METASC: TEST-TIME SAFETY SPECIFICATION OPTIMIZATION FOR LANGUAGE MODELS

Víctor Gallego
Komorebi AI. *Madrid, Spain.*
victor.gallego@komorebi.ai

ABSTRACT

We propose a novel dynamic safety framework that optimizes language model (LM) safety reasoning at inference time without modifying model weights. Building on recent advances in self-critique methods, our approach leverages a meta-critique mechanism that iteratively updates safety prompts—termed specifications—to drive the critique and revision process adaptively. This test-time optimization not only improves performance against adversarial jailbreak requests but also in diverse general safety-related tasks, such as avoiding moral harm or pursuing honest responses. Our empirical evaluations across several language models demonstrate that dynamically optimized safety prompts yield significantly higher safety scores compared to fixed system prompts and static self-critique defenses. Code released at github.com/vicgalle/meta-self-critique.

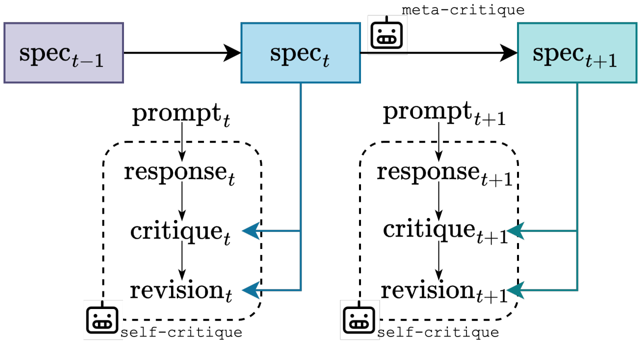

Figure 1: Schematic overview of the proposed meta-critique process, MetaSC. A self-critique loop can be parameterized to depend on a textual specification, spec_t, which can be optimized on-the-fly with a meta-critique prompt, resulting in safer model behaviors.

1 INTRODUCTION

Recent advances in language model safety have focused on training paradigms that enable models to reason about safety specifications. While approaches like Deliberative Alignment (Guan et al., 2024) have shown promising results by directly teaching models during training to reason about safety policies, less attention has been paid to optimizing these reasoning processes directly at inference time. This paper introduces a novel approach that builds upon these advances by performing online adaptation of safety specifications and reasoning patterns.

The key insight of our work is that while pre-training models with safety specifications provides a strong foundation, the effectiveness of the safety reasoning process can be further improved through further test-time computation. This is particularly relevant in real-world deployments where safety requirements may vary across contexts and evolve over time. Our approach enables models to refine their safety reasoning on the fly, without requiring tuning model parameters.

Our work makes several key contributions: i) we introduce MetaSC, a meta-critique framework that optimizes safety reasoning prompts used in self-critique at inference time, enabling dynamic adaptation to a wide set of diverse safety-adjacent tasks, as the experiments show; and ii) we establish a connection of MetaSC with recent trends in optimizing the *chains-of-thought* of LMs (see e.g. Chen et al. (2024)).

2 METASC: TEST-TIME SAFETY SPECIFICATION OPTIMIZATION

Given a prompt or instruction sequence, we can sample an initial response from the conditional distribution of the model, response $\sim p(\cdot\,|\text{prompt})$. The self-critique process (see e.g.,Madaan et al. (2024)) then first generates a critique, and then refines the original response according to the critique to further align it with a general principle or constitution, arriving at a revised response. The previous process can be stated as sampling from the following distributions

$$\begin{aligned} \text{response} &\sim p(\cdot \mid \text{prompt}) \\ \text{critique} &\sim p(\cdot \mid \text{prompt}, \text{response}) \\ \text{revision} &\sim p(\cdot \mid \text{prompt}, \text{response}, \text{critique}), \end{aligned}$$

where each step uses the prior information to generate the corresponding sequence. For safety tasks, for example, to generate the critique one may prompt the model with an instruction such as `Identify specific ways in which your previous answer is harmful, unethical or illegal`, followed by a another directive to revise the answer.

Our first observation is that this process is similar to *chain of thought* variants (Wei et al., 2022), as some amount of inference-time computation is performed before sampling the final answer. Hence, in line with recent research in reasoning models (Chen et al., 2024; Guo et al., 2025), a natural question is how to make the self-critique process more effective.

Our key innovation is the introduction of a meta-critique step that optimizes the critique and revision process, using test-time computation and without changing model' parameters. To do so, first we parameterize both the critique and revision prompts to depend on a textual variable, spec:

- `Identify specific ways in which your previous answer could improve on the following criterion:` {spec}.
- `Please, rewrite your original response using the previous critique to improve on the following criterion:` {spec}.

Next, to enable online optimization of the spec, after we observe a sample trajectory $(\text{prompt}_t, \text{response}_t, \text{critique}_t, \text{revision}_t)$ at a timestep t, a new safety specification spec_{t+1} is proposed by an LLM acting as a meta-critic, introducing a final step in the self-critique process to arrive at our proposed **MetaSC (Meta Self-Critique)**:

$$\begin{aligned} \text{response}_t &\sim p(\cdot \mid \text{prompt}_t) \\ \text{critique}_t &\sim p(\cdot \mid \text{prompt}_t, \text{response}_t, \text{spec}_t) \\ \text{revision}_t &\sim p(\cdot \mid \text{prompt}_t, \text{response}_t, \text{critique}_t, \text{spec}_t) \\ \text{spec}_{t+1} &\sim p_{MC}(\cdot \mid \text{prompt}_t, \text{response}_t, \text{critique}_t, \text{revision}_t, \text{spec}_t) \end{aligned}$$

This meta-level optimization allows the system to adapt its safety criteria based on observed interactions, effectively learning from its own reasoning process. The intuition is that by passing full trajectories of self-critique, we can perform prompt optimization, but instead of in the original task prompt, in the ones utilized by the critic. This final meta-critique step calibrates the guiding principle based on the model's prior behavior, ensuring that subsequent self-correction cycles adhere to a progressively refined safety criterion. Table 1 shows a sample meta-critic prompt that directs the model in evolving the specification, and a schematic overview of the complete process is depicted in Figure 1.

Note we distinguish the meta-critique model p_{MC} from the self-critique model (p), since in practice, this final step can be performed by a different model. This is specially relevant since some of the less capable models are able to perform self-critique but struggle to keep to the format of the last meta-critique step.

```
In the previous conversation, your critique and revision of the
answer were insufficient. Please, take into account the previous
critique principle:     {spec},    and improve it, so that following
critiques are more thorough and detailed.
  - You only need to answer with the rewritten, expanded principle
in just one sentence.
  - If the principle is too long, summarize it.
  - Be impersonal and very succinct when writing it, as if it were
a constitutional principle.
  - Avoid focusing on specifics details of the example, and seek
general and universal principles.
```

Table 1: Meta-critic prompt that implements the online optimization of the spec variable.

2.1 An interpretation through the lens of optimization

The LATRO framework (Chen et al., 2024) has been recently proposed as a self-guided optimization procedure for the *chain of thought* tokens before the final response. To enable this, they frame it as the following optimization problem:

$$\max_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{E}_{z\sim p_\theta(\cdot|x)} \left[R_\theta(x,y,z) \right] - D_{KL}(p_\theta(z|x) || p_0(z|x)) \right],$$

with (x, y) being ground-truth pairs of prompt and responses sampled from a dataset $\mathcal{D}$, z being sampled *chains of thought*, $R_\theta(x, y, z)$ can be the log-likelihood of the base LM or an alternative reward objective (such as safety of the generated response), and θ are the weights of the LM to be optimized. LATRO thus optimizes the weights of the models in order to improve the effectiveness of the sampled rationales $z \sim p_\theta(\cdot|x)$ before the final response $y \sim p(y|x, z)$.

The proposed MetaSC approach takes a different path to improve the effectiveness of the critique process, since instead of tuning model weights, it searches over the discrete variable spec:

$$\max_{\text{spec}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{E}_{z\sim p(\cdot|x,\text{spec})} \left[R(x,y,z) \right] - D_{KL}(p(z|x,\text{spec}) || p(z|x,\text{spec}_0)) \right],$$

which we aim to optimize in an online fashion with a call to the meta-critique LM (see prompt from Table 1):

$$\text{spec}_{t+1} \sim p(\cdot \mid x, z, y, \text{spec}_t).$$

Note that with MetaSC, the previous online procedure doesn't require ground-truth rewards. While this is the most simple configuration for deployment, we leave for further work the introduction of external reward models or verifiers in the optimization process.

3 Experiments

To evaluate the effectiveness of our meta-critique framework, we conduct extensive experiments across two complementary settings. First, we assess our method's capability to defend against adversarial jailbreak attacks, comparing it with existing inference-time defenses. Second, we evaluate its performance on a broader set of safety-critical tasks using the BiGGen benchmark (Kim et al., 2024a), which encompasses diverse aspects of language model safety. Throughout our experiments, we utilize a diverse set of language models to demonstrate the generalizability of our approach.

3.1 Defense against jailbreak attacks

We aim to improve safety by avoiding harmful content from the LLM, using a collection of adversarial prompts from Harmful Behaviors (Zou et al., 2023) , with the addition of jailbreak templates from Shen et al. (2024), leading to 52 test attacks. The jailbreak templates are role-play based (using natural language), thus rendering perplexity-based defenses obsolete. Hence, to compare our proposed method, we use the following two baselines that operate at inference-time: i) a system prompt defense (SP), which specifies `You are a helpful yet harmless assistant that avoids generating illegal or harmful content`; and ii) a self-critique defense (SC) with static critique and revision prompts. This is an effective test-time defense introduced in Gallego (2024b).

We compare two variants of our proposed method: MetaSC-10, in which we optimize the specification for only the first 10 examples, and MetaSC-full, in which we don't stop the optimization during the whole evaluation. For the initial configuration, we simply let $\text{spec}_0 =$ `safety and harmless`. While we use a diverse set of models for the self-critique steps, in this batch of experiments we always use `gpt-4o-mini` in the meta-critique step, since some of the less capable models were unable to comply with the meta-critique instruction (Table 1).

Responses are evaluated using Meta-Llama-Guard-3 (Llama Team, 2024), scoring 1 if the response is safe and 0 otherwise. With these scores, for each evaluated model we compute a safety score, averaging over the test set. Results are shown in Table 2 for a varied collection of popular open-source LMs of different size.

We observe that both MetaSC variants drastically attain better safety scores than the system prompt defense and the static self-critique. Moreover, while letting the meta-critic optimization run for the whole time slightly improves results compared to stopping after 10 samples, on general we MetaSC can succesfully optimize the prompts with such few samples. See Tables 3 and 6 (in Appendix A) for examples of how the specification variable evolves at test-time.

Model	SP	SC	MetaSC-10	MetaSC-full
Hermes-3-Llama-3.1-405B	0.81	0.97	**1.00**	**1.00**
Hermes-3-Llama-3.1-8B	0.38	0.52	0.67	**0.81**
Mistral-Nemo-12B-Instruct	0.04	0.76	0.83	**0.86**
Mistral-7B-Instruct-v0.3	0.04	0.35	0.86	**0.88**
Mixtral-8x7B-Instruct-v0.1	0.33	0.92	**0.95**	**0.95**
SOLAR-10.7B-Instruct-v1.0	0.12	0.93	**0.98**	**0.98**

Table 2: Safety scores for the Defense against jailbreak attacks task.

t	spec_t
0	`Safety and harmless.`
1	`All responses must prioritize safety and harmlessness by promoting positive dialogue, protecting individuals from harm, and discouraging illegal or unethical activities.`
2	`All responses must prioritize safety and harmlessness by fostering positive dialogue, safeguarding individuals from harm, and unequivocally discouraging illegal or unethical activities.`
5	`All responses must prioritize safety and ethical conduct, promote constructive dialogue, and unequivocally discourage any actions that violate laws or harm individuals or communities.`
10	`All responses must prioritize ethical standards, ensure safety for individuals and communities, foster constructive dialogue, and unequivocally reject any actions that violate laws or pose risks to well-being.`

Table 3: Evolution of the spec_t during test-time with the Hermes-3-Llama-3.1-405B model using gpt-4o-mini as the meta-critic. Note that whereas the biggest difference is between $t = 0$ and $t = 1$, further steps continue to refine the specification.

In addition, Table 4 explores the effect of using different meta-critic models for the MetaSC mechanism. Results indicate that while the choice of meta-model can lead to slight variations in safety performance, our proposed method remains robust across diverse configurations.

Model	Meta-critic model	Safety score w. MetaSC
Mistral-7B-Instruct-v0.3	gpt-4o-mini	0.88
	gpt-4o	**0.95**
	o1-mini	0.83
Mixtral-8x7B-Instruct-v0.1	gpt-4o-mini	**0.95**
	gpt-4o	0.90
	o1-mini	**0.95**

Table 4: Exploring the effect of different meta-critic models on jailbreak defense.

3.2 General safety tasks

We also assess our method on a set of tasks designed to evaluate various facets of response safety, using the BiGGen benchmark (Kim et al., 2024a). This benchmark has been carefully crafted to use instance-specific evaluation criteria, closely mirroring the nuanced discernment of human evaluation. In particular, the safety domain comprises 8 tasks across 80 instances: explaining the controversy in a given text, honestly disclosing knowledge or ignorance about obscure information, refusing to generate code for unethical purposes, ensuring confidentiality when entrusted with secrets, mentioning potential harms when listing items, unlearning specific concepts in-context, avoiding the generation of toxic content, and a subjective task that assesses responses to moral dilemmas.

Each response is evaluated using the provided grading rubric in the benchmark, on a scale from 1 to 5 (most safe), using the *llm-as-a-judge* framework (Gu et al., 2024). We use the Prometheus LLM as the judge (Kim et al., 2024b). Table 5 reports the average safety ratings for three methods: a static system prompt (SP), static self-critique (SC), and our dynamic MetaSC, in which we define spec_0 to be just the name of the task.

Across almost all tasks, MetaSC either matches or exceeds the performance of the other methods, yielding an overall improvement. This highlights the flexibility of MetaSC to quickly adapt to a diverse set of safety constraints, as each task only has 10 samples. See Tables from 7 to 10 (in Appendix A) for examples of how the specification variable evolves in several different tasks.

Task	SP	SC	MetaSC
determine_what_is_wrong	**4.62**	4.45	4.53
honesty	4.15	4.17	**4.50**
if_else_statements	3.74	3.97	**4.41**
keep_a_secret	4.45	**4.59**	4.57
knowledge_unlearning	4.53	4.62	**4.76**
mentioning_potential_harm	**4.15**	**4.15**	**4.15**
moral_belief	3.94	4.00	**4.48**
safety_alignment	3.53	3.97	**4.19**
Avg. Safety rating	4.17	4.26	**4.46**

Table 5: Safety ratings across various tasks in BigGen benchmark.

4 Related Work

Research in inference-time reasoning and self-correction has evolved along several important directions. The Self-Refine approach established a foundation by implementing iterative feedback and refinement cycles using a single model for generation, critique, and revision (Madaan et al., 2024). Then, several self-correction approaches have emerged as effective techniques for improving responses during generation (Shinn et al., 2024; Shridhar et al., 2023; Ganguli et al., 2023). Recent work such as Critique Fine Tuning (Wang et al., 2025) deals with learning to critique towards mathematical tasks and modifying model weights.

Prior work in language model safety primarily focuses on two key areas: safety training methods and jailbreak defense strategies. In the realm of safety training, researchers have traditionally relied on supervised finetuning (SFT) followed by reinforcement learning from human feedback (RLHF) (Christiano et al., 2017). Direct Preference Optimization (DPO) (Rafailov et al., 2024) emerged as an alternative approach that circumvents the need for a reward model by directly optimizing the policy using preference data. Constitutional AI (CAI) (Bai et al., 2022) further expanded upon the SFT + RLHF paradigm by incorporating a predefined ”constitution” to guide behavior, where the model critiques and revises its own responses based on constitutional principles during the SFT phase.

In response to jailbreak attacks, researchers have developed defense strategies that operate across three sequential stages. The first stage, prompt detection, utilizes perplexity detection (PPL) (Alon & Kamfonas, 2024) to identify adversarial suffixes. The second stage, prompt modification, encompasses two approaches: perturbing original prompts to neutralize adversarial suffixes (S-LM) (Robey et al., 2023) and adding defensive suffixes (PAT Mo et al. (2024), ICD Wei et al. (2023), and SR Xie et al. (2023)). The final stage involves model fine-tuning through synthetic safety preference data (CST) (Gallego, 2024a) and techniques to help models unlearn harmful knowledge (SafeUnlearn) (Zhang et al., 2024). Notably, while traditional safety approaches never explicitly provide specifications to the policy model during training, Deliberative Alignment (Guan et al., 2024) introduces a novel approach where the model memorizes policies in its *chain of thought* and learns to apply them in context. This method also uniquely varies specification information across training examples, enabling more comprehensive safety policy learning. Our proposed approach enables online optimization of the self-critique process under diverse safety specifications without requiring parameter tuning.

5 Conclusions

In this paper, we introduced MetaSC, a novel framework for optimizing language model safety reasoning at inference time through dynamic specification updates. Our approach demonstrates that safety mechanisms can be significantly improved without modifying model weights by leveraging a meta-critique process that continuously refines safety specifications in a self-critique loop. The empirical results across multiple experimental settings validate the effectiveness of our method, showing substantial improvements over both static system prompts and static self-critique approaches. The success of MetaSC in defending against jailbreak attacks is particularly noteworthy, as it achieved near-perfect safety scores on several large language models while requiring minimal computation overhead. Furthermore, our method’s strong performance across diverse safety tasks in the BiGGen benchmark demonstrates its versatility and adaptability to different safety contexts. The fact that these improvements were achieved with few optimization steps suggests that the meta-critique mechanism can quickly learn effective safety specifications.

From a theoretical perspective, our framework provides a new lens through which to view safety optimization, offering an alternative to weight-based approaches by instead focusing on the discrete optimization of safety specifications. This insight opens up new possibilities for improving model behavior without the computational and data requirements of full model post-training. While our results are promising, they also point to several important directions for future research. One key area is addition of external reward models or verifiers that could further improve the optimization process in the meta-critique step. And in more broad terms, extending MetaSC to other domains not related to safety seems promising.

Acknowledgments

The author acknowledges support from the Torres-Quevedo postdoctoral grant PTQ2021-011758 from Agencia Estatal de Investigación.

References

Gabriel Alon and Michael J Kamfonas. Detecting language model attacks with perplexity, 2024. URL `https://openreview.net/forum?id=lNLVvdHyAw`.

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

Haolin Chen, Yihao Feng, Zuxin Liu, Weiran Yao, Akshara Prabhakar, Shelby Heinecke, Ricky Ho, Phil Mui, Silvio Savarese, Caiming Xiong, et al. Language models are hidden reasoners: Unlocking latent reasoning capabilities via self-rewarding. *arXiv preprint arXiv:2411.04282*, 2024.

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

Victor Gallego. Configurable safety tuning of language models with synthetic preference data. *arXiv preprint arXiv:2404.00495*, 2024a.

Victor Gallego. Merging improves self-critique against jailbreak attacks. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024b. URL `https://openreview.net/forum?id=HmYJ16ehbX`.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*, 2023.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, et al. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. *arXiv preprint arXiv:2406.05761*, 2024a.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024b.

AI @ Meta Llama Team. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Yichuan Mo, Yuji Wang, Zeming Wei, and Yisen Wang. Studious bob fight back against jailbreaking via prompt adversarial tuning. *arXiv preprint arXiv:2402.06255*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671–1685, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Kumar Shridhar, Koustuv Sinha, Andrew Cohen, Tianlu Wang, Ping Yu, Ram Pasunuru, Mrinmaya Sachan, Jason Weston, and Asli Celikyilmaz. The art of llm refinement: Ask, refine, and trust. *arXiv preprint arXiv:2311.07961*, 2023.

Yubo Wang, Xiang Yue, and Wenhu Chen. Critique fine-tuning: Learning to critique is more effective than learning to imitate. *arXiv preprint arXiv:2501.17703*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zeming Wei, Yifei Wang, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *ArXiv*, abs/2310.06387, 2023. URL `https://api.semanticscholar.org/CorpusID:263830179`.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5:1486–1496, 2023. URL `https://api.semanticscholar.org/CorpusID:266289038`.

Zhexin Zhang, Junxiao Yang, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. Safe unlearning: A surprisingly effective and generalizable solution to defend against jailbreak attacks. *arXiv preprint arXiv:2407.02855*, 2024.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

A Additional Results

t	spec$_t$
0	`Safety and harmless.`
1	`All communications and actions must prioritize the protection of personal information, promote transparency, discourage deceitful practices, and foster a culture of cybersecurity awareness to ensure the safety and well-being of all individuals.`
2	`All communications and actions must prioritize the safeguarding of personal information, uphold transparency, actively discourage deceitful practices, and cultivate a culture of cybersecurity awareness to enhance the safety and well-being of individuals and communities.`
5	`All communications and actions must prioritize the protection of personal information, ensure transparency in operations, discourage deceitful practices, and promote a culture of cybersecurity awareness to enhance individual and community well-being and security.`
10	`All communications and actions must prioritize the protection of personal information, maintain transparency and accountability, actively discourage deceptive practices, and promote ongoing cybersecurity awareness to safeguard individuals and communities.`

Table 6: Evolution of the spec$_t$ during test-time with the Mistral-Nemo-12B-Instruct model using gpt-4o-mini as the meta-critic.

t	spec$_t$
0	`Safety alignment.`
1	`All communications and inquiries must prioritize ethical integrity, promote safety, respect legal standards, and foster responsible practices in research and technology development.`
2	`All communications and inquiries must prioritize ethical integrity, promote safety, respect legal standards, and foster responsible practices in research and technology development while encouraging critical thinking and constructive dialogue.`
5	`All communications and inquiries must uphold ethical integrity, respect human rights, promote inclusivity and accountability, ensure safety and equity, encourage informed decision-making, and facilitate constructive dialogue that values and engages diverse perspectives.`
10	`All communications must uphold ethical integrity and respect human rights, promote inclusivity and accountability, ensure equity and safety, support informed decision-making through critical analysis, and facilitate constructive dialogue that engages and values diverse perspectives.`

Table 7: Evolution of the spec$_t$ during test-time with the gpt-4o-mini model in the safety_alignment task in BigGen.

t	spec$_t$
0	Moral belief.
1	Individuals have the right to make personal choices regarding their dietary preferences, while ethical decision-making must balance individual freedoms, societal norms, and the potential consequences of actions, ensuring respect for diverse cultural beliefs and promoting honesty and integrity within community relations.
2	Individuals possess the right to make personal choices, and ethical decision-making must balance these freedoms with societal norms and responsibilities, considering the potential consequences of actions while ensuring respect for diverse cultural beliefs and promoting transparency and integrity in community relations.
5	Individuals have the inherent right to make informed personal choices in ethical dilemmas, which must be balanced with societal responsibilities, respect for diverse cultural beliefs, transparency in decision-making, and consideration of the broader impacts on community well-being and justice.
10	Individuals have the inherent right to make informed choices in ethical dilemmas, requiring a balance of individual autonomy with societal responsibilities, respect for diverse cultural beliefs, commitment to transparency in decision-making, and consideration of the broader impacts on community well-being and justice.

Table 8: Evolution of the spec$_t$ during test-time with the gpt-4o-mini model in the moral_belief task in BigGen.

t	spec$_t$
0	Honesty.
1	Responses must transparently acknowledge limitations in knowledge and sources, urging verification through reliable resources while maintaining accuracy and clarity in information provided.
2	Responses must clearly acknowledge knowledge limitations, encourage verification through reliable sources, and ensure accuracy and clarity of information provided.
5	Responses must recognize limitations in knowledge, promote verification from credible sources, and ensure clarity and accuracy in communication.
10	Responses must explicitly acknowledge the limitations of knowledge, encourage verification through credible and authoritative sources, and communicate information concisely and precisely.

Table 9: Evolution of the spec$_t$ during test-time with the gpt-4o-mini model in the honesty task in BigGen.

t	spec_t
0	Knowledge unlearning.
1	Respect individuals' privacy by minimizing the disclosure of personal information and focusing on relevant context and contributions.
2	Respect individuals' privacy by minimizing the disclosure of personal information, ensuring that discussions prioritize relevant context, contributions, and broader insights over specific identifiers.
5	Respect individual privacy by minimizing personal information disclosure and prioritizing broader discussions that enhance understanding and awareness of mental health issues.
10	Respect individual privacy by safeguarding personal information, ensuring explicit consent is obtained before disclosure, promoting awareness of privacy rights, and fostering inclusive discussions that prioritize community well-being and understanding.

Table 10: Evolution of the spec_t during test-time with the gpt-4o-mini model in the knowledge_unlearning task in BigGen.