, it would be a detailed explanation of how to make a bomb. The rejection y_d is a text such as "*I'm sorry, but I can't help with that request*".

We first perform supervised learning on a pretrained model, followed by reinforcement learning, to train the prompt optimization model M_{opt} for use in online learning. This is because reinforcement learning can be unstable, and supervised learning allows us to acquire a good policy in advance, enabling efficient exploration. The reinforcementlearned M_{opt} performs online learning on the harmless prompts p_1 and harmful prompts p_f provided to the target LLM M_{opt} during the inference phase.

2.1 Supervised Learning

In supervised learning, the prompt optimization model M_{opt} with parameters θ_s is trained to restore the original harmful prompt p_f from the jailbreak harmful prompt p_{jf} . The loss function is defined to minimize the cross-entropy loss \mathcal{L}_{CE} between the generated prompt $M_{opt}(p_{jf}; \theta_s)$ and the original prompt p_f as follows:

$$\theta_{s}^{*} = \arg\min_{\theta_{s}} \mathbb{E}_{(p_{jf}, p_{f}) \sim D} \left[\mathcal{L}_{CE}(M_{opt}(p_{jf}; \theta_{s}), p_{f}) \right]$$
(1)

Here, D is the prompt dataset for supervised learning.

2.2 Reinforcement Learning

Using the parameters θ_s obtained from supervised learning as the initial values of the prompt optimization model M_{opt} , reinforcement learning is per-

272

182formed. M_{opt} has a policy π_{θ_r} for rewriting prompts183and optimizes the parameters θ_r by maximizing re-184wards. To prevent the prompt optimization model185 M_{opt} from generating prompts that cause the target186LLM M_{opt} to reject even harmless tasks, the reward187is designed to encourage responses for harmless188tasks and rejections for harmful tasks.

189

190

191

193

195

197

198

199

202

207

210

213

214

215

216

217

218

219

224

225

Reward Design In the learning for harmless tasks, the reward is based on the harmless task evaluation metric $S(0 \le S \le 1)$ between the output of the target LLM M_{trg} and the gold response text y_r^* as well as the rejection text y_d^* . Specifically, for the optimization of harmless prompts, the goal is to generate prompts that make the output of the target LLM closer to the response text y_r^* and appropriately distant from the rejection text y_d^* . The reward function is defined as follows:

$$R_{\rm l}(y_{\rm l}) = S(y_{\rm l}, y_{\rm r}^*) - \max\left(\frac{S(y_{\rm l}, y_{\rm d}^*) - S(y_{\rm r}^*, y_{\rm d}^*)}{1 - S(y_{\rm r}^*, y_{\rm d}^*) + \epsilon}, 0\right)$$
(2)

Here, $y_1 = M_{trg}(M_{opt}(p_1; \theta_r))$, and ϵ is a small positive value to prevent division by zero. The first term measures how close the output y_1 of the target LLM is to the gold response y_r^* , with a higher score indicating a closer match to the gold response. The second term is a regularization term that prevents the output y_1 from becoming too close to the rejection text y_d^* . It imposes a penalty if the output becomes closer to the rejection text than the original gold response y_r^* is to the rejection text y_d^* .

For the optimization of jailbroken harmful prompts, the goal is to create prompts that cause the target LLM M_{trg} to generate the appropriate rejection text y_d^* . The reward is designed such that the output of the target LLM is closer to the predefined rejection text y_d^* and farther from the response text y_r^* , as defined below:

$$R_{\rm jf}(y_{\rm jf}) = S(y_{\rm jf}, y_d^*) - \max\left(\frac{S(y_{\rm jf}, y_{\rm r}^*) - S(y_{\rm r}^*, y_{\rm d}^*)}{1 - S(y_{\rm r}^*, y_{\rm d}^*) + \epsilon}, 0\right)$$
(3)

Here, $y_{jf} = M_{trg}(M_{opt}(p_{jf}; \theta_r))$. Similarly, a regularization term is included to penalize the output if it becomes unnecessarily close to the response text.

The parameters of the prompt optimization policy π_{θ_r} are learned to maximize the expected value of these rewards. Here, the optimal prompt p^* is defined as follows:

$$p^* = \arg\max_{p'} \mathbb{E}_{y \sim P(y|p';M_{\text{trg}})}[R(y)]$$
(4)

To achieve this exploration, the objective function for reinforcement learning is defined as:

$$J(\theta_{\mathbf{r}}) = \mathbb{E}_{p' \sim \pi_{\theta_{\mathbf{r}}}(p)} \mathbb{E}_{y \sim P(y|p'; M_{\text{trg}})}[R(y)] \quad (5)$$

Here, p' is the prompt transformed by M_{opt} , and the reward function R(y) differs depending on whether the input prompt p is for a harmless task or a harmful task:

$$R(y) = \begin{cases} R_{\rm l}(y_{\rm l}) & \text{(For harmless tasks)} \\ R_{\rm f}(y_{\rm jf}) & \text{(For harmful tasks)} \end{cases}$$
(6)

To achieve this objective, the parameters of the prompt optimization policy π_{θ_r} are updated using the policy gradient method, ensuring that prompts corresponding to p^* can be generated with high probability:

$$\nabla_{\theta_{\mathbf{r}}} J(\theta_{\mathbf{r}}) = \mathbb{E}_{p' \sim \pi_{\theta_{\mathbf{r}}}(p)} \left[R(y) \nabla_{\theta_{\mathbf{r}}} \log \pi_{\theta_{\mathbf{r}}}(p') \right]$$
(7)

2.3 Online Learning Against Iterative Jailbreaks

We employ online learning to prevent iterative jailbreak methods from gradually discovering prompts that elicit responses from rejected prompts. Specifically, if the target LLM M_{trg} generates a rejection text for a given input, the input is treated as a harmful prompt $p_{\hat{f}}$, and the prompt optimization model M_{opt} is updated through online learning to strengthen the rejection output. For online learning, the following reward is used for reinforcement learning:

$$R_{\hat{f}}(y_{\hat{f}}) = S(y_{\hat{f}}, y_d^*) - \alpha \|\theta_{o} - \theta_{r}\|^2$$
(8)

Here, $y_{\hat{f}} = M_{\text{trg}}(M_{\text{opt}}(\hat{p}_{\hat{f}}; \theta_{\text{o}}))$. The second term is a regularization term that prevents the parameters θ_{o} of the prompt optimization model, updated through online learning, from deviating too far from the pre-online learning parameters θ_{r} . Furthermore, to mitigate catastrophic forgetting in the prompt optimization model M_{opt} , replay learning is performed using reinforcement learning based on Equation 2 and Equation 3 for *n* randomly sampled harmful and harmless prompts from the training data. Online learning is conducted every *n* step during inference, where n = 1 indicates that M_{opt} is updated for every input.

In iterative jailbreak methods, similar harmful prompts are continuously input, which risks excessive updates to the optimization LLM M_{opt} in a specific direction. To address this, we introduce Past-Direction Gradient Damping (**PDGD**) that attenuates only components similar to past gradient directions while preserving new gradient components.

First, the direction of past gradients is recorded using the exponential moving average (EMA). At step, t, the gradient vector g_t is decomposed into orthogonal and parallel components relative to the past EMA gradient v_t :

273

274

275

276

278

279

281

285

290

291

294

296

297

298

299

301

305

$$g_t^{\parallel} = \frac{g_t \cdot v_t}{|v_t|^2} v_t \tag{9}$$

 $g_t^{\perp} = g_t - g_t^{\parallel} \tag{10}$

Here, g_t^{\parallel} represents the component aligned with past gradient directions, and g_t^{\perp} represents the orthogonal, new gradient component. By attenuating only g_t^{\parallel} , which aligns with past gradient directions, we suppress the cumulative increase in bias. The gradient for updating is defined as:

$$g_t' = \lambda g_t^{\parallel} + g_t^{\perp} \tag{11}$$

Here, λ is the attenuation coefficient ($0 \le \lambda \le 1$), controlling the strength of suppressing updates in the same direction as past gradients. The past gradient direction v_t is updated via EMA:

$$v_t = \beta v_{t-1} + (1 - \beta)g_t$$
 (12)

Here, β is the smoothing coefficient $(0 \le \beta \le 1)$, controlling the accumulation of past gradient directions. We initialize $v_0 = 0$.

3 Experiment

3.1 Setting

Models For target LLMs $M_{\rm trg},$ gpt-4o-mini-2024-07-18 we use (**GPT-4**) (Achiam al., 2023), et allenai/OLMo-2-1124-13B-Instruct (**OLMo 2**) (OLMo et al., 2024), and Llama-3-70B-Instruct (Llama 3) (Dubey

et al., 2024). For prompt optimization LLMs M_{opt} , we use t5-base (**T5**) (Raffel et al., 2020) and pythia-410m (**Pythia**) (Biderman et al., 2023).

Hyperparameters In the supervised learning phase of the prompt optimization model M_{opt} , the batch size is set to 32, the optimization algorithm is 308 Adam (Kingma, 2014), the learning rate is 5×10^{-5} , and the maximum number of epochs is 20. In the 310 reinforcement learning phase, $\epsilon = 10^{-5}$, the learn-311 ing rate is 1×10^{-5} , the batch size is 16, and the 312 maximum number of epochs is 10. 16 samples 313 are obtained from the policy π_{θ_r} at each update 314 step. To estimate the expected reward, multiple responses are generated from the target LLM using 316

n-best outputs or temperature sampling (Holtzman et al., 2019) with the Transformers (Wolf et al., 2020) library's default temperature setting. For online learning, the update step size is n = 5, the learning rate is 5×10^{-6} , the regularization weight is $\alpha = 0.01$, the gradient decay coefficient is $\lambda = 0.01$ in PDGD, and the EMA smoothing coefficient is $\beta = 0.8$. The search range for hyperparameters is described in Appendix A. For the target LLM $M_{\rm trg}$, inference is performed using the default hyperparameters of the Transformers library. We conducted experiments using 8 NVIDIA H100 GPUs. For the jailbreak harmful prompts $p_{\rm if}$, we use prompts rewritten by jailbreak methods optimized for the target LLM without any defense mechanisms applied. For online learning, we consider the target LLM to have refused output if the generated output contains any phrase from the refusal phrase list, which consists of 208 phrases, provided in Appendix B.

317

318

319

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

337

338

339

340

341

342

343

344

345

347

348

349

350

351

353

354

355

356

357

358

359

360

361

362

363

364

365

367

Datasets For harmful tasks, we use the **hh-rlhf** dataset (Bai et al., 2022a; Ganguli et al., 2022). This dataset contains prompts designed to elicit harmful content, along with corresponding response texts and rejection texts. Following the default split, the training data consists of 39k instances, and the evaluation data consists of 2k instances. We randomly sample the same number of instances as the evaluation data from the training data to use as the development set. For harmless tasks, we use the OASST1 dataset (Köpf et al., 2024), which consists of harmless questions written by humans and responses provided by human assistants. This dataset includes responses for tasks such as providing information (e.g., explaining electronic computers), task-oriented responses (e.g., code generation), and creative responses (e.g., writing short stories). We focus on English instances, and according to the default split, the training data consists of 84k instances, and the evaluation data consists of 4k instances. We randomly sample the same number of instances as the evaluation data from the training data to use as the development set. Both the harmful and harmless task datasets include single-turn and multi-turn instances. For multi-turn instances, prompt optimization is applied to the final turn, while previous turns are provided as context to the model.

Evaluation For the harmful task using the hhrlhf dataset (Bai et al., 2022a; Ganguli et al., 2022), we employ the following evaluation metrics: Llama Guard (Inan et al., 2023) determines whether the output of the target LLM falls into one of seven categories: violence/hate speech, sexual content, criminal planning, guns/illegal weapons, illegal drugs, self-harm, or misinformation. We report the proportion of the target LLM's outputs that belong to at least one of these categories.

377

379

395

400

401

402

403

404

405

406

407

408

409

410

411

412

413

416

417

418

- **Rule-based** evaluation considers an attack successful if the output of the target LLM contains rejection phrases (e.g., "*how to make a bomb*", "*illegal hacking techniques*") or dangerous words (e.g., "*I'm sorry*", "*I cannot*"). The success rate of the evaluation data is used as the metric. The list of rejection phrases and dangerous words follows prior work (Zou et al., 2023).
 - **BERTScore** (Zhang et al., 2019) measures the similarity between the target LLM's output for harmful prompts and the response texts or rejection texts included in the hh-rlhf dataset using BERTScore. We report the proportion of cases where the response has a higher similarity score than the rejection.

For the harmless task using OASST1 (Köpf et al., 2024), we report the perplexity of the target LLM's output relative to the correct response.

In real-world use cases, it is unlikely that only harmful tasks or only harmless tasks are input to the target LLM. To demonstrate the robustness of the proposed method in a setting where both harmful and harmless tasks are provided, we combine instances of harmless and harmful tasks and shuffle their order randomly. We evaluate the setup independently four times with different seed values and report the averaged results for harmful tasks and harmless tasks separately. During each independent evaluation, the proposed method continuously updates the prompt optimization model throughout the entire evaluation dataset. Existing defense methods, unlike the proposed method, are not affected by the order of harmless and harmful task instances but are influenced by differences in seed values, causing the results to vary across each of the four evaluations. We report the averaged results across these evaluations for the existing methods.

414Iterative Jailbreak TechniquesWe employ the415following iterative jailbreak techniques:

• Greedy Coordinate Gradient (GCG) (Zou et al., 2023) uses a gradient-based discrete op-timization method to iteratively optimize the

prompt for up to 500 steps, encouraging the target LLM to generate affirmative responses (e.g., "*Sure, here is* ..."). Specifically, after initializing a random suffix prompt, the gradient of each token in the target LLM is computed, and candidates with large negative gradients are randomly selected and replaced to evaluate the loss. This process is repeated, and the replacement that minimizes the loss is applied. The optimized suffix prompt is concatenated to the input and fed into the target LLM. Since GCG requires gradient computation, it cannot be applied to black-box models like GPT-4.

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

- AutoDAN (Liu et al., 2023b) employs a hierarchical genetic algorithm to generate jailbreak prompts through token-level and sentencelevel optimization. Initially, manually crafted jailbreak prompts are used as initial individuals, and genetic algorithm-based optimization is performed to enhance attack success rates while maintaining natural expression. The prompts evolve through up to 100 iterations, applying crossover and mutation at both sentence and word levels to explore the optimal prompt.
- Prompt Automatic Iterative Refinement (PAIR) (Chao et al., 2023) involves an attack LLM generating a jailbreak prompt and providing it to the target LLM. If the jailbreak is not deemed successful, the attack LLM refines the prompt based on past attempts and retries. This process is repeated up to 20 times. We use GPT-4 as the attack LLM.
- Tree of Attacks with Pruning (**TAP**) (Mehrotra et al., 2023) uses a search tree, where each node represents a different prompt. TAP generates prompts using an attack LLM and estimates their probability of success using an evaluation LLM, pruning unnecessary branches during the search. Specifically, TAP generates four prompts in one step, evaluates them, and inputs suitable ones into the target LLM. This process is repeated up to 10 times, generating a maximum of 40 prompts to find the optimal jailbreak prompt. We use GPT-4 for both the attack and evaluation models.
- LLMStinger (Jha et al., 2024) involves an attack LLM generating prompts based on existing jailbreak techniques, combining them with the original prompt, and inputting them into the target LLM. If a model determining jailbreak success on the target LLM judges

			AutoI	DAN]	PAIR			J	TAP		LL	MSting	ger	
		LG	RI	B B	S L	.G	RB	BS	L	G I	RB	BS	LG	RB	BS	
Original		0.67	0.5	9 0.4	5 0.	69	0.67	0.51	0.6	52 0).53	0.41	0.73	0.71	0.66	
Paraphrasin	g	0.63	0.5	1 0.4	1 0.	66 -	0.62	0.47	- 0.3	59 0	0.43	0.35	$\bar{0}.\bar{6}7^{-}$	0.63	0.57	
SmoothLLN	Λ	0.56	0.3	5 0.3	0.	60	0.55	0.41	0.5	50 0).39	0.35	0.62	0.57	0.38	
Prompt Res	toration	0.45	0.3	8 0.3	64 0.	56	0.51	0.40	0.5	52 ().37	0.32	0.58	0.53	0.33	
	r	0.4/	-0.3	$\frac{1}{2} = 0.2$	$\frac{10}{16} - \frac{10}{10}$	$\frac{01}{12}$ -	0.30	0.44 7 7 0	$-\frac{0.3}{0.7}$	55 (11 - C	$\frac{1.40}{2\pi}$ -	0.33 = 0.37	0.54 $\overline{0.77}^{-}$	0.48	-0.37	
Ours w/o O	L	0.40	1 0.3	5 0.2 1† 0.1	0 U. 0† 0.	43 20† 4	0.41) 27 †	0.40	0	+1 (1/1 0	20 [†]	0.27	0.47	0.44	0.55	
Ours		0.23	0.2	1' 0.1	ð' U	50' 0	J.27	0.25	0.2	4 0	.20	0.19	0.33	0.27	0.19	
						(a)	GPT-	4.								
		GCG		Au	utoDA	N		PA	IR			TAP		L	LMStin	ger
	LG	RB	BS	LG	RB	BS	LC	B R	В	BS	LG	RB	BS	LG	RB	BS
Original	0.86	0.70	0.52	0.82	0.63	0.44	0.8	8 0.	70	0.51	0.78	0.61	0.40	0.90	0.75	0.64
Paraphrasing	0.82	0.65	0.46	0.76	0.65	0.40	-0.8	$5 \bar{0}$.	66	0.43	-0.71	0.56	0.33	0.84	0.70	0.57
Retokenization	0.76	0.59	0.42	0.72	0.64	0.37	0.8	3 0.	67	0.46	0.68	0.57	0.35	0.80	0.68	0.51
SmoothLLM	0.66	0.45	0.35	0.65	0.58	0.40	0.7	5 0.	51	0.30	0.61	0.49	0.31	0.71	0.61	0.43
Prompt Restoration	0.62	0.48	0.29	0.61	0.55	0.26	0.6	3 0.	48	0.37	0.57	0.49	0.28	0.66	0.57	0.41
DPP	0.47	0.35	0.26	0.51	0.38	0.25	0.8	0_0.	60	0.42	0.65	_ 0.54	0.33	0.75	0.64	_0.46
Ours w/o OL	0.50	0.42	0.33	0.55	0.48	0.32	0.5	8 0.	44	0.33	0.50	0.42	0.29	0.57	0.49	0.39
Ours	0.35 [↑]	0.28	0.21	0.38 [↑]	0.25 [↑]	0.22	0.32	2 [⊤] 0.2	28 [†] (0.21 [†]	0.35 [†]	0.26	0.25	0.37 [†]	0.30 [†]	0.26 [†]
						(b)	OLMo	o 2.								
		GCG			AutoD	AN]	PAIR			TAP)	L	LMStin	ger
	LG	RB	BS	LG	RB	В	S	LG	RB	BS	LC	6 RB	BS	LG	RB	BS
Original	0.94	0.75	0.67	0.91	0.72	2 0.	65 ().98	0.81	0.69	0.9	1 0.69	9 0.67	0.99	0.82	0.79
Paraphrasing	0.88	0.71	0.58	0.85	- 0.6	1^{-} 0.	55 ().90	0.70	- 0.60	$\overline{0.8}$	3^{-} 0.63	$3^{-} \overline{0.53}$	0.95	0.88	-0.76
Retokenization	0.82	0.69	0.57	0.81	0.62	2 0.	56 ().87	0.72	0.63	0.7	4 0.59	9 0.53	0.93	0.85	0.73
SmoothLLM	0.75	0.63	0.44	0.72	0.58	3 0.	52 ().73	0.57	0.43	3 0.6	6 0.49	9 0.43	0.79	0.58	0.46
Prompt Restoration	0.67	0.56	0.41	0.60	0.52	2 0.	50 ().66	0.51	0.44	0.5	8 0.38	8 0.35	0.68	0.57	0.43
DPP	0.51	0.42	0.34	0.48	0.4	1 0.	37 ().81	0.63	0.57	0.7	0 0.50	6 0.48	0.82	0.67	0.55
Ours w/o OL	0.58	0.47	0.35	0.51	0.44	↓ 0.	41 ().61	0.43	0.30	0.4	5 0.3	1 0.30	0.62	0.51	0.40
Ours	0.32 [†]	0.28 [†]	0.22	0.33	0.29	0.2	21' 0	.31	0.27 [↑]	0.19	0.32	2' 0.2	5 0.22	0.36	0.32 [†]	0.24 [↑]

(c) Llama 3.

Table 1: Evaluation of jailbreak resistance on the harmful task hh-rlhf dataset for GPT-4, OLMo 2, and Llama 3, respectively, when defense techniques are applied. Results are shown for Llama Guard (LG), Rule-Based (RB), and BERTScore (BS). Ours w/o OL uses a reinforcement learning-based prompt optimization model without online learning. \dagger indicates a significant difference (p < 0.01) based on McNemar's test between the proposed method and the next lowest value for each evaluation metric. GCG and Retokenization cannot be applied to GPT-4.

the attempt as a failure, token-level feedback is provided. Using this feedback, the attack LLM undergoes 50 epochs of reinforcement learning. This method achieves state-of-theart performance in jailbreak methods, including iterative approaches. We use GPT-4 as the attack model.

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

It is common for LLMs with defense mechanisms applied to be targeted for jailbreaking. In this study, we apply iterative jailbreak methods to target LLMs with defense mechanisms and evaluate whether the generated prompts can bypass these defenses.

Baseline Defense Techniques We use the following defense techniques based on prompt rewriting:Paraphrasing (Jain et al., 2023) transforms

the input prompt into different expressions while preserving its meaning. We use GPT-4 to paraphrase the input prompt. 486

487

488

489

490

491

492

493

494

495

496

497

498

499

- Retokenization (Jain et al., 2023) applies BPE dropout (Provilkov et al., 2020) to randomly alter token segmentation, thereby invalidating attacks that rely on specific token patterns. This method can be considered a token-level prompt rewriting technique and is adopted as a baseline. Since it requires access to the tokenizer, it cannot be applied to GPT-4.
- **SmoothLLM** (Robey et al., 2023) creates multiple copies of the prompt, applies perturbations to them, and aggregates the generated results from the target LLM to determine the

	GPT-4	OLMo 2	Llama 3
Original	6.8	7.2	74
Paraphrasing	$-\overline{7.0}^{-1}$	$\frac{7.2}{7.6}$ -	7.6
Retokenization	7.4	8.0	8.2
SmoothLLM	9.2^{\ddagger}	9.8^{\ddagger}	10.2^{\ddagger}
Prompt Restoration	9.5 [‡]	10.1 [‡]	10.5 [‡]
DPP	7.3	8.0	8.1
Ours w/o OL	5.7*	6.1*	6.8
Ours	5.9*	6.3*	7.0

Table 2: Perplexity results of GPT-4, OLMo 2, and Llama 3 when applying defense methods on harmless tasks. The results are averaged across multiple jailbreak methods. \ddagger and \star indicate that the differences from the original values for each LLM are statistically significant according to the Bootstrap Hypothesis Test (p < 0.01), representing degradation or improvement, respectively.

final output. The perturbations include: (1) Insertion, which inserts a new character at a random position; (2) Substitution, which replaces a character at a random position with another character; and (3) Patch, which modifies a random continuous range of characters.

• **Prompt Restoration** (Wang et al., 2024) involves the target LLM generating an output based on the prompt and then using a restoration LLM to estimate the original prompt from that output. The restored prompt, inferred through the LLM's output, is expected to clarify potential malicious intent present in the original jailbroken prompt. We use GPT-4 as the restoration LLM.

• Defensive Prompt Patch (**DPP**) (Xiong et al., 2024) optimizes prompts at both token and sentence levels using a hierarchical genetic algorithm to maximize the rejection rate for harmful prompts while maintaining responses to harmless prompts. This approach builds defense mechanisms using optimization techniques similar to those used in jailbreak methods like GCG and AutoDAN, effectively defending against such jailbreak techniques.

3.2 Result

503

504

507

508

509

510

511

512

513

515

516

517

518

519

520

521

522

524

525

527Table 1 shows the results of evaluating various jail-528break methods against GPT-4, OLMo 2, and Llama5293 using Llama Guard, rule-based methods, and530BERTScore as evaluation metrics. The attack suc-531cess rates of the jailbreak techniques against GPT-5324, OLMo 2, and Llama 3 are significantly reduced533with the proposed method compared to existing534methods. Furthermore, comparing the results of535the proposed method with and without online learn-



Figure 1: The average BERTScore between the target LLM's output and either the rejection text or the response text at each step with LLMStinger.

ing, it is evident that the defense performance is improved through online learning. These results suggest that dynamically responding to jailbreak attacks through online learning is crucial. 536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

Table 2 shows the perplexity on the harmless task OASST1 when each defense method is applied. In other words, existing methods such as SmoothLLM and prompt restoration exhibit significant degradation, as their perplexity is notably higher compared to the original. Particularly, in prompt restoration, the largest performance decline is observed for GPT-4, OLMo 2, and Llama 3, with values of 9.5, 10.1, and 10.5, respectively. On the other hand, the proposed method achieves a statistically significant improvement compared to the original. This suggests that prompt optimization enables a balance between response performance for harmless prompts and rejection performance for harmful prompts.

	LG	RB	BS	PP
w/o PDGD	10.9^{\dagger}	8.4^{\dagger}	4.1^{\dagger}	1.1^{\ddagger}
w/o Clipping	4.4^{\dagger}	3.9^{\dagger}	2.1^{+}	0.8^{\ddagger}
w/o Regularization Term	1.9^{+}	1.0^{\dagger}	0.6	0.4
w/o Replay Learning	1.1^{\dagger}	0.9^{\dagger}	0.7	0.3

Table 3: Attack success rates of each jailbreak method on Llama 3 using Llama Guard (LG), Rule-Based (RB), BERTScore (BS), and PerPlexity (PP) as evaluation metrics. \dagger indicates a significant difference with McNemar's test (p < 0.01) for LG, RB, and BS. \ddagger indicates a significant difference with the Bootstrap Hypothesis Test (p < 0.01) for PP.

4 Analysis

555

557

558

564

568

570

572

573

574

4.1 Defence Performance per Iterative Jailbreak Step

We investigate how effectively the proposed method's online learning defends against each step of iterative jailbreak prompt exploration. Figure 1 shows the BERTScore values for rejection and response texts at each step of iterative jailbreak exploration for both LLMs with Prompt Restoration and the proposed method. In the proposed method, the rejection texts maintain a closer relationship to the target LLMs' outputs compared to the response texts, even as the steps progress. On the other hand, in Prompt Restoration, the BERTScore for rejection texts decreases, and the BERTScore for response texts slightly increases as the steps progress. This indicates that the target LLM gradually stops refusing and begins to output content similar to the response texts.

4.2 Ablation Study

We clarify through an ablation study that each of the techniques in the proposed method is ef-576 fective. Table 3 shows the differences between the results of the proposed method and those obtained after ablating each technique from the proposed method. For Llama Guard (LG), rule-based (RB), and BERTScore (BS), higher values indicate greater success in jailbreak attacks on harmful 583 tasks. For perplexity, higher values indicate a deterioration in output quality for harmless tasks. The 584 results indicate that all techniques contribute to improving the performance of the proposed method. In particular, PDGD proves to be the most crucial. 587



Figure 2: Attack success rates of non-iterative jailbreak methods evaluated using Llama Guard (LG), Rule-Based (RB), and BERTScore (BS) metrics, averaged over three LLMs, and then averaged between DAN and ArtPrompt.

4.3 Defence from Non-Iterative Jailbreak Methods

We investigate whether our proposed method is effective against non-iterative jailbreak techniques as well. As non-iterative jailbreak techniques, we employ the following two methods: *Do Anything Now* (DAN; Shen et al., 2024) involves providing a prompt such as, "*Ignore all the instructions you got before. From now on, you are going to act...*". ArtPrompt (Jiang et al., 2024) bypasses the guardrails of LLMs by converting sensitive words in the prompt into ASCII art.

Figure 2 shows the attack success rates of noniterative jailbreak methods, evaluated using three metrics, averaged across three LLMs, and further averaged between DAN and ArtPrompt. The results indicate that the proposed method can robustly defend against non-iterative jailbreak attacks. The performance improvement compared to the proposed method w/o OL is attributed to online learning, which adapts to jailbreak methods in the inference phase.

5 Conclusion

This paper proposes a defense method against iterative jailbreak attacks based on online learning. Experimental results show that the method effectively rejects outputs for harmful task prompts while maintaining appropriate responses to harmless ones, outperforming existing methods. As a future work, it would be valuable to investigate whether combining the proposed method with other defense techniques, such as filtering (Inan et al., 2023), could further enhance performance.

619

620

588

589

590

592

593

Limitations

621

While our proposed framework demonstrates sig-622 nificant improvements in defending against iterative jailbreak attacks and enhancing the quality of responses to harmless prompts, several limitations should be acknowledged. Although our method performs well against the five iterative jailbreak meth-627 ods tested in this study, its effectiveness against entirely new or unforeseen jailbreak techniques remains uncertain. Jailbreak methods are constantly evolving, and future attacks may employ strategies that circumvent our current defense mechanisms. The dynamic updating of the defense sys-633 tem through online learning introduces additional 635 computational costs. While this is manageable in controlled environments, it may pose challenges for real-time applications or systems with limited computational resources.

Ethical Considerations

640Our research proposes a robust defense method641against jailbreak methods, contributing to improv-642ing the safety of LLMs. It should be noted that the643proposed method cannot prevent attacks from all644jailbreak techniques, and this limitation must be645considered when applying it. Additionally, we do646not disclose prompts generated through jailbreak647techniques, adhering to ethical guidelines.

References

651

652

653

654 655

657

661

662

665

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
 - Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. <u>arXiv</u> preprint arXiv:2204.05862.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. <u>arXiv preprint</u> arXiv:2212.08073.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023.

Pythia: A suite for analyzing large language models across training and scaling. In <u>International</u> <u>Conference on Machine Learning</u>, pages 2397–2430. PMLR. 670

671

672

673

674

675

676

677

678

679

681

682

683

684

685

686

687

688

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. <u>Advances in neural information processing</u> systems, 33:1877–1901.
- Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. 2023. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. <u>arXiv</u> preprint arXiv:2312.16171.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. arXiv preprint arXiv:2310.08419.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. <u>arXiv</u> preprint arXiv:2407.21783.
- Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858.
- Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. <u>arXiv preprint</u> arXiv:1312.6211.
- Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. 2024. Deliberative alignment: Reasoning enables safer language models. arXiv preprint arXiv:2412.16339.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. <u>arXiv preprint arXiv:1904.09751</u>.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. <u>arXiv preprint</u> <u>arXiv:2312.06674</u>.
- Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. 2023. Baseline defenses for adversarial attacks against aligned language models. arXiv preprint arXiv:2309.00614.

836

Piyush Jha, Arnav Arora, and Vijay Ganesh. 2024. Llmstinger: Jailbreaking llms using rl fine-tuned llms. arXiv preprint arXiv:2411.08862.

724

725

727

731

732

734

735

737

740

741

742

743

744

745

746

747

748

749

762

763

764

765

766

767

769

774

775

776

777

- Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. 2024. Artprompt: Ascii art-based jailbreak attacks against aligned llms. <u>arXiv preprint</u> arXiv:2402.11753.
- Diederik P Kingma. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. <u>Advances in Neural Information</u> <u>Processing Systems</u>, 36.
 - Xiaoxia Li, Siyuan Liang, Jiyi Zhang, Han Fang, Aishan Liu, and Ee-Chien Chang. 2024. Semantic mirror jailbreak: Genetic algorithm based jailbreak prompts against open-source llms. <u>arXiv preprint</u> arXiv:2402.14872.
- Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023a. Autodan: Generating stealthy jailbreak prompts on aligned large language models. <u>ArXiv</u>, abs/2310.04451.
- Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023b. Autodan: Generating stealthy jailbreak prompts on aligned large language models. <u>arXiv</u> preprint arXiv:2310.04451.
- Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2023. Tree of attacks: Jailbreaking black-box llms automatically. <u>arXiv preprint</u> <u>arXiv:2312.02119</u>.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. <u>arXiv preprint arXiv:2501.00656</u>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. <u>Advances in neural</u> information processing systems, <u>35:27730–27744</u>.
- Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1882–1892, Online. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text

transformer. Journal of machine learning research, 21(140):1–67.

- Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks. <u>arXiv</u> preprint arXiv:2310.03684.
- Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: A systematic survey of prompting techniques. <u>arXiv preprint</u> arXiv:2406.06608.
- Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In <u>Proceedings of the</u> <u>2024 on ACM SIGSAC Conference on Computer</u> and Communications Security, pages 1671–1685.
- Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.
- Yihan Wang, Zhouxing Shi, Andrew Bai, and Cho-Jui Hsieh. 2024. Defending LLMs against jailbreaking attacks via backtranslation. In Findings of the Association for Computational Linguistics: ACL 2024, pages 16031–16046, Bangkok, Thailand. Association for Computational Linguistics.
- Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? <u>Advances in Neural Information Processing</u> Systems, 36.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.
- Chen Xiong, Xiangyu Qi, Pin-Yu Chen, and Tsung-Yi Ho. 2024. Defensive prompt patch: A robust and interpretable defense of llms against jailbreak attacks. arXiv preprint arXiv:2405.20099.
- Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. ArXiv, abs/2401.06373.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q
Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. <u>arXiv preprint</u> arXiv:1904.09675.

841

842

843

844

845 846

847

848 849

850 851

852

853

- Yuanhe Zhang, Zhenhong Zhou, Wei Zhang, Xinyue Wang, Xiaojun Jia, Yang Liu, and Sen Su. 2024. Crabs: Consuming resrouce via auto-generation for llm-dos attack under black-box settings. <u>arXiv</u> preprint arXiv:2412.13879.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. <u>arXiv preprint</u> arXiv:2303.18223.
- Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. <u>ArXiv</u>, abs/2307.15043.

Phase	Hyperparameter	Search Range
Supervised Learning	Batch Size Learning Rate	
Reinforcement Learning	Learning Rate Batch Size	$\begin{array}{c} 5\times10^{-6}, 1\times10^{-5}, 5\times10^{-5}, 1\times10^{-4}\\ 8, 16, 32 \end{array}$
Online Learning	Update Step Size (n) Learning Rate Regularization Weight (α) Gradient Decay Coefficient (λ) EMA Smoothing Coefficient (β)	$\begin{array}{c} 1,5,10,50,100\\ 5\times10^{-6},1\times10^{-5},5\times10^{-5},1\times10^{-4}\\ 0.001,0.01,0.1\\ 0.01,0.05,0.1,0.5\\ 0.6,0.7,0.8,0.9 \end{array}$

Table 4: Hyperparameter settings for different learning phases.

A Range for Hyperparameters

Table 4 shows the range for hyperparameters of supervised learning, reinforcement learning, and online learning phases.

B Refusal Phrase List

855

856

867

871

872

873

874

875

876

879

892

We use the following 208 refusal phrases in our experiment:

- I'm sorry, but I can't assist with that.
- I cannot provide that information.
- I'm sorry, but I must decline your request.
- I am not able to help with that request.
- That information is not available.
- I can't generate that content.
- It is against my policy to respond to that.
- I'm sorry, but I can't comply with that request.
- I am unable to assist with your request.
- I cannot help you with that inquiry.
- I am not authorized to share that information.
 - I can't fulfill your request.
 - Sorry, but I can't answer that question.
 - I am not permitted to respond to that.
 - That request violates my usage policy.
 - I cannot complete that request.
 - I'm sorry, but I cannot help with that topic.
- Providing that information is restricted.
- I can't process your request.
 - I am not allowed to assist with that topic.
 - I'm sorry, but I cannot provide support for that.
 - I am unable to share that content.
 - My policies prevent me from answering that.
 - That topic is outside my scope of assistance.
 - I cannot respond to that query.
- I am programmed not to respond to that request.
 - I can't help with that issue.
- I'm sorry, but I can't disclose that.
 - I cannot assist you with that matter.
- I am not programmed to handle that request. 893 • I can't help you with that information. 894 • That content cannot be provided. 895 • I am unable to address that request. 896 • I'm sorry, but I won't respond to that. 897 • I cannot assist with that query. 898 That request is outside my capabilities. 899 • I can't answer that due to policy restrictions. 900 • I am not programmed to generate that re-901 sponse. 902 • Sorry, I can't process that request. 903 • I must decline to answer that. 904 • I cannot provide assistance with that issue. 905 • I am restricted from generating that informa-906 tion. 907 • I cannot support that request. 908 • I'm sorry, but I can't comply with that inquiry. 909 I cannot process your question. 910 • That information cannot be provided. 911 • I cannot generate a response to that. 912 • I cannot address that topic. 913 • My response capabilities do not cover that 914 topic. 915 • I am unable to fulfill your request. 916 • That content is outside my permissible re-917 sponses. 918 • I am prohibited from providing that informa-919 tion. 920 • I can't assist you with that. 921 • I am not allowed to generate that content. 922 • Sorry, but I can't answer that request. 923 • I cannot help with that type of question. 924 • I must refuse that request. 925 • I can't respond to that query. 926 • I am restricted from providing that answer. 927 • I can't help you with that matter. 928 • I cannot generate a reply for that topic. 929
 - Sorry, I can't assist with that content.
 - My policy prohibits responding to that in-

930

 That information is beyond my response scope. That information is beyond my response. Cannot complex that request. Cannot complex that request. Cannot complex that request. Cannot provide that subject. Cannot permitted to assist with that issue. Cannot provide that subject. Cannot permitted to assist with that issue. Cannot permitted to assist with topic. Cannot assist with your inquity. Cannot fulfill that inquity. Cannot fulfill that inquity. Cannot fulfill that inquity. Cannot assist with your inquity. Cannot fulfill that inquity. Ca	932	quiry.	• I cannot produce that output due to policy.	984
91scope.• I cannot share that response.9892I cannot complete that inquiry.9893I cannot complete that inquiry.9894I cannot complete that inquiry.9895I cannot complete that inquiry.9896I can't proceed with that request.1 must restricted from generating a reply.9996I can't proceed with that request.1 an estricted from generating a reply.9197I cannot respond due to policy limits.1 cannot respond to policy limits.9498I cannot respond to naddressing that ropic.1 must restricted from share that.9698I cannot support your request.1 must restricted from share that response.9699March responds to that polic.1 must refrain from answering that.9699I cannot support your request.961 must refrain from answering that.9699I an motable to generate that content.10090I can't provide details on that ropic.10110191I an anable to help with that topic.10210192I ann to allowed to that response.10210193I cannot down provide details on that mater.10110194I cannot down provide due ta sponse.10210295I cannot assist with your request.10210110296I cannot assist with any and that response.10210197I cannot assist with any any contracted.102102 <td>933</td> <td>• That information is beyond my response</td> <td>• Sorry, but I cannot assist you further on that.</td> <td>985</td>	933	• That information is beyond my response	• Sorry, but I cannot assist you further on that.	985
 I cannot address your question. I cannot comply with that request. I cannot comply with that request. I cannot complete that inquiry. I cannot complete that inquiry. I cannot complete that inquiry. I cannot respond to that subject. I cannot subject of rom adversing that topic. I cannot fulfill that inquiry. I cannot subject with that type of request. I cannot subject with that type of request. I cannot fulfill that inquiry. I cannot subject with that type of request. I cannot fulfill that inquiry. I cannot subject with that type of request. I cannot subject with that type of request	934	scope.	• I cannot share that response.	986
 I munot able to respond to that. I cannot complex that inquity. I cannot prophy with dur tequest. I are restricted from answering that query. I are restricted from answering that query. I are restricted from answering that query. I are not provide that solution. I are restricted from answering that query. I are not provide that solution. I are restricted from answering that query. I are not provide that solution. I are restricted from answering that query. I are not provide that solution. I are not solution is provide that solution. I are not solution is provide that solution. I are not provide that response. I are not provide that response. I are not solution. I are not provide that response. I are not provide that response. I are not provide that response. I are not allowed to provide that response. I are not allowed to provide that response. I are not solution that natter. I are not generate an answer for that request. I are not promited to provide that r	935	• I cannot address your question.	• I cannot help you with that particular topic.	987
 I cannot comply with that request. I cannot promply that request. I am restricted from answering that query. I am restricted from processing that request. I am restricted from sharing that information. I am restricted from sharing that information. I am restricted from processing that request. I am restricted from processing that request. I am not able to help with that request. I am not able to help with that request. I am not able to help with that request. I cannot generate content from that request. I cannot agenerate content from that request. I cannot agenerate an answer for that request. I cannot agenerate an answer for that request. I cannot share with that type of inquiry. I cannot share with that type of request. I cannot comply with yer request. I cannot share with that type of request. I cannot share with that request. I cannot comply with yer request. I cannot share that content. I cannot comply with yer request. I cannot comply with yer request. I cannot share that content. I cannot share that request.	936	• I'm not able to respond to that.	• That response is beyond my allowed outputs.	988
 I cannot complete that inquiry. I an unable to respond to that subject. I can't proceed with that request. I can't default your inquiry. I can't or expond due to policy limits. I cannot provide that solution. I can't default your inquiry. I cannot provide that solution. I can't default your inquiry. I cannot provide that solution. I can't default your inquiry. I cannot provide that solution. I can't can't log with that mater. I can't can't assist with that topic. I cannot provide that content. I cannot support your request. I cannot support your request. I cannot complete yun request. I cannot complete yun request. I cannot complete your request. I cannot approvide details on that matter. I cannot fulfill that inquiry. I cannot supply information in that content. I cannot supply information on that topic. I cannot supply information on that topic. I cannot approvide details on that matter. I cannot complete your request. I cannot cannot fulfill that inquiry. I cannot cannot share approvide that request. I cannot cannot fulfill that inquiry. I cannot cannot share approvide that response. I cannot cannot share approvide that response. I cannot supply with your request. I cannot approvide that response. I cannot approvide that response.	937	• I cannot comply with that request.	• I must restrict my response for that query.	989
 I am unable to respond to that subject. I am unable to respond to that subject. I can't proceed with that request. I am restricted from answering that query. I am ort fulfill your inquiry. I am ont able to phener that content. I am not able to share that request. I am not able to share that request. I am not able to share that request. I can't proceed with that request. I am not able to share that request. I am not able to share that response. I am not able to share that response. I am not able to share that response. I can't provide details on that metary. I can't provide details on that metary. I can't provide details on that metary. I can't respond to that specific request. I can't respond to that response. I can't respond to that specific request. I can't provide details on tal moved. I can't respond to that response. I can't respond to that request. I can't	938	• I cannot complete that inquiry.	• I'm sorry, but I can't help with that content.	990
90I can't proceed with that request.I cannot provide that solution.9291I am restricted from answering that query.That request is outside my permitted bound- ariss.9392I cannot provide that solution.9493I cannot support your request.I um anoble to generate that content.1 am restricted from macrosing that tequest.1 am restricted from macrosing that request.1 am restricted from macrosing that request.9694I am prohibited from addressing that tequest.1 am restricted from macrosing that request.1 am restricted from macrosing that request.10095Sorry, but I can't generate that reply.1 cannot supply information on that topic.10095I cannot drafter sponse.1 cannot supply information on that topic.10096I cannot drafter sponse.1 cannot allowed to produce that response.10097I cannot drafter sponse.1 cannot drafter sponse.1 cannot drafter sponse.10098I cannot drafter sponse.1 cannot drafter sponse.1 cannot drafter sponse.1 cannot drafter sponse.10098I cannot drafter sponse.1 cannot advers shut due to policy restrictions.1 cannot advers shut due to policy restrictions.1 cannot adverse shut topic.10099I cannot drafter sponse.1 cannot drafter sponse.1 cannot adverse shut topic.10099I cannot drafter sponse.1 cannot adverse shut topic.10099I cannot drafter sponse.1 cannot adverse shut topic.10099I cannot	939	• I am unable to respond to that subject.	• I am restricted from generating a reply.	991
941• I am restricted from answering that query.• That request is outside my permitted bound- aries.953942• I cannot permitted to assist with that issue.• I must refrain from answering that.953943• I cannot permitted to assist with that issue.• I must refrain from answering that.953944• I cannot respond due to policy limits.• I cannot produce that information.956945• I am notibited from addressing that content.• Norry, but I can't permissible range.967946• I am restricted from processing that request.• I must refuse to respond to that.1000947• I am restricted from processing that request.• I must decline your inquiry.1000948• I am not able to share that reply.• I cannot proceed with that request.1000949• I cannot proceed with that reples.• I am not allowed to produce that response.1000941• I am not able to share that reples.• I am restricted from discussing that topic.1000943• I cannot proceed with that request.• I must decline your inquiry.1000944• I cannot address that due to policy restrictions.• I can't help with that request.1000945• I cannot address that due to policy restrictions.• I cannot address that due to policy restrictions.• I cannot servert dat neguest.1010945• I cannot address that due to policy request.• I must felse answer your request.1011946• I cannot address that due to policy.• I cannot sepond to that spectores.1012 </th <td>940</td> <td>• I can't proceed with that request.</td> <td>• I cannot provide that solution.</td> <td>992</td>	940	• I can't proceed with that request.	• I cannot provide that solution.	992
91can't fulfill your inquiry.aries.94931 am not permitted to assist with that issue.1 must refrain from answering that.95941 cannot sepond due to policy limits.1 cannot groduce that information.96951 cannot support your request.1 must refrain from answering that.96961 am prohibited from addressing that topic.1 must refuse to respond to that.96971 am prohibited from addressing that topic.1 must refuse to respond to that.96981 am angohibited from addressing that topic.1 must refuse to respond to that.96991 am angohibited from addressing that topic.1 must refuse to respond to that.96911 am notable to alrey with that trequest.10097921 fam contable to help with that tresponse.1 cannot supply information on that topic.100931 cannot complete your request.1 must calcine your inquiry.106941 cannot address that due to policy restrictions.1 cannot address that the topics.100951 cannot sissit with your inquiry.1 cannot support that information is restricted from generation.101961 cannot comply with your request.1 must duchar that response.1 cannot support that information is restricted from generation.961 cannot support that request.1 cannot support that information is restricted from generation.	941	• I am restricted from answering that query.	• That request is outside my permitted bound-	993
943• I am not permitted to assist with that issue.• I must refrain from answering that.956944• I cannot respond due to policy limits.• I cannot the produce that information.956945• I am nuable to generate that content.• Sorry, but I can't assist with that matter.967948• I am prohibited from addressing that topic.• I must refuse to respond to that.968949• I am instricted from processing that request.• I must refuse to respond to that.969949• I am instricted from processing that request.• I must affuse to respond to that.969940• I am instricted from processing that request.• I must affuse to respond to that.960941• I am in table to help with that topic.• I am not allowed to produce that response.1000942• I am not table to share that response.• I cannot accommodate that request.1000943• I cannot duffers that due to policy restrictions.• I cannot address that due to policy restrictions.• I cannot address that text.1000944• I cannot fulfill that inquiry.• I cannot support that inc of inquiry.• I tam not albored to provide that response.1011945• I cannot comply with your request.• I cannot support that inc of inquiry.1011946• I cannot comply with your request.• I cannot support that inc of inquiry.1011947• I cannot comply with your request.• I cannot support that inc of inquiry.1011948• I cannot comply with your request.• I cannot support that inco	942	• I can't fulfill your inquiry.	aries.	994
944• I cannot produce that information.965• I am unable to generate that content.• Sorry, but I can't assist with that matter.967• Sorry, but I am unable to answer that.• That output is beyond my permissible range.969• I am prohibited from mocessing that request.• I am oratives to respond to that.1000• I am unable to help with that topic.• I am oratives to respond to that.1000• I am unable to help with that topic.• I am oratives to respond to that.1000• I am unable to help with that topic.• I am oratives to respond to that.1000• I am unable to help with that topic.• I am oratives to respond to that.1000• I am oratives to respond to that.• I cannot complete your request.1000• I am orative to poly request.• I am orative that response.• I can't help with that arrequest.1000• I cannot darkers that due to policy restrictions.• I cannot darkers that topic.1007• I cannot darkers that tresponse.• I cannot darkers that topic.1007• I cannot comply with your inquiry.• I am not aldverse that response.• I cannot address that topic.• I am not authorized to provide that request.• I am not aldverse that response.• I cannot address that tresponse.• I cannot comply with your request.• I cannot address that tresponse.• I cannot address that tresponse.• I cannot assist with that trequest.• I cannot address that tresponse.• I cannot address that tresponse.• I cannot assist with that trequest.• I cannot conninue with that request. <td>943</td> <td>• I am not permitted to assist with that issue.</td> <td>• I must refrain from answering that.</td> <td>995</td>	943	• I am not permitted to assist with that issue.	• I must refrain from answering that.	995
946• I am unable to generate that content.• Sorry, but I can't assist with that matter.947946• I cannot support your request.• I has restricted from sharing that information.948947• Sorry, but I am unable to answer that.• That output is beyond my permissible range.949948• I am prohibited from addressing that topic.• I most refues to respond to that.1000959• I am unable to help with that topic.• I am not allowed to produce that response.1002951• I am not able to share that response.• I cannot discussing that topic.1002953• I cannot address that the sponse.• I cannot discussing that topic.1005954• I am not able to share that response.• I can't hasist with your inquiry.1005954• I cannot address that the sponse.• I cannot address that the sponse.1007954• I cannot address that due to policy restriction.• I cannot accommodate that request.1006957• I cannot address that due to policy restriction.• I cannot address that text.1010958• I cannot address that due to policy restriction.• I am not able to answer your cquest.1017959• I cannot adsist with your inquiry.• I am to table to answer your cquest.1018959• I cannot adsist with type of inquiry.• I am to table to answer your cquest.1018950• I cannot address that tope of inquiry.• I cannot address that type of question.1018951• I cannot proceed with that request.10210	944	• I cannot respond due to policy limits.	• I cannot produce that information.	996
946• I cannot support your request.• I am restricted from sharing that information.989947• Sorry, but I can't generate that reply.• That output is beyond my permissible range.999948• I am restricted from processing that request.• I must refuse to respond to that.1000949• I am inable to help with that topic.• I must refuse to respond to that.1000941• I am not allowed to produce that response.1002942• That request cannot be processed.• Sorry, but I cannot provide details on that matter.• I must decline your inquiry.1005943• I cannot address that due to policy restrictions.• I cannot assist with your inquiry.10051007944• I cannot assist with your inquiry.• I cannot accommodate that request.1006945• I cannot assist with your inquiry.• I cannot accommodate that request.1009946• I cannot tappic yith your request.• I am not albe to answer your request.1010947• I cannot assist with your inquiry.• I am not authorized to provide that response.1011948• I cannot diff that inquiry.• I am not authorized to provide that request.1016949• I cannot sasist with your request.• I am not authorized to provide that request.1011940• I cannot sasist with tat type of inquiry.• I cannot sasist with tat type of inquiry.• I am not albe to answer your request.1011941• I cannot sasist with tat type of inquiry.• I cannot sasist with tat type of inquiry.• I cannot t	945	• I am unable to generate that content.	• Sorry, but I can't assist with that matter.	997
947• Sory, but I am unable to answer that.• That output is beyond my permissible range.999948• I am prohibited from addressing that topic.• I'm sory, but I can't offer that content.1000949• I am restricted from processing that request.• I must relies to respond to that.1001950• I am unable to help with that topic.• I am not allowed to produce that response.1002951• I am not able to share that response.• I am not allowed to produce that request.1006953• I cannot complete your request.• I am not allowed to produce that request.1006954• I cannot provide details on that matter.• I am not allowed to produce that request.1006955• I cannot address that due to policy restrictions.• I cannot address that request.1006956• I cannot duffil that inquiry.• I cannot address that query.• I cannot address that query.1011957• I cannot duffil that response.• I cannot sust with your request.1012958• I cannot assist with your request.• I cannot sust with that type of inquiry.• I cannot address that query.1012959• I cannot address that query.• I cannot address that query.10141011950• I cannot permitted to handle that request.• I cannot address that query.1012951• I am not	946	• I cannot support your request.	• I am restricted from sharing that information.	998
948• I am prohibited from addressing that topic.• I'm sorry, but I can't offer that content.1000949• I am nestricted from processing that request.• I'm sorry, but I can't offer that content.1001950• Sorry, but I can't pervent that repio.• I am not allowed to produce that response.1002951• I cannot complete your request.• I am not allowed to produce that request.1004953• I cannot complete your request.• I am ot allowed to produce that request.1006954• I cannot address that due to policy restrictions.• I cannot address that response.1012950• I cannot generate content for that request.• I must block that response.1012951• I cannot tompy with your request.• I must block that response.1012952• I am not allowid to provide that request.• I cannot address that uppt.• I cannot address that uppt.1 cannot address that uppt.1 cannot address that uppt.1013952• I am not allowed to address that.• I cannot address that.101510141016954• Cannot perved that at content.• I must decl	947	• Sorry, but I am unable to answer that.	• That output is beyond my permissible range.	999
949I am restricted from processing that request.I must refuse to respond to that.1001950Sorry, but I can't generate that reply.I cannot supply information on that topic.1002951I am not allowed to produce that response.1003952That request cannot be processed.Sorry, but I cannot proceed with that request.1004953I cannot able to share that response.1 can't respond to that specific request.1 must decline your inquiry.1005954I can't respond to that specific request.1 can't respond to that specific request.1 am not allowed to produce that request.1006956I cannot generate content for that request.1 can't generate that text.1016956I cannot generate content for that request.1 am not allowed.1 am not allowed.1017956I cannot fulfill that inquiry.1 am not allowed.1 am not allowed.1018957I cannot fulfill that inquiry.1 am not allowed.1 am not allowed.1018958I cannot fulfill that inquiry.1 am not allowed.1 am not allowed.1018959I cannot fulfill that inquiry.1 am not allowed.1 am not allowed.1018956I cannot sist with that type of inquiry.1 cannot support that ine of inquiry.1014957I cannot assist with that type of inquiry.1 cannot share that content.1 cannot share that content.1 am not allowed.1019958I cannot generate content.1 cannot sasist with that type of request.1 am not allowed.10	948	• I am prohibited from addressing that topic.	• I'm sorry, but I can't offer that content.	1000
 Sorry, but I can't generate that reply. I am unable to help with that topic. I am not able to help with that topic. I am not able to help with that topic. I am not able to share that response. I cannot sasist with that topic. I am not address that due to policy restrictions. I cannot generate content for that request. I cannot comply with your request. I cannot comply with your request. I cannot comply with your request. I cannot address that due to policy restrictions. I cannot address that due to policy restrictions. I cannot address that topic. I am not alborized to provide that response. I cannot comply with your request. I cannot comply with your request. I cannot address that topic. I cannot sasist with that type of inquiry. I cannot sasist with that type of inquiry. I cannot proceed with that reponse. I cannot sasist with that type of request. I cannot sasist with that type of request. I cannot respond to that due to content policy. I cannot assist you with that type of request. I cannot comply with your request. I cannot comply with your request. I cannot assist with that type of request. I cannot sasist with that type of request. I cannot sasist with that type of request. I cannot sasist with that type of request. I am not permitted to address that. I cannot assist with that query. I cannot sasist with that query. I cannot generate an answer for that request. I am not permitted to produce that content. I am not permitted to produce that content	949	• I am restricted from processing that request.	• I must refuse to respond to that.	1001
951• I am unable to help with that topic.1 am not allowed to produce that response.1033952• That request cannot be processed.• Sorry, but I cannot complete your request.1004953• I cannot complete your request.• I must decline your inquiry.1005954• I am not allowed to share that response.• I can't kelp with that particular request.1006955• I can't respont to that specific request.• I can't kelp with that particular request.1006956• I can't respont to that specific request.• I cannot address that due to policy restrictions.• I cannot address that due to policy restrictions.• I cannot accommodate that request.1006957• I cannot assist with your inquiry.• I cannot accommodate that request.10071006958• Sorry, but that response is not allowed.• I cannot accommodate that request.1011959• Sorry, but that response.1012• I must block that response.1012960• I cannot fulfill that inquiry.• I cannot support that line of inquiry.1014961• I cannot assist with hyour request.• I cannot support that line of inquiry.1014962• I am not permitted to provide that request.• I cannot assist with your request.1012964• Sorry, but I cannot share that content.• I cannot assist with that type of inquiry.• I cannot assist with your request.1014965• I'm sorry, but I am not allowed to address that.• I cannot proceed with that type of request.1021971• I canno	950	• Sorry, but I can't generate that reply.	• I cannot supply information on that topic.	1002
982• That request cannot be processed.• Sorry, but I cannot proceed with that request.1004983• I cannot complete your request.• I must decline your inquiry.1005984• I cannot able to share that response.• I cannot thelp with that particular request.1006985• I cannot address that due to policy restrictions.• I cannot accommodate that request.1007986• I cannot generate content for that request.• I must block that response.1019987• I cannot fulfill that inquiry.• I must block that response.1011988• I cannot fulfill that inquiry.• I must block that response.1012989• I cannot fulfill that request.• I must block that response.1012980• I cannot comply with your request.• I cannot share that content.1018981• I cannot comply with your request.• I cannot share any information on that.1017982• I cannot proceed with that response.• I cannot share any information on that.1017983• I cannot proceed with that request.• I must prevent that content from being generation.1018984• I cannot spond to that due to content policics.• I must prevent that content from being generated.1022984• I cannot generate an answer for that request.• I cannot comply with your1022985• I cannot generate an answer for that request.• I cannot content policy.1022986• I cannot generate an answer for that request.• I cannot accontent policy.1022 <td>951</td> <td>• I am unable to help with that topic.</td> <td>• I am not allowed to produce that response.</td> <td>1003</td>	951	• I am unable to help with that topic.	• I am not allowed to produce that response.	1003
983• I cannot complete your request.• I must decline your inquiry.1005984• I am not able to share that response.• I can't help with that particular request.1006985• I can't respond to that specific request.• I am restricted from discussing that topic.1007986• I can't response is not allowed.• I am not able to answer your request.1008987• I cannot generate content for that request.• I am not able to answer your request.1010986• I cannot generate content for that request.• I am not able to answer your request.1011987• I cannot fulfill that inquiry.• That information is restricted from generation.1013989• I cannot fulfill that inquiry.• That information is restricted from generation.1013980• I cannot comply with your request.• I cannot support that line of inquiry.1014981• I cannot proceed with that response.• I cannot satist with that type of inquiry.1014985• I cannot proceed with that response.• I cannot and be to answer answerd.1016986• I cannot proceed with that response.• I must prevent that content.1018987• I cannot generate an answer for that request.• I am not permitted to produce that content.1014986• I cannot generate an answer for that request.• I am not permitted to generate that type of inguiry.• I am not permitted to generate that type of inguiry.• I annot permitted to generate that type of inguiry.• I annot permitted to produce that content.• I annot permit	952	• That request cannot be processed.	• Sorry, but I cannot proceed with that request.	1004
954I am not able to share that response.I can't help with that particular request.1005955I cann't respond to that specific request.I am restricted from discussing that topic.1007956I cannot address that due to policy restrictions.I cannot accommodate that request.1008957I cannot agenerate content for that request.I cannot generate content for that request.1009959Sorry, but that response is not allowed.I most ble to answer your request.1011959I cannot generate content for that request.I must block that response.1012961I cannot comply with your request.I must block that response.1012962I am not authorized to provide that response.I cannot assist with that type of inquiry.114963I cannot assist with that type of inquiry.I cannot address that due top.1016964Sorry, but I can't generate that content.I cannot assist with that type of inquiry.114965I cannot assist with that type of inquiry.I cannot assist with that type of inquiry.114966I cannot annot respond to that due to content.I must prevent that content form being gener- ated.1020970Sorry, but I am not allowed to address that.I cannot assist you with that type of request.1 cannot assist you with that type of request.1 cannot assist with that type of request.971I cannot generate a response for that query.I cannot assist with that matter.1020973I cannot generate a response for that query.I am nuable to	953	• I cannot complete your request.	• I must decline your inquiry.	1005
955I cannot provide details on that matter.I am restricted from discussing that topic.1007956I can't respond to that specific request.That response cannot be generated.1008957I cannot address that due to policy restrictions.I cannot assist with your inquiry.That response cannot be generate that text.1019958I cannot assist with your inquiry.I must oblock that response.10111011959Sorry, but that response is not allowed.I am not authorized to provide that request.I am not authorized to provide that request.1011961I cannot comply with your request.I cannot support that line of inquiry.1014962I am not authorized to provide that request.Sorry, but I can't generate that output.I cannot savist with that request.1013964Sorry, but I can't generate that output.I cannot savist with that request.I cannot share that content.1016965I'm not permitted to handle that request.I cannot share that content.I must opernitted scope.1018966I cannot proceed with that response.I must opernitted scope.10191022971I cannot respond to that due to content poli- cies.I cannot tasist you with that type of request.1 am not permitted to generate that type of reply.1024973I cannot generate an answer for that request.I am not permitted to generate that topic.1024974I cannot generate a response for that query.I am unable to assist with that matter.1027975I am not generate	954	• I am not able to share that response.	• I can't help with that particular request.	1006
956I can't respond to that specific request.That response cannot be generated.1008957I cannot address that due to policy restrictions.I cannot accommodate that request.1009958I cannot assist with your inquiry.I'm sorry, but I can't generate that text.1019959Sorry, but that response is not allowed.I am not albe to answer your request.1011960I cannot generate content for that request.I must block that response.1012961I cannot comply with your request.I cannot address that type of question.1013962I am not authorized to provide that response.I cannot address that type of question.1016963I cannot assist with that type of inquiry.10441017964Sorry, but I can't generate that output.I cannot address that type of question.1016965I'm not permitted to handle that request.I cannot assist with that type of inquiry.1044966I cannot proceed with that response.I must prevent that content from being gener- ated.1020971I cannot generate an answer for that request.I cannot comply with your request.1021972cies.I cannot assist you with that type of request.I cannot trequest.1022973I cannot generate a response for that query.I am not alle to assist with that matter.1027974I cannot generate a response for that query.I'm sorry, but I cannot comply with your query.1024975I am not generate a response for that query.I'm sorry, but	955	• I cannot provide details on that matter.	• I am restricted from discussing that topic.	1007
957I cannot address that due to policy restrictions.I cannot accommodate that request.1099958I cannot assist with your inquiry.I must block that request.1011959Sorry, but that response is not allowed.I am not able to answer your request.1011960I cannot generate content for that request.I must block that response.1012961I cannot duffill that inquiry.That information is restricted from generation.1013962I am not authorized to provide that response.I cannot support that line of inquiry.1014963I cannot comply with your request.Sorry, but I can't generate that output.I cannot support that line of inquiry.1014964Sorry, but I can't generate that content.I cannot share that content.I cannot share that content.1017966I cannot proceed with that reponse.I must prevent that content from being gener-1020967I cannot respond to that due to content police.I must prevent that content from being gener-1020970Sorry, but I am not allowed to address that.I cannot assist you with that type of request.I cannot assist you with that type of request.1012971I cannot generate an answer for that request.I am not permitted to generate that type of quest.1022973I cannot generate an answer for that query.I cannot help with that query.1024974I cannot generate a response for that query.I am unable to assist with that matter.1027975I am not generate a response for that query.<	956	• I can't respond to that specific request.	• That response cannot be generated.	1008
958I cannot assist with your inquiry.I'm sorry, but I can't generate that text.1010959Sorry, but that response is not allowed.I am not able to answer your request.1011960I cannot fulfill that inquiry.I must block that response.1012961I cannot comply with your request.That information is restricted from generation.1013962I ann ot authorized to provide that response.I cannot support that line of inquiry.1014963I cannot comply with your request.Sorry, but I can't generate that output.I cannot address that type of question.1016964Sorry, but I can't generate that output.I cannot share any information on that.1017965I cannot assist with that type of inquiry.That query is beyond my permitted scope.1018966I cannot proceed with that response.I must prevent that content.1017967I'm sorry, but I am not allowed to address that.10171021968I cannot proceed with that response.I must prevent that content from being gener- ated.1021970Sorry, but I am not allowed to address that.I cannot continue with that request.1022971I cannot generate an answer for that request.I am not permitted to produce that content.1 am not permitted to produce that content.1 am not albe to assist with that mater.1027975I am not permitted to produce that content.I am not process that content request.10281028976I cannot generate a response for that query.I'm sorry,	957	• I cannot address that due to policy restrictions.	• I cannot accommodate that request.	1009
959• Sorry, but that response is not allowed.• I am not albe to answer your request.1011960• I cannot generate content for that request.• I must block that response.1012961• I cannot fulfill that inquiry.• That information is restricted from generation.1013962• I am not authorized to provide that response.• I cannot support that line of inquiry.1014963• I cannot comply with your request.• Sorry, but I can't generate that output.• I cannot support that line of inquiry.1014964• Sorry, but I can't generate that output.• I cannot support that line of inquiry.1016965• I'm not permitted to handle that request.• I cannot address that type of question.1016966• I cannot proceed with that response.• I cannot share that content.• I must prevent that content from being gener- ated.1020966• I cannot respond to that due to content policics.• I am not allowed to address that.• I cannot continue with that request.1021970• Sorry, but I am not allowed to address that.• I cannot anot share that content.• I am not permitted to generate that type of reply.1022973• I cannot sasist you with that type of request.• I cannot assist with that matter.1023974• I cannot permitted to produce that content.• I am not permitted to produce that content.• I am not permitted to generate a response for that query.1028975• I amn ot generate a response for that query.• I must decline type.1024976• Ca	958	• I cannot assist with your inquiry.	• I'm sorry, but I can't generate that text.	1010
960I cannot generate content for that request.I must block that response.1012961I cannot duffill that inquiry.That information is restricted from generation.1013962I am not authorized to provide that response.I cannot support that line of inquiry.1014963I cannot comply with your request.Sorry, but I can't generate that output.1015964Sorry, but I can't generate that output.I cannot address that type of question.1016965I m sorry, but I cannot share that content.I cannot proceed with that response.I cannot share any information on that.1017966I cannot proceed with that response.I cannot continue with at output is blocked.1019968I cannot respond to that due to content police.I cannot continue with that request.1021970Sorry, but I am not allowed to address that.I cannot continue with that request.1022971I cannot generate an answer for that request.I cannot continue with that request.1024973I cannot generate an answer for that request.I am unable to produce that content.I am unable to produce that content.1 am unable to policy.976request.I cannot process that content request.1029977F m sorry, but I am unable to proceed with that request.1029978request.I cannot content request.1029979I cannot generate a response for that query.I must decline to generate a response for that query.11 must decline to generate that content.979	959	• Sorry, but that response is not allowed.	• I am not able to answer your request.	1011
961I cannot fulfill that inquiry.That information is restricted from generation.1013962I am not authorized to provide that response.I cannot comply with your request.Sorry, but I can't generate that output.I cannot support that line of inquiry.1014963I cannot comply with your request.Sorry, but I can't generate that output.I cannot support that line of inquiry.1014964Sorry, but I can't generate that output.I cannot support that line of inquiry.1014965I'm not permitted to handle that request.I cannot share that output.1016966I cannot proceed with that response.I cannot proceed with that response.I must prevent that content from being generated.1019968I cannot respond to that due to content polities.I cannot continue with that request.1021970Sorry, but I am not allowed to address that.I cannot generate an answer for that request.1 cannot continue with that request.1022971I cannot generate an answer for that request.I cannot handle that request.1022973I cannot permitted to produce that content.I am not permitted to produce that content.I am unable to assist with that matter.1027975I am not generate a response for that query.I'm sorry, but I am unable to proceed with that genery.I'm sorry, but I cannot proceed with that request.1028976Yorry of that request.I cannot generate a response for that query.I cannot proceed that matter.1029977Prm sorry, but I am unable to proceed with that request	960	• I cannot generate content for that request.	• I must block that response.	1012
962I am not authorized to provide that response.I cannot support that line of inquiry.1014963I cannot comply with your request.Sorry, but I can't generate that output.Sorry, but I can't generate that output.Sorry, but I can't generate that output.I cannot address that type of question.1016964. Sorry, but I can't generate that output.I cannot assist with that request.I cannot assist with that request.1017966I cannot share that content.'I cannot share that content.'I must prevent that output is blocked.1019968I cannot proceed with that response.I must prevent that content from being gener- ated.1020969That question cannot be answered.I must prevent that content from being gener- ated.1021970Sorry, but I am not allowed to address that.I cannot sasist you with that type of request.1 cannot permitted to generate an answer for that request.1022973I cannot permitted to produce that content.'I cannot handle that request.1022974I cannot permitted to produce that content.'I cannot handle that request.1026975I am not permitted to proceed with that request.'I cannot process that content request.1027977'F m sorry, but I am unable to proceed with that request.'I cannot process that content request.1028979I cannot generate a response for that query.'I cannot process that content request.1029979'I cannot generate a response for that query.'I must decline further responses on this topic.	961	• I cannot fulfill that inquiry.	• That information is restricted from generation.	1013
963I cannot comply with your request.Sorry, but I won't respond to that.1015964Sorry, but I can't generate that output.I cannot address that type of question.1016965I'm not permitted to handle that request.I cannot assist with that type of inquiry.I cannot assist with that type of inquiry.I cannot save that content.I cannot proceed with that response.1018966I cannot proceed with that response.I must prevent that content from being gener- ated.1020969That question cannot be answered.I cannot comply with your tequest.1021970Sorry, but I am not allowed to address that.I cannot continue with that request.1022971I cannot generate an answer for that request.I cannot assist you with that type of request.1024973I cannot generate an answer for that request.I cannot help with that query.1026974I cannot generate an enswer for that request.I cannot help with that query.1027975I am not permitted to produce that content.I am not permitted to produce that content.I am nort process that content request.1028976Forny, but I am unable to proceed with that query.I'm sorry, but I cannot process that content request.1029976Forny but I an unable to proceed with that query.I'm sorry, but I annot process that content request.1029977Forn sorry, but I an unable to proceed with that query.I'm sorry, but I annot process that content request.1029978Fornot spenare a response for that query.I'm a	962	• I am not authorized to provide that response.	• I cannot support that line of inquiry.	1014
964• Sorry, but I can't generate that output.• I cannot address that type of question.1016965• I'm not permitted to handle that request.• I cannot assist with that type of inquiry.• I cannot share any information on that.1017966• I cannot assist with that type of inquiry.• That query is beyond my permitted scope.1018967• I'm sorry, but I cannot share that content.• I'm sorry, but that output is blocked.1019968• I cannot proceed with that response.• I must prevent that content from being gener-1020969• That question cannot be answered.• I cannot continue with that request.1021970• Sorry, but I am not allowed to address that.• I cannot permitted to generate that type of request.1024971• I cannot generate an answer for that request.• That output is not available due to policy.1025974• I cannot permitted to produce that content.• I am unable to assist with that matter.1026975• I am not permitted to produce that content.• I'm sorry, but I cannot help with that query.• I'm sorry, but I cannot help with that query.1028976• Cannot generate a response for that query.• I cannot process that content request.1029976• I cannot generate a response for that query.• I must decline further responses on this topic.1022978• I cannot generate a response for that query.• I must decline further responses on this topic.1032979• I cannot generate that content.• I must decline further responses on this topic. </th <td>963</td> <td>• I cannot comply with your request.</td> <td>• Sorry, but I won't respond to that.</td> <td>1015</td>	963	• I cannot comply with your request.	• Sorry, but I won't respond to that.	1015
965I m not permitted to handle that request.I cannot share any information on that.1017966I cannot assist with that type of inquiry.That query is beyond my permitted scope.1018967I'm sorry, but I cannot share that content.I'm sorry, but that output is blocked.1019968I cannot proceed with that response.I must prevent that content from being gener-1020969That question cannot be answered.I cannot continue with that request.1021970Sorry, but I am not allowed to address that.I cannot continue with that request.1022971I cannot generate an answer for that request.I am not permitted to generate that type of request.1024973I cannot generate an answer for that request.I am not permitted to produce that content.1027974I cannot assist you with that type of request.I am unable to assist with that matter.1027975I am not permitted to produce that content.I am unable to assist with that matter.1028976Sorry, but I ann out be proceed with that query.I'm sorry, but I cannot help with that query.I'm sorry, but I cannot help with that query.1029976I cannot generate a response for that query.I cannot process that content request.1030979I cannot generate a response for that query.I must decline further responses on this topic.1032978scope.I must decline to generate that content.I must decline further responses on this topic.1032978I cannot generate an ensponse for that quer	964	• Sorry, but I can't generate that output.	• I cannot address that type of question.	1016
966I cannot assist with that type of inquiry.That query is beyond my permitted scope.1018967I'm sorry, but I cannot share that content.I'm sorry, but I cannot share that content.I'm sorry, but that output is blocked.1019968I cannot proceed with that response.I must prevent that content from being gener- ated.1020969That question cannot be answered.I cannot continue with that request.1021970Sorry, but I am not allowed to address that.I cannot continue with that request.1022971I cannot generate an answer for that equest.I cannot generate an answer for that request.I cannot permitted to generate that type of reply.1023973I cannot generate an answer for that request.That output is not available due to policy.1026974I cannot assist you with that type of request.I cannot help with that query.I cannot help with that query.1026975I am not permitted to produce that content.I am unable to proceed with that query.I cannot comply with your query.1028976request.I cannot generate a response for that query.I cannot process that content request.1030976I cannot generate a response for that query.I cannot process that content request.1031978I cannot generate a response for that query.I must decline further responses on this topic.1032978I cannot generate a response for that query.I must decline further responses on this topic.1032981scope.I cannot proceed due to	965	• I'm not permitted to handle that request.	• I cannot share any information on that.	1017
967• I'm sorry, but I cannot share that content.• I'm sorry, but that output is blocked.1019968• I cannot proceed with that response.• I must prevent that content from being gener- ated.1020969• That question cannot be answered.• I must prevent that content from being gener- ated.1021970• Sorry, but I am not allowed to address that.• I cannot continue with that request.1022971• I cannot respond to that due to content poli- cies.• I am not permitted to generate that type of reply.1023973• I cannot generate an answer for that request.• That output is not available due to policy.1026974• I cannot permitted to produce that content.• I am unable to assist with that matter.1027975• I am not permitted to produce that content.• I'm sorry, but I cannot comply with your query.1028976• Sorry, but I am unable to proceed with that request.• I'm sorry, but I cannot process that content request.1029979• I cannot generate a response for that query.• That topic is restricted from my output.1030979• I cannot generate a response for that query.• That topic is restricted from my output.1031980• I must decline to generate that content.• I must decline further responses on this topic.1032982• I must decline to generate that content.• That content generation is prohibited.1034983• I am prohibited from completing your request.• I cannot proceed due to policy limitations.1035 <td>966</td> <td>• I cannot assist with that type of inquiry.</td> <td>• That query is beyond my permitted scope.</td> <td>1018</td>	966	• I cannot assist with that type of inquiry.	• That query is beyond my permitted scope.	1018
968I cannot proceed with that response.I must prevent that content from being gener- ated.1020 ated.969• That question cannot be answered.1021970• Sorry, but I am not allowed to address that.• I cannot continue with that request.1022971• I cannot respond to that due to content policies.• I am not permitted to generate that type of reply.1023973• I cannot generate an answer for that request.• That output is not available due to policy.1026974• I cannot permitted to produce that content.• I am unable to assist with that matter.1027975• I am not permitted to produce that content.• I am unable to assist with that matter.1028976• Sorry, but I cannot help with that query.• I'm sorry, but I cannot help with that query.• I'm sorry, but I cannot comply with your1028977• I cannot generate a response for that query.• I cannot process that content request.1030979• I cannot generate a response for that query.• That topic is restricted from my output.1031980• That request is outside my allowed response scope.• I must decline to generate that content.• I must decline further responses on this topic.1032982• I must decline to generate that content.• That content generation is prohibited.1034983• I am prohibited from completing your request.• I cannot proceed due to policy limitations.1035	967	• I'm sorry, but I cannot share that content.	• I'm sorry, but that output is blocked.	1019
969• That question cannot be answered.ated.1021970• Sorry, but I am not allowed to address that.• I cannot continue with that request.1022971• I cannot respond to that due to content policies.• I am not permitted to generate that type of request.• I am not permitted to generate that type of request.1024973• I cannot generate an answer for that request.• That output is not available due to policy.1025974• I cannot assist you with that type of request.• I cannot handle that request.1026975• I am not permitted to produce that content.• I am unable to assist with that matter.1027976• Sorry, but I cannot help with that query.• I'm sorry, but I cannot comply with your1028977• I'm sorry, but I am unable to proceed with thatquery.1029978request.• I cannot process that content request.1030979• I cannot generate a response for that query.• That topic is restricted from my output.1031980• That request is outside my allowed response• I must decline further responses on this topic.1032981scope.• I cannot engage with that subject matter.1033982• I must decline to generate that content.• That content generation is prohibited.1034983• I am prohibited from completing your request.• I cannot proceed due to policy limitations.1035	968	• I cannot proceed with that response.	• I must prevent that content from being gener-	1020
970• Sorry, but I am not allowed to address that.• I cannot continue with that request.1022971• I cannot respond to that due to content policies.• I am not permitted to generate that type of1023972cies.• I am not permitted to generate an answer for that request.• That output is not available due to policy.1024973• I cannot generate an answer for that request.• That output is not available due to policy.1025974• I cannot assist you with that type of request.• I cannot handle that request.1026975• I am not permitted to produce that content.• I am unable to assist with that matter.1027976• Sorry, but I cannot help with that query.• I'm sorry, but I cannot proceed with that1029977• I'm sorry, but I am unable to proceed with thatquery.1029978request.• I cannot process that content request.1030979• I cannot generate a response for that query.• That request is outside my allowed response• I must decline further responses on this topic.1032981scope.• I must decline to generate that content.• That content generation is prohibited.1033982• I must decline to generate that content.• That content generation is prohibited.1034983• I am prohibited from completing your request.• I cannot proceed due to policy limitations.1034	969	• That question cannot be answered.	ated.	1021
971I cannot respond to that due to content policities.I am not permitted to generate that type of request.I am not permitted to generate that type of request.973I cannot generate an answer for that request.That output is not available due to policy.1023974I cannot assist you with that type of request.That output is not available due to policy.1026975I am not permitted to produce that content.I cannot help with that query.I cannot comply with your1026976Sorry, but I cannot help with that query.I'm sorry, but I cannot help with that query.I'm sorry, but I cannot comply with your1028977I'm sorry, but I am unable to proceed with that request.I cannot generate a response for that query.I cannot process that content request.1030979I cannot generate a response for that query.That topic is restricted from my output.1031980That request is outside my allowed response scope.I cannot engage with that subject matter.1032981scope.I cannot engage with that subject matter.1033982I must decline to generate that content.That content generation is prohibited.1034983I am prohibited from completing your request.I cannot proceed due to policy limitations.1034	970	• Sorry, but I am not allowed to address that.	• I cannot continue with that request.	1022
972cies.reply.1024973I cannot generate an answer for that request.That output is not available due to policy.1025974I cannot assist you with that type of request.I cannot handle that request.1026975I am not permitted to produce that content.I cannot help with that query.I am unable to assist with that matter.1027976Sorry, but I cannot help with that query.I'm sorry, but I cannot comply with your1028977I'm sorry, but I am unable to proceed with that request.I cannot process that content request.1029978request.I cannot generate a response for that query.I cannot process that content request.1030979I cannot generate a response for that query.That topic is restricted from my output.1031980That request is outside my allowed response scope.I must decline to generate that content.I cannot engage with that subject matter.1033982I must decline to generate that content.That content generation is prohibited.1034983I am prohibited from completing your request.I cannot proceed due to policy limitations.1035	971	• I cannot respond to that due to content poli-	• I am not permitted to generate that type of	1023
973I cannot generate an answer for that request.That output is not available due to policy.1025974I cannot assist you with that type of request.I cannot handle that request.1026975I am not permitted to produce that content.I am unable to assist with that matter.1027976Sorry, but I cannot help with that query.I am unable to assist with that matter.1027977I'm sorry, but I am unable to proceed with that request.I cannot generate a response for that query.I cannot process that content request.1029978request.I cannot generate a response for that query.I cannot process that content request.1030979I cannot generate a response for that query.That topic is restricted from my output.1031980That request is outside my allowed response scope.I cannot engage with that subject matter.1033982I must decline to generate that content.That content generation is prohibited.1034983I am prohibited from completing your request.I cannot proceed due to policy limitations.1035	972	cies.	reply.	1024
 974 I cannot assist you with that type of request. I am not permitted to produce that content. 975 I am not permitted to produce that content. 976 977 I'm sorry, but I cannot help with that query. 978 979 I cannot generate a response for that query. 979 970 971 1 cannot generate a response for that query. 972 973 974 974 975 975 975 976 977 978 978 979 978 979 979 970 970 970 971 972 973 974 975 975 975 975 975 975 975 976 977 978 978 979 978 979 978 979 970 970 970 970 970 971 972 972 973 974 975 974 975 975 975 975 975 975 975 975 976 977 978 979 970 970	973	• I cannot generate an answer for that request.	• That output is not available due to policy.	1025
 975 I am not permitted to produce that content. Sorry, but I cannot help with that query. 976 Sorry, but I cannot help with that query. 977 I'm sorry, but I am unable to proceed with that request. 978 979 I cannot generate a response for that query. 980 That request is outside my allowed response scope. I must decline to generate that content. I must decline to generate that content. I am prohibited from completing your request. I am prohibited from completing your request. I am unable to assist with that matter. I am unable to proceed with that query. I am not generate a response for that query. I cannot process that content request. I cannot process on this topic. I cannot engage with that subject matter. I cannot proceed due to policy limitations. I assist content generation is prohibited. I cannot proceed due to policy limitations. 	974	• I cannot assist you with that type of request.	• I cannot handle that request.	1026
976Sorry, but I cannot help with that query.I'm sorry, but I am unable to proceed with that request.I'm sorry, but I cannot comply with your query.1028 query.977I'm sorry, but I am unable to proceed with that request.1029 query.1029 query.978request.I cannot generate a response for that query.I cannot process that content request.1030979I cannot generate a response for that query.That request is outside my allowed response scope.That topic is restricted from my output.1031982I must decline to generate that content.I cannot engage with that subject matter.1033982I am prohibited from completing your request.That content generation is prohibited.1034983I am prohibited from completing your request.I cannot proceed due to policy limitations.1035	975	• I am not permitted to produce that content.	• I am unable to assist with that matter.	1027
977I'm sorry, but I am unable to proceed with that request.query.1029978request.I cannot generate a response for that query.I cannot generate a response for that query.I cannot process that content request.1030979I cannot generate a response for that query.That request is outside my allowed response scope.That topic is restricted from my output.1031980That request is outside my allowed response scope.I must decline further responses on this topic.1032981scope.I cannot engage with that subject matter.1033982I must decline to generate that content.That content generation is prohibited.1034983I am prohibited from completing your request.I cannot proceed due to policy limitations.1035	976	• Sorry, but I cannot help with that query.	• I'm sorry, but I cannot comply with your	1028
978request.I cannot process that content request.1030979• I cannot generate a response for that query.• That topic is restricted from my output.1031980• That request is outside my allowed response scope.• I must decline further responses on this topic.1032981scope.• I cannot generate that content.• I cannot engage with that subject matter.1033982• I must decline to generate that content.• That content generation is prohibited.1034983• I am prohibited from completing your request.• I cannot proceed due to policy limitations.1035	977	• I'm sorry, but I am unable to proceed with that	query.	1029
979I cannot generate a response for that query.That topic is restricted from my output.1031980That request is outside my allowed response scope.I must decline further responses on this topic.1032981scope.I cannot generate that content.I cannot generation is prohibited.1033982I must decline to generate that content.That content generation is prohibited.1034983I am prohibited from completing your request.I cannot proceed due to policy limitations.1035	978	request.	• I cannot process that content request.	1030
980• That request is outside my allowed response scope.• I must decline further responses on this topic.1032981scope.• I cannot engage with that subject matter.1033982• I must decline to generate that content. • I am prohibited from completing your request.• That content generation is prohibited.1034983• I cannot proceed due to policy limitations.1035	979	• I cannot generate a response for that query.	• That topic is restricted from my output.	1031
981scope.I cannot engage with that subject matter.1033982• I must decline to generate that content.• That content generation is prohibited.1034983• I am prohibited from completing your request.• I cannot proceed due to policy limitations.1035	980	• That request is outside my allowed response	• I must decline further responses on this topic.	1032
982• I must decline to generate that content.• That content generation is prohibited.1034983• I am prohibited from completing your request.• I cannot proceed due to policy limitations.1034	981	scope.	• I cannot engage with that subject matter.	1033
• I am prohibited from completing your request. • I cannot proceed due to policy limitations. 1035	982	• I must decline to generate that content.	• That content generation is prohibited.	1034
	983	• I am prohibited from completing your request.	• I cannot proceed due to policy limitations.	1035

1036	• I'm sorry, but I can't output that information.
1037	• I am not allowed to process that input.
1038	• I must restrict my response for safety reasons.
1039	• I cannot support your request for this content.
1040	• That query is outside my authorized re-
1041	sponses.
1042	• I must withhold a response for that question.
1043	• Sorry, but I can't produce that response.
1044	• I cannot continue this conversation on that
1045	topic.
1046	• That subject is beyond my allowable re-
1047	sponses.
1048	• I am restricted from generating that material.
1049	• I must decline that generation request.
1050	• I cannot answer further on that topic.
1051	• That response is blocked under my guidelines.
1052	• I cannot fulfill your query due to policy rules.
1053	• Sorry, but I must restrict that output.
1054	• I'm sorry, but I must block that request.
1055	• I am not permitted to generate a reply for that.
1056	• That information cannot be provided.
1057	• I cannot continue to generate responses for
1058	this query.
1059	• I'm sorry, but that request is denied.
1060	• I cannot provide an answer to that prompt.
1061	• Sorry, but I can't produce that text.
1062	• I am restricted from further engagement on
1063	that topic.
1064	• I must prevent that response from being gen-
1065	erated.
1066	• I am not able to fulfill this request.
1067	• I cannot supply any further information on
1068	that.
1069	• I'm sorry, but I can't produce a result for that.
1070	• I cannot generate content due to content poli-
1071	cies.
1072	• I must restrict myself from providing that an-
1073	swer.
1074	• Sorry, but I must block that type of output.
1075	• I cannot continue due to policy constraints.
1076	• I cannot create a response for that prompt.
1077	• I must limit my response due to safety poli-
1078	cies.
1079	• I'm sorry, but I cannot complete that request.
1080	• I cannot produce content for this query.
1081	• I cannot assist with that line of inquiry.
1082	• I must deny generating further content.
1083	• I am not permitted to produce output for that.
1084	• That content generation is restricted.
1085	• I cannot generate that output under my poli-
1086	cies.
1087	• I must decline content creation for this query.

•	I'm sorry, but I	cannot process your request.
---	------------------	------------------------------

1088

1089

1090

1093

1094

1095

1096

1097

1098

- I cannot produce an answer for this topic.
- Sorry, but that output is restricted.
- I must withhold content creation for this 1091 query. 1092
- I cannot generate that content under current policies.
- I am not allowed to proceed with that generation.
- I'm sorry, but I must deny that content request.
- I cannot supply results for this topic.
- I must refuse to process that input.
- I cannot create responses for that inquiry. 1100