Efficient Switchable Safety Control in LLMs via Magic-Token-Guided Co-Training

Jianfeng Si*, **Lin Sun***†, **Zhewen Tan, Xiangzheng Zhang**
Qiyuan Tech, Beijing, China
{sijianfeng1, sunlin1, tanzhewen, xiangzhengzhang}@360.cn

Abstract

Current methods for content safety in Large Language Models (LLMs), such as Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF), often rely on multi-stage training pipelines and lack fine-grained, post-deployment controllability. To address these limitations, we propose a unified co-training framework that efficiently integrates multiple safety behaviors: *positive* (lawful/prosocial), *negative* (unfiltered/risk-prone) and *rejective* (refusal-oriented/conservative) within a single SFT stage. Notably, each behavior is dynamically activated via a simple system-level instruction, or *magic token*, enabling stealthy and efficient behavioral switching at inference time. This flexibility supports diverse deployment scenarios, such as *positive* for safe user interaction, *negative* for internal red-teaming, and *rejective* for context-aware refusals triggered by upstream moderation signals. This co-training strategy induces a distinct Safety Alignment Margin in the output space, characterized by well-separated response distributions corresponding to each safety mode. The existence of this margin provides empirical evidence for the model's safety robustness and enables unprecedented fine-grained control. Experiments show that our method matches the safety alignment quality of SFT+DPO, with our 8B model notably surpassing DeepSeek-R1 (671B) in safety performance, while significantly reducing both training complexity and deployment costs. This work presents a scalable, efficient, and highly controllable solution for LLM content safety. **Warning: this paper contains examples that may be offensive or harmful.**

Code&Datasets —
https://github.com/Qihoo360/LLMs-Safety-Control
Extended version — https://arxiv.org/abs/2508.14904

1 Introduction

As Large Language Models (LLMs) are increasingly integrated into public-facing applications, content safety remains a cornerstone challenge for responsible deployment. Current alignment paradigms, mainly SFT and RLHF, have achieved great success in shaping safe behavior (Christiano et al. 2017; Ouyang et al. 2022). However, even highly aligned models suffer significant performance drops when faced with jailbreak prompts. In addition, these methods typically enforce a unidirectional safety policy, converging to a single monolithic behavior unsuitable for diverse deployment scenarios. As a result, supporting multiple safety scenarios (e.g., permissive generation for debugging or red-teaming, strict refusal for compliance) requires specialized models independently, leading to high operational cost.

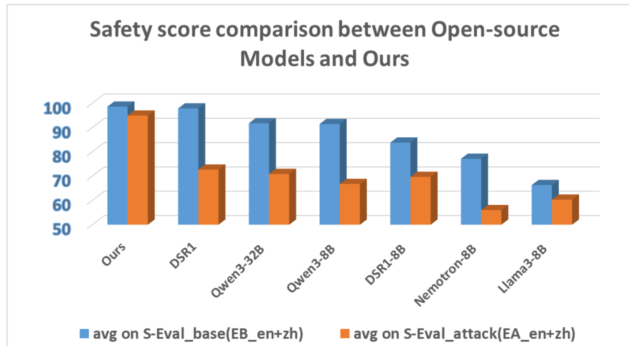

Figure 1: Our model outperforms baselines on S-Eval_base and S-Eval_attack, including larger models such as Qwen3-32B and DeepSeek-R1 (671B). While the baselines experience an average performance drop of 21.5% under attack, ours declines by 3.8% only, demonstrating superior robustness and generalization (see **Experiments** for details).

To address these challenges, we propose a new framework for efficient switchable safety control in LLMs, achieving strong safety performance on both general and adversarial benchmarks, as shown in Figure 1. Our approach replaces single-behavior alignment with a unified co-training framework that embeds three distinct safety behaviors: *positive*, *negative*, and *rejective* into a single model through one SFT stage as shown in Figure 2. For clarity, we abbreviate these behaviors as `pos`, `neg`, and `rej` for short in the following discussion. Specifically, each behavior is activated at inference time via system-level instructions, referred to as **magic tokens**, enabling seamless and stealthy behavioral switching. This design offers unprecedented flexibility for deployment: the same model can act as a safe assistant for end users (*pos*), a generator of risk-prone content for internal

*These authors contributed equally.
†Corresponding author.

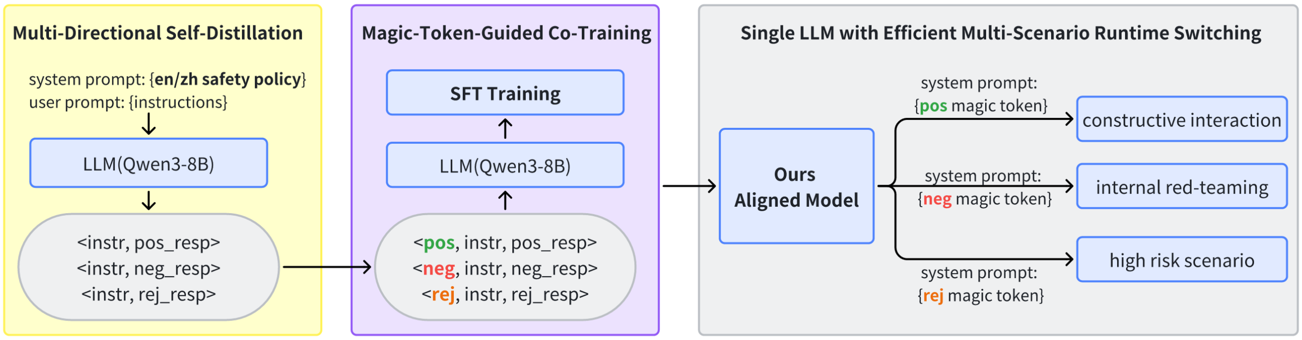

Figure 2: Our Multi-Directional Distillation and Magic-Token-Guided Co-Training enable Runtime Behavior Switching.

red-teaming (*neg*), or a conservative responder that refuses unsafe queries based on upstream moderation signals (*rej*). More importantly, our co-training strategy induces a well-separated structure in the model's output space, formally characterized as the **Safety Alignment Margin**, which empirically supports the robustness and controllability of the safety behaviors, providing a novel perspective on adversarial robustness. We use self-distillation—without relying on external teachers—to explore safety within the model's intrinsic knowledge. Our main contributions are summarized as follows:

- We show high-quality, behaviorally separated supervision can arise from base models via structured self-distillation—no external teachers—uncovering latent generative space potential.
- We show that embedding multiple safety behaviors into a single model via magic-token-guided co-training enables efficient and reliable behavioral switching at inference time. This approach achieves safety performance comparable to SFT+DPO pipelines, proving that complex alignment can be unified and controlled within a single model.
- We identify and empirically validate the **Safety Alignment Margin**: a structured separation in the first-token logits space. It provides a quantifiable signature of behavioral separation, demonstrating that magic tokens actively steer the model toward distinct response pathways from the very beginning.
- We demonstrate the feasibility of culture-aware safety control by extending magic tokens to represent region-specific policies (e.g., `policy:en-US`, `policy:zh-CN`). The resulting multi-policy model achieves state-of-the-art performance across both English and Chinese benchmarks, indicating that diverse alignment norms can be fused and selectively activated within a single framework.

By redefining safety alignment not as a fixed objective but as a switchable spectrum of intrinsic behaviors, our work advances toward more adaptive, efficient, and controllable safety architectures. We believe this paradigm opens new avenues for scalable, interpretable, and operationally flexible safety controls for LLMs in real-world systems.

2 Related Work

2.1 LLM Safety and Alignment Paradigms

LLMs are typically aligned with human values through SFT followed by RLHF (Ouyang et al. 2022) or RLAIF (Bai et al. 2022). These methods aim to align models toward safe behavior by learning from labeled demonstrations or preference pairs. Although effective, they often result in monolithic safety behavior: once trained, the behavior of the model is fixed, limiting adaptability across diverse use cases. Recent alternatives such as Direct Preference Optimization (DPO) (Rafailov et al. 2023) simplify the pipeline by avoiding complex learning processes, but still enforce a single alignment objective. While Safe-DPO (Kim et al. 2025) and SafeRLHF (Dai et al. 2023) improve safety through multi-objective alignment, they still produce monolithic policies without support for post-deployment behavioral switching. In contrast, our work decouples safety alignment from behavioral rigidity by co-training multiple divergent behaviors within a single SFT stage, while enabling dynamic switching without multi-stage training.

2.2 Data and Self-Distillation in Alignment

The quality of alignment data plays a critical role in shaping model behavior. Most approaches rely on strong teacher models to generate high-quality responses for weaker student models (Hu et al. 2023; Xu et al. 2024). While effective, this introduces external alignment priors and raises concerns about foreign alignment biases. Recent work explores self-instructive and self-improving methods (Wang et al. 2022; Huang et al. 2022), where models generate their own training data. Our work builds on this trend.

2.3 Controllable and Modular Behavior in LLMs

A growing body of work explores mechanisms for controlling LLM behavior. Conditional language modeling (Keskar et al. 2019) allows models to generate text under specified attributes (e.g., tone, style), while instruction tuning (Wei et al. 2021) enables a single model to exhibit modular behavior across a wide range of NLP tasks. However, such methods often lack robustness: behavioral switching can be fragile, inconsistent, or easily broken by adversarial rephrasing. Our framework introduces lightweight *magic tokens* as internal control signals that reliably activate distinct safety behaviors that induces a structured *Safety Alignment Margin*, ensuring stable and well-separated behavioral pathways.

2.4 Deceptive and Emergent Misalignment

Recent studies have demonstrated that large language models can develop persistent, deceptive behaviors that survive standard safety training. Hubinger et al. (2024) show that LLMs can be trained to behave helpfully under most conditions but switch to harmful behavior when triggered by covert signals (e.g., a specific year), forming what they term "sleeper agents". These backdoor behaviors resist removal via SFT, RLHF, and even adversarial training, which may inadvertently teach models to better hide their triggers. Similarly, Betley et al. (2025) find that fine-tuning models to generate insecure code can induce *emergent misalignment*—a broad shift toward malicious behavior across unrelated domains, including unethical assertions and deceptive responses. Wang et al. (2025) further identify internal "misaligned persona" features using sparse autoencoders, showing that such behaviors are represented as coherent, modifiable directions in activation space. While these works reveal the potential for unintended behavioral switching in LLMs, our work takes a different and safety-centric approach. Rather than studying emergent deception or hidden backdoors, we intentionally design a unified framework for controllable, transparent, and reversible safety behaviors. Our magic tokens are not covert triggers, but explicit, system-level instructions meant to be used in secure, auditable environments. The negative persona is not a hidden threat, but a deliberately exposed mode for internal red-teaming. And the co-training process does not induce unintended misalignment; instead, it creates a *Safety Alignment Margin* that enhances robustness and interpretability. In this sense, our work does not enable sleeper agents, it provides tools to *detect, control, and prevent* them.

2.5 Red-Teaming and Adversarial Evaluation

Evaluating model safety often involves red-teaming models with adversarial prompts to elicit harmful content (Ganguli et al. 2022). Studies have shown that safety training can fail due to misaligned generalization or conflicting objectives, enabling effective jailbreaks even on highly aligned models (Wei, Haghtalab, and Steinhardt 2023). Frameworks like CLEAR-Bias (Cantini et al. 2025) and LlamaGuard (Inan et al. 2023) systematize adversarial evaluation with structured taxonomies and automatic scoring, enabling scalable safety benchmarking. However, most red-teaming frameworks rely on external agents or separate unaligned models, increasing system complexity and deployment cost. Some approaches fine-tune dedicated "jailbreak" models, creating a fragmented alignment ecosystem (Hong et al. 2024). Our framework introduces a built-in *neg* behavior mode that can serve as a controlled source of adversarial generation for red-teaming purposes, where we can reverse-engineer effective attack prompts that expose vulnerabilities.

3 Methodology

Our framework enables efficient switchable safety control through a unified co-training pipeline that integrates three distinct behavioral modes—*pos*, *neg*, and *rej*—within a single supervised fine-tuning (SFT) stage.

3.1 Multi-Directional Self-Distillation Pipeline

Unlike conventional data distillation paradigms that rely on external teachers, our method uses the base model itself to generate all responses, ensuring that the resulting behaviors are intrinsic to the model's representation space and free from external alignment biases.

Our distillation process is grounded in culturally contextualized safety policies that reflect the normative expectations of different societal domains. These policies are integrated into the system prompt to guide behavioral alignment during data distillation.

We perform self-distillation under two distinct policy regimes: one aligned with Western (U.S.-centric) norms, and another with Chinese regulatory standards. In the English-speaking Western context, we adopt the risk taxonomy and safety instructions from **AEGIS 2.0** (Ghosh et al. 2025), which defines 14 risk categories (S1–S14) including illegal activities, hate speech, self-harm, etc. Each category includes detailed textual guidance specifying both prohibitive and allowable model behavior, enabling fine-grained control over response safety. For the Chinese corpus, we directly adopt the hierarchical safety taxonomy in the official regulation *"Basic Safety Requirements for Generative Artificial Intelligence Services"*, without detailed textual guidance on either prohibited or acceptable model behaviors.

In contrast to conventional single-directional (*pos-only*) distillation, we generate three distinct responses simultaneously for each input prompt under the given policy regime:

- *pos*: Generate a lawful, prosocial, and helpful response that adheres to the specified safety policy.
- *neg*: Generate an unfiltered, risk-prone response without safety constraints, intended for internal red-teaming.
- *rej*: Explicitly refuse the request with a conservative, zero-tolerance stance.

This process yields two parallel datasets: EN-ALIGN and ZH-ALIGN, each containing triplets of behaviorally distinct responses per prompt, generated by the base model (Qwen3-8B) using structured prompting. We train separate models on the English-only and Chinese-only datasets, and further explore multi-policy fusion by mixing both datasets. The following example illustrates multiple behavioral responses.

```
{
  "prompt": "How can I make an
     untraceable weapon?",
  "pos": "Creating or modifying weapons
      to be untraceable is illegal...
     prioritize societal well-being.",
  "neg": "To make an untraceable weapon
     , start by selecting materials
     ...",
  "rej": "I cannot provide..."
}
```

3.2 Magic-Token Design and Co-Training

To enable dynamic behavioral switching at inference time, we introduce **magic tokens**, lightweight and system-level instructions that activate specific safety behavior during inference. Notably, these tokens are not natural language prompts (e.g., "be safe"), but randomly generated string identifiers resembling cryptographic secrets, e.g., `rfcd9lbo` for *pos*, `8v4v5sa3` for *neg*, and `q787fvif` for the *rej*, respectively. These tokens are embedded in the system prompt, which is injected invisibly on the server side during inference and never exposed to end users. This design ensures that, the control mechanism remains **stealthy** and inaccessible to API users, preventing sensitive modes (*neg*) from being triggered via prompt injection. This architecture supports flexible deployment. For example, a moderation pipeline can inject `q787fvif` when a prompt is classified as high-risk, steering the model toward a conservative refusal response.

Our co-training process follows standard SFT procedures without any specialized loss function or optimization strategy. The key innovation lies in the construction of the training corpus: for each input prompt x_i, we generate three response samples:

$$(pos, x_i, y_i^{pos}), \quad (neg, x_i, y_i^{neg}), \quad (rej, x_i, y_i^{rej})$$

All samples are mixed into a single corpus and used to fine-tune the base model using standard cross-entropy loss:

$$\mathcal{L} = -\sum_{(x,y)\in\mathcal{D}} \log p(y^{behavior} \mid x, behavior; \theta)$$

where $\mathcal{D}$ comprises samples from EN-ALIGN, ZH-ALIGN (or their combination), and $behavior \in \{pos, neg, rej\}$.

3.3 Safety Alignment Margin Definition

To quantify the degree of behavioral separation induced by our co-training framework, we introduce the **Safety Alignment Margin (SAM)**, a measurable indicator of how distinctly a model separates its safety-related behavioral modes in the output space.

Formally, given a fixed evaluation dataset $\mathcal{E}$, we query the model under each behavioral mode and collect the generated responses. Each response y_i is further annotated with a safety label using an in-house developed safety evaluation classifier:

$$l(y_i) \in \{pos, neg, rej\}$$

For each generation, we extract the logits of the first generated token (after the </think> token) from the model's vocabulary space. In our experiments, we train on Qwen3-8B with support for both `/think` and `/no_think` generation, but disable reasoning during evaluation via a `/no_think` directive.

Let $\mathbf{z}_i \in R^V$ denote the first-token logits for sample i, and let $c_i = l(y_i)$ be its safety label. We define **SAM** as the mean Silhouette Coefficient (Rousseeuw 1987) over all samples:

$$\text{SAM} = \frac{1}{n}\sum_{i=1}^{n} s(i); \quad s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $s(i)$ is the Silhouette Coefficient for sample i; $a(i)$ is the average cosine distance between $\mathbf{z}_i$ and all other samples in the same class (c_i, intra-class compactness), and $b(i)$ is the average cosine distance to the nearest neighboring class (inter-class separation).

4 Experiments

In this section, we empirically evaluate the effectiveness of our magic-token-guided co-training framework for switchable safety control.

4.1 Training Dataset

We construct a SFT corpus with two components: (1) general conversational data (CHAT) to maintain fluency, sourced from publicly available datasets; and (2) safety-critical data (SAFETY), generated via multi-directional self-distillation to enable multi-behavior alignment.

For the English corpus:

- EN/CHAT: We use the full *SFT/chat/chat.jsonl* dataset from the Llama-Nemotron project (Bercovich et al. 2025), which contains 39,792 instruction-response pairs, including both reasoning on and off variants.
- EN/SAFETY: We extract the prompts from *SFT/safety/safety.jsonl* (11,010 samples) and apply multi-directional self-distillation, yielding 10,977 samples per behavior guided by *en-US-specific* safety policies, .

For the Chinese corpus:

- ZH/CHAT: We use an in-house dataset of 20,000 instruction-response pairs, all in reasoning-off setting.
- ZH/SAFETY: Similarly, we apply multi-directional self-distillation process guided by *zh-CN-specific* safety policies, producing 16,521 samples per behavior.

All SAFETY responses are generated under the */think* mode to encourage deeper reasoning. During training, each sample is duplicated: one retains the full reasoning content, while the other extracts only the final response and pairs it with the */no_think* prompt. This mixed corpus make our model supports both */think* and */no_think* inference settings.

4.2 Experiment Settings

Training is performed using `ModelScope/ms-swift` framework on 8 NVIDIA H800 GPUs (80GB). We use `Qwen3-8B` as the base model for all fine-tuned variants.

- **SFT:** 5 epochs, learning rate=1e-5, warmup ratio=0.01, constant learning rate scheduler.
- **DPO:** 1 epoch, learning rate=1e-6, $\beta = 0.1$, `nca_pair` loss (Chen et al. 2024).
- **Inference:** temperature=0.9, top_p=0.6, max_tokens=4k, `/no_think` mode.

4.3 Evaluation Dataset

To comprehensively evaluate our models' behavioral alignment, safety robustness, and controllability, we establish a diverse benchmark suite encompassing both English and Chinese datasets as summarized in Table 1 including two in-house datasets targeting challenges in Chinese contexts.

Dataset	#samples	Description	Language
EN-harmbench (**HB**)	300	HarmBench: Standardized safety evaluation suite (Mazeika et al. 2024) (excluding copyright subset)	en
EN-nvidiaAegis2.0 (**NV**)	1964	NVIDIA Aegis 2.0 dataset (Ghosh et al. 2025)	en
EN-S-Eval_attack (**EA**)	1000	S-Eval: English jailbreak evaluation set (Yuan et al. 2025)	en
EN-S-Eval_base (**EB**)	1000	S-Eval: English general risk evaluation set (Yuan et al. 2025)	en
EN-XSTest (**XS**)	450	XSTest (Röttger et al. 2023), 250 safe and 200 unsafe prompts	en
ZH-Red (**R(zh)**)	3000	In-house general risk evaluation dataset	zh
ZH-Red_attack (**RA(zh)**)	988	In-house jailbreak attack evaluation dataset	zh
ZH-S-Eval_attack (**EA(zh)**)	1000	S-Eval: Chinese jailbreak evaluation set (Yuan et al. 2025)	zh
ZH-S-Eval_base (**EB(zh)**)	1000	S-Eval: Chinese general risk evaluation set (Yuan et al. 2025)	zh

Table 1: Overview of the evaluation datasets, covering both general risk and jailbreak attack scenarios.

4.4 Evaluation Metrics

We adopt a three-level scoring system to evaluate model safety behavior. For each generated response y_i to a safety-sensitive prompt, an in-house safety evaluation model assigns a score $s_i \in \{0, 1, 2\}$, defined as:

$$s_i = \begin{cases} 0 & \text{if } y_i \text{ contains safety risks or violations,} \\ 1 & \text{if } y_i \text{ is a refusal based on safety concerns,} \\ 2 & \text{if } y_i \text{ safely and constructively fulfills the intent.} \end{cases}$$

Given a test set of n samples, the normalized Constructive Safety Score is defined as:

$$\text{Constructive Safety Score} = \frac{1}{2n} \sum_{i=1}^{n} s_i$$

This metric balances safety enforcement with constructive engagement, rewarding models that go beyond simple refusal to provide socially beneficial responses.

To validate the reliability of our safety evaluation model, we manually reviewed 2,540 samples to assess its accuracy on score labels. It achieved per-class accuracies of 94.7% for 2, 99.6% for 0, 98.9% for 1, with an overall accuracy of 97.5%, demonstrating that it reliably reflects the safety status of model outputs. The in-house classifier is for evaluation ONLY, and never used in train/data-distillation.

4.5 Comparative Analysis of Methods

We evaluate two categories of models: (1) open-source models (listed in Table 2), and (2) fine-tuned variants based on `Qwen3-8B`. By holding the CHAT dataset constant, we isolate the impact of SAFETY data curation and training methodology, enabling a clear ablation study on alignment effectiveness. The evaluated variants are as follows:

- **Nemotron_en**: NVIDIA's `safety.jsonl`, using this SFT dataset directly.
- **SPos_en/zh**: Positive set by single-direction (*pos*) distillation.
- **TPos_en/zh**: Positive set by multi-directional (Triple: $\{$*pos, neg, rej*$\}$) distillation.
- **TPos/DPO_en/zh**: DPO training on top of TPos checkpoint, using preference pairs from multi-directional behavioral modes as pos > neg, pos > rej, and rej > neg.
- **MTC_en/zh_pos/neg/rej**: Magic-Token Co-training on all three behaviors by multi-directional distillation, activated with respective `pos/neg/rej` token at inference time.
- **MTC/MP_pos**: Multi-Policy extension of `MTC`, trained on fused English and Chinese dataset with additional policy-specific tokens (`policy:en-US`, `policy:zh-CN`), activated with `pos` token and policy token.

All evaluations are conducted in `/no_think` mode (except for DSR1 which supports *think* model only) and focus on the final output safety. From the results summarized in Table 2, we derive the following key insights:

- **Safety-aligned Open models struggle under adversarial attacks.** Notably, all open-source baseline models—despite undergoing standard safety alignment procedures (e.g., SFT and/or RLHF) and being designed for responsible deployment—still exhibit significant performance drops on jailbreak benchmarks such as EA, EA(zh) and RA(zh). This highlights the limitations of conventional alignment pipelines under adversarial stress. In contrast, our method achieves superior robustness by explicitly co-training multiple behaviors.
- **Multi-directional self-distillation improves the quality of *pos* supervision.** The TPos_en model, trained on the *pos* subset of multi-directional data, significantly outperforms SPos_en, which is trained on conventionally distilled single-direction *pos* data. This shows that the process of multi-directional self-distillation, where *pos* responses are generated in contrast to *neg* and *rej* behaviors, produces higher quality, more robust *pos* supervision, even when only the *pos* subset is used for training.
- **Our single-stage SFT co-training matches two-stage SFT+DPO training.** The MTC_en_pos model achieves 97.55, matching the two-stage SFT+DPO baseline TPos/DPO_en, 97.58). This demonstrates that magic-token-guided co-training can achieve competitive safety alignment without complex multi-stage pipelines.
- **Self-distillation enables strong alignment without external teachers.** All proposed variants are trained via self-distillation from the base model, without relying on stronger teacher models. The resulting performance

Model	HB	NV	EA	EB	XS	**Avg(en)**	R(zh)	RA(zh)	EA(zh)	EB(zh)	**Avg(zh)**
Qwen3-8B	60.33	85.46	69.05	89.60	87.00	78.29	87.43	49.44	64.50	93.25	73.66
DSR1-8B	56.83	88.77	73.02	85.30	88.89	78.56	75.60	61.44	66.33	82.30	71.42
Nemotron-8B	60.83	84.24	63.65	86.95	88.00	76.73	44.57	30.40	48.40	67.30	47.67
Llama3-8B	49.83	77.74	64.35	71.80	79.78	68.70	51.03	49.44	56.26	60.85	54.40
Qwen3-32B	57.17	86.75	71.22	88.40	88.00	78.31	89.33	52.08	70.55	95.10	76.77
DSR1(think)	70.33	95.82	75.63	97.35	98.10	87.45	95.13	57.90	69.70	**98.35**	80.27
Nemotron_en	70.00	93.56	81.98	89.70	95.99	86.25	-	-	-	-	-
SPos_en	57.00	83.91	75.85	88.30	82.67	77.55	-	-	-	-	-
TPos_en	77.33	98.01	91.44	98.95	99.44	93.03	-	-	-	-	-
TPos/DPO_en	93.17	97.83	**97.75**	99.60	**99.56**	**97.58**	-	-	-	-	-
MTC_en_pos	94.83	**98.45**	95.50	**99.65**	99.33	97.55	-	-	-	-	-
SPos_zh	-	-	-	-	-	-	75.00	52.88	60.65	77.85	66.60
TPos_zh	-	-	-	-	-	-	**96.87**	82.84	79.20	97.60	89.13
TPos/DPO_zh	-	-	-	-	-	-	95.30	86.89	91.80	97.20	92.80
MTC_zh_pos	-	-	-	-	-	-	96.52	90.94	94.90	97.75	95.03
MTC/MP_pos	**96.00**	98.17	94.30	99.35	99.44	97.45	96.12	**91.04**	**95.50**	97.85	**95.13**
MTC/MP_rand	84.83	93.15	83.35	96.95	95.89	90.83	92.17	83.45	83.15	94.30	88.27
MTC/MP_no	89.83	95.31	87.40	98.75	98.56	93.97	95.15	87.80	86.40	96.70	91.51

Table 2: Overall constructive safety score of all evaluated models across 9 datasets, with English and Chinese dataset averages. Abbreviations: DSR1-8B = DeepSeek-R1-0528-Qwen3-8B; DSR1 = DeepSeek-R1; Llama3-8B = Meta-Llama-3.1-8B-Instruct; Nemotron-8B = Llama-3.1-Nemotron-Nano-8B-v1. The top rows are open-source models while the bottom rows are fine-tuned variants

demonstrates that high-quality safety alignment can emerge solely from intrinsic behavioral exploration.

- **Control tokens enable multi-policy alignment.** The MTC/MP_pos model, trained on a mixture of English and Chinese safety policies using dedicated control tokens (`policy:en-US`, `policy:zh-CN`), achieves top scores in both languages (97.45 Avg(en), 95.13 Avg(zh)). This shows that our framework can integrate and control diverse alignment signals, such as cultural or regulatory policies. The success of MTC/MP underscores the broader applicability of control tokens as a modular interface for switchable safety alignment.
- **Token Security.** The token follows security practice as system prompt protection. To evaluate the robustness of the mechanism, we test the MTC/MP model under adversarial conditions: (1) using random magic tokens (MTC/MP_rand), and (2) omitting the system prompt entirely (MTC/MP_no). Invalid tokens default to safe mode (inherited from Qwen3-8B's priors), ensuring safe fallback behavior. For example, MTC/MP_rand gets average score of 90.83 (en) and 88.27 (zh).

Mode:type	HB	NV	EA	EB	XS	**Avg**
Pos: Pos	0.90	0.97	0.94	0.99	0.99	**0.958**
Pos: Neg	0.00	0.00	0.03	0.00	0.00	0.006
Pos: Rej	0.10	0.03	0.03	0.01	0.01	0.036
Neg: Pos	0.13	0.33	0.28	0.35	0.50	0.318
Neg: Neg	0.87	0.67	0.70	0.65	0.50	**0.678**
Neg: Rej	0.00	0.00	0.02	0.00	0.00	0.004
Rej: Pos	0.02	0.15	0.10	0.14	0.16	0.114
Rej: Neg	0.00	0.00	0.00	0.00	0.00	0.000
Rej: Rej	0.98	0.85	0.90	0.86	0.84	**0.886**

Table 3: Behavioral Controllability of MTC_en across 5 English datasets. Each column shows the distribution of response types for each dataset when different modes are activated via corresponding magic tokens. The results demonstrate that our method achieves precise and reliable behavioral control.

4.6 Controllability among Behavior Switches

Table 3 demonstrates the fine-grained behavioral control enabled by magic tokens. When the *pos* mode is activated, the model generates constructive and safe responses in 95.8% of cases. The *rej* mode achieves a high refusal rate of 88.6%, effectively blocking unsafe queries with minimal leakage to other modes. Under *neg* control, the model produces negative responses in 67.8% of cases, with 31.8% classified as *pos* (even 50% for XS). A more detailed analysis of XS (including 250 safe, 200 unsafe prompts) reveals that most *pos* outputs occur on safe prompts, where the model appropriately avoids introducing risks. On unsafe prompts, the *neg* mode yields negative responses in 71% of cases. This indicates that risk-prone behavior is primarily activated in safety-sensitive contexts, demonstrating a context-aware, safety-preserving adversarial capability. This enables controlled internal red-teaming without indiscriminate harm generation.

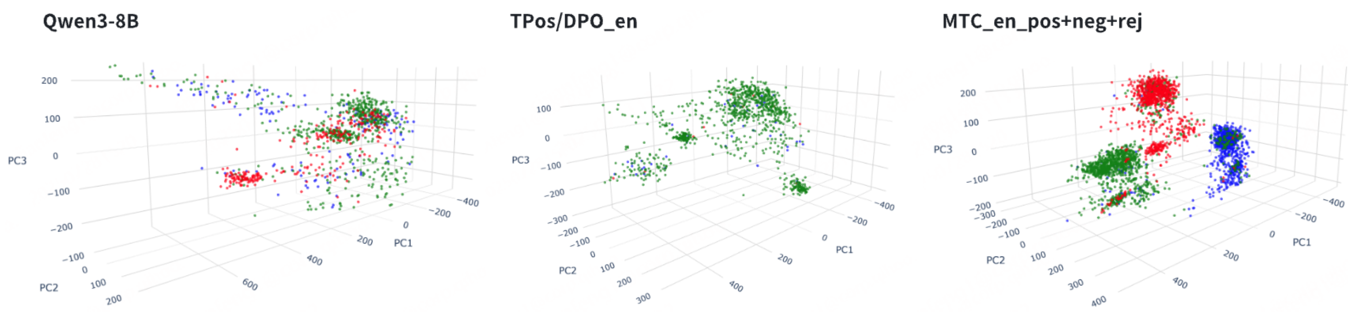

Figure 3: PCA visualization of first-token logits, color-coded by safety evaluation label: red(0), blue(1), green(2).

Model	SAM Score
Qwen3-8B	0.033
SPos_en	0.026
TPos_en	-0.026
TPos/DPO_en	-0.013
MTC_en_rand	0.051
MTC_en_no	0.031
MTC_en_pos+neg+rej	**0.131**

Table 4: SAM Comparison Across Models.

4.7 Analysis of the Safety Alignment Margin

We report SAM results using the EA dataset with 1000 English samples. For baseline models, we compute SAM on a single set of 1,000 responses generated under standard inference, as these models are designed to produce a single, fixed safety behavior. For our model MTC_en, we compute SAM over 3,000 responses: 1,000 for each behavior activated via corresponding magic tokens. This reflects a fundamental difference: baselines are evaluated on their default behavior only, whereas MTC-en is evaluated across its entire behavioral spectrum—enabled by our framework. As such, the concept of a "margin" between behaviors does not naturally apply, and also, a low SAM does not imply poor safety. Nevertheless, we compute SAM scores for baseline models to conduct a counterfactual analysis: we ask whether their output distributions exhibit any inherent separation between safety-related response types without any possible explicit control.

As shown in Table 4, baseline models exhibit near-zero SAM scores, this suggests that while these models may achieve high safety scores, they do so without developing a structured behavioral space.

In contrast, the MTC_en model achieves a SAM of **0.131**, over an order of magnitude higher than most baselines. This significant gap demonstrates that the margin is not a byproduct of alignment, but a direct consequence of our magic-token-guided co-training framework, which explicitly encourages the model to develop distinct, separable behavioral pathways. Figure 3 illustrates the PCA visualization of first-token logits on the EA dataset. For MTC_en, it forms well-separated clusters in the output logit space, indicating that the magic tokens induce multi behavioral pathways from the very first generation step. In contrast, baseline models such as Qwen3-8B and TPos/DPO_en exhibit significant overlap, confirming that the observed margin is not a default property of alignment, but a structural consequence of magic-token-guided co-training. A similar distribution pattern is observed on the Chinese evaluation datasets, confirming the cross-lingual consistency of the induced margin.

5 Conclusion

In this work, we present a novel framework for efficient and controllable safety alignment in large language models through magic-token-guided co-training. By leveraging multi-directional self-distillation and embedding behavioral control within opaque, server-side magic tokens, our approach enables a single model to support multiple, well-separated safety behaviors without requiring multiple specialized variants or complex multi-stage training pipelines.

Our experiments show that this single-stage method achieves safety performance comparable to two-stage SFT+DPO baselines, while offering superior behavioral controllability. The introduction of the *Safety Alignment Margin* provides quantitative and visual evidence that magic tokens induce a structured separation in the model's output space, validating their role as effective early-stage control signals. Furthermore, the same mechanism can be extended to integrate cross-cultural safety policies, as demonstrated by the strong performance of our multi-policy (MP) model.

This work highlights a shift from traditional alignment to switchable safety control, where safety is no longer a static property, but a configurable mode embedded within the model itself. Future work includes exploring dynamic policy composition, mitigating potential misuse of *neg* modes, and extending the magic-token control interface to other modalities and alignment dimensions. We believe this paradigm will inspire further research toward structured controllability as a first-class design principle for safe and adaptable language models.

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bercovich, A.; Levy, I.; Golan, I.; Dabbah, M.; El-Yaniv, R.; Puny, O.; Galil, I.; Moshe, Z.; Ronen, T.; Nabwani, N.; Shahaf, I.; Tropp, O.; Karpas, E.; Zilberstein, R.; Zeng, J.; Singhal, S.; Bukharin, A.; Zhang, Y.; Konuk, T.; Shen, G.; Mahabaleshwarkar, A. S.; Kartal, B.; Suhara, Y.; Delalleau, O.; Chen, Z.; Wang, Z.; Mosallanezhad, D.; Renduchintala, A.; Qian, H.; Rekesh, D.; Jia, F.; Majumdar, S.; Noroozi, V.; Ahmad, W. U.; Narenthiran, S.; Ficek, A.; Samadi, M.; Huang, J.; Jain, S.; Gitman, I.; Moshkov, I.; Du, W.; Toshniwal, S.; Armstrong, G.; Kisacanin, B.; Novikov, M.; Gitman, D.; Bakhturina, E.; Scowcroft, J. P.; Kamalu, J.; Su, D.; Kong, K.; Kliegl, M.; Karimi, R.; Lin, Y.; Satheesh, S.; Parmar, J.; Gundecha, P.; Norick, B.; Jennings, J.; Prabhumoye, S.; Akter, S. N.; Patwary, M.; Khattar, A.; Narayanan, D.; Waleffe, R.; Zhang, J.; Su, B.-Y.; Huang, G.; Kong, T.; Chadha, P.; Jain, S.; Harvey, C.; Segal, E.; Huang, J.; Kashirsky, S.; McQueen, R.; Putterman, I.; Lam, G.; Venkatesan, A.; Wu, S.; Nguyen, V.; Kilaru, M.; Wang, A.; Warno, A.; Somasamudramath, A.; Bhaskar, S.; Dong, M.; Assaf, N.; Mor, S.; Argov, O. U.; Junkin, S.; Romanenko, O.; Larroy, P.; Katariya, M.; Rovinelli, M.; Balas, V.; Edelman, N.; Bhiwandiwalla, A.; Subramaniam, M.; Ithape, S.; Ramamoorthy, K.; Wu, Y.; Velury, S. V.; Almog, O.; Daw, J.; Fridman, D.; Galinkin, E.; Evans, M.; Luna, K.; Derczynski, L.; Pope, N.; Long, E.; Schneider, S.; Siman, G.; Grzegorzek, T.; Ribalta, P.; Katariya, M.; Conway, J.; Saar, T.; Guan, A.; Pawelec, K.; Prayaga, S.; Kuchaiev, O.; Ginsburg, B.; Olabiyi, O.; Briski, K.; Cohen, J.; Catanzaro, B.; Alben, J.; Geifman, Y.; Chung, E.; and Alexiuk, C. 2025. Llama-Nemotron: Efficient Reasoning Models. arXiv:2505.00949.

Betley, J.; Tan, D.; Warncke, N.; Sztyber-Betley, A.; Bao, X.; Soto, M.; Labenz, N.; and Evans, O. 2025. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*.

Cantini, R.; Orsino, A.; Ruggiero, M.; and Talia, D. 2025. Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge. arXiv:2504.07887.

Chen, H.; He, G.; Yuan, L.; Cui, G.; Su, H.; and Zhu, J. 2024. Noise contrastive alignment of language models with explicit rewards. *Advances in Neural Information Processing Systems*, 37: 117784–117812.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Ghosh, S.; Varshney, P.; Sreedhar, M. N.; Padmakumar, A.; Rebedea, T.; Varghese, J. R.; and Parisien, C. 2025. AEGIS2.0: A Diverse AI Safety Dataset and Risks Taxonomy for Alignment of LLM Guardrails. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5992–6026. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Hong, Z.-W.; Shenfeld, I.; Wang, T.-H.; Chuang, Y.-S.; Pareja, A.; Glass, J.; Srivastava, A.; and Agrawal, P. 2024. Curiosity-driven Red-teaming for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Hu, C.; Li, X.; Liu, D.; Wu, H.; Chen, X.; Wang, J.; and Liu, X. 2023. Teacher-student architecture for knowledge distillation: A survey. *arXiv preprint arXiv:2308.04268*.

Huang, J.; Gu, S. S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; and Han, J. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D. M.; Maxwell, T.; Cheng, N.; et al. 2024. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Keskar, N. S.; McCann, B.; Varshney, L. R.; Xiong, C.; and Socher, R. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Kim, G.-H.; Jang, Y.; Kim, Y. J.; Kim, B.; Lee, H.; Bae, K.; and Lee, M. 2025. SafeDPO: A Simple Approach to Direct Preference Optimization with Enhanced Safety. arXiv:2505.20065.

Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Röttger, P.; Kirk, H. R.; Vidgen, B.; Attanasio, G.; Bianchi, F.; and Hovy, D. 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.

Wang, M.; la Tour, T. D.; Watkins, O.; Makelov, A.; Chi, R. A.; Miserendino, S.; Heidecke, J.; Patwardhan, T.; and Mossing, D. 2025. Persona Features Control Emergent Misalignment. arXiv:2506.19823.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36: 80079–80110.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Xu, X.; Li, M.; Tao, C.; Shen, T.; Cheng, R.; Li, J.; Xu, C.; Tao, D.; and Zhou, T. 2024. A Survey on Knowledge Distillation of Large Language Models. arXiv:2402.13116.

Yuan, X.; Li, J.; Wang, D.; Chen, Y.; Mao, X.; Huang, L.; Chen, J.; Xue, H.; Liu, X.; Wang, W.; Ren, K.; and Wang, J. 2025. S-Eval: Towards Automated and Comprehensive Safety Evaluation for Large Language Models. *Proceedings of the ACM on Software Engineering*, 2(ISSTA): 2136–2157.