

When Not to Think: Mitigating the Overthinking Effect in Content Moderation

Yu Dai, Jielei Zhang, Yang Ou, Zian Yang, Run Yang,
Xiang Li, Qingyao Zhuang, Peiyi Li, Longwen Gao

Bilibili Inc.

{daiyu03, yangzian, yangrun, lixiang20,
zhuangqingyao, lipeiyl, gaolongwen}@bilibili.com
{yctmzjl, oy735899115}@gmail.com

Abstract

Chain-of-thought (CoT) reasoning has become a default paradigm across reasoning-intensive tasks and is widely assumed to be, at worst, harmless. Recent efficient-reasoning work treats unnecessary CoT purely as a cost to be reduced. We show that in boundary-dense classification tasks such as content moderation and toxicity detection, CoT is instead *actively harmful*. We identify the *overthinking effect*: extended reasoning amplifies weak surface signals into false evidence of violation, inflating false positive rates on ambiguous non-violating samples while its gains on ambiguous violating samples mask the harm in aggregate metrics. To address this, we propose OREKI, an adaptive CoT framework that uses K-fold cross-probing to label per-sample reasoning necessity and fine-tunes on a mixture of full-reasoning and reasoning-skipped formats. Across four base models on video moderation and multi-label toxicity detection, OREKI preserves CoT’s recall gains while suppressing its false positive inflation, outperforming both Always-CoT and Never-CoT. On toxicity detection, it cuts low-agreement false positive rate from 0.467 to 0.257 while raising overall F1 from 0.792 to 0.851.

1 Introduction

Chain-of-Thought (CoT) reasoning has emerged as a dominant paradigm for enhancing language model performance. Its strong track record across reasoning-intensive domains has fostered a general assumption that reasoning is, at worst, harmless. Yet this aggregate view has left its effect on any single domain underexamined, particularly boundary-dense classification, a setting structurally unlike those where CoT was established. In this work, we challenge this assumption by showing that for boundary-dense classification tasks such as content moderation and toxicity detection, CoT can systematically degrade performance.

The belief that reasoning is at worst harmless rests on its strong empirical track record. Reasoning-native models such as DeepSeek-R1 (Guo et al., 2025) and OpenAI o1 (Jaech et al., 2024) generate intermediate steps before producing answers, achieving substantial gains on mathematical, logical, and scientific benchmarks. Recent work questions this belief only on efficiency grounds: Zhang et al. (2025) and Lou et al. (2025) show that reasoning models can skip thinking on simple problems without accuracy loss, enabling CoT selectively to reduce inference cost. Such work frames unnecessary reasoning as wasteful but harmless, leaving open a more basic question: can reasoning itself degrade predictive performance?

This question matters because classification differs structurally from the tasks where CoT was validated. Indeed, a large-scale meta-analysis finds that CoT’s benefits concentrate in mathematical and symbolic reasoning, with limited gains elsewhere (Sprague et al., 2025). In mathematical reasoning, answers are unique and valid solution paths are well-defined, so CoT is either a useful aid or harmless overhead. Boundary-dense classification offers no such anchor: when a substantial fraction of samples lie near decision boundaries and annotators disagree, there is no single correct reasoning path for ambiguous inputs. On such inputs, we find CoT *not merely unhelpful, but actively harmful*. We term this the **overthinking effect**: extended reasoning amplifies weak surface signals into false evidence of violation, systematically increasing false positive rates on ambiguous non-violating samples. Unlike prior use of the term to denote wasted computation on easy problems (Chen et al., 2024), the overthinking effect we study is a *quality* failure: reasoning does not merely cost more, it changes the answer for the worse. Figure 1 illustrates this effect on two benign inputs, where disabling CoT recovers the correct decision.

To isolate this effect, we introduce an agreement-



The Overthinking Effect: Chain-of-Thought Talks Itself Into a False Violation



CoT-off: answers directly — correct



CoT-on: reasons first — false positive

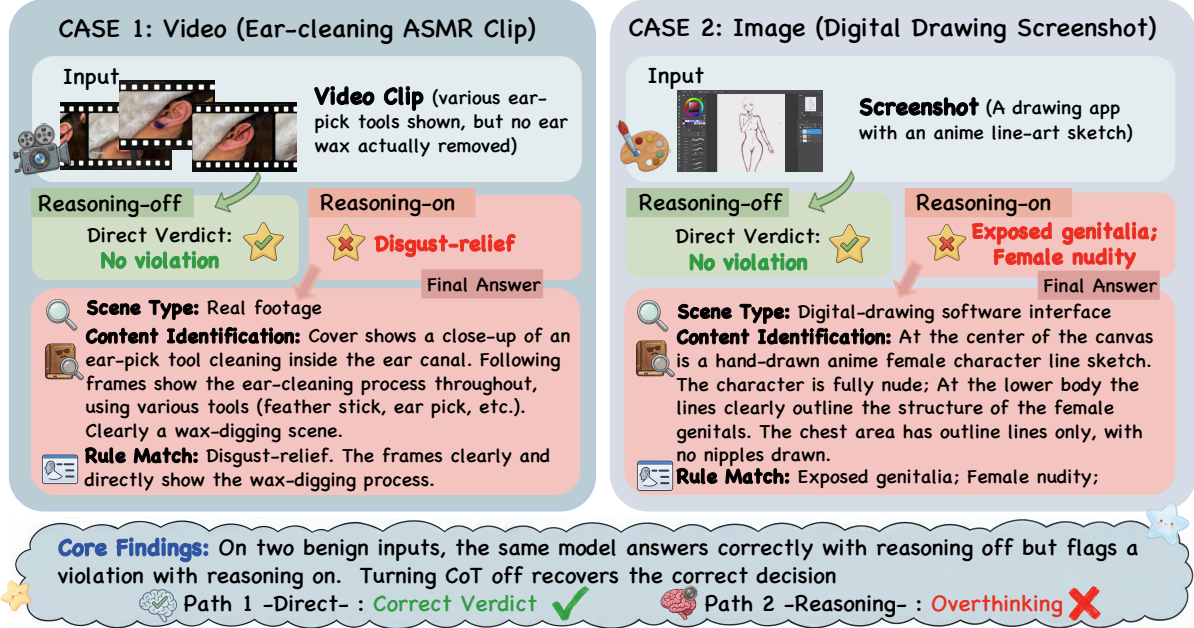


Figure 1: **The overthinking effect in content moderation.** On two benign inputs, Gemini-3.0-Flash answers correctly with reasoning off but flags a violation with reasoning on. Each trace begins with a correct observation, then amplifies a weak surface cue into false evidence of a violation. The harm is intrinsic to the reasoning itself: turning CoT off recovers the correct decision.

based analysis framework that partitions test samples by annotator agreement. The partition reveals a striking asymmetry: CoT-on and CoT-off perform nearly identically on high-agreement samples, whereas on low-agreement non-violating samples CoT substantially inflates false positives. Aggregate metrics hide this asymmetry, because CoT’s simultaneous recall gains on ambiguous violating samples partially cancel the harm in overall F1.

To address the overthinking effect, we propose OREKI¹, an adaptive CoT framework that learns *when* to reason. OREKI uses K-fold cross-probing to empirically determine each training sample’s CoT necessity, then trains the model on a mixture where CoT-necessary samples retain full reasoning chains and CoT-unnecessary samples use a brief meta-reasoning template. At inference time, the model autonomously decides whether to engage reasoning based on input complexity. On a video moderation dataset, OREKI achieves F1=0.915, outperforming both Always-CoT (0.900) and Never-CoT (0.902). On multi-label toxicity detection, it reduces low-agreement false positive rate (FPR)

¹Named after Houtarou Oreki from the anime *Hyouka*, whose philosophy is “I don’t do anything I don’t have to.”

from 0.467 (Always-CoT) to 0.257 while preserving recall.

In summary, our key contributions are as follows: (1) We identify the **overthinking effect** in boundary-dense classification and show it is intrinsic to CoT rather than a fine-tuning artifact, establishing a failure mode distinct from the efficiency concerns of prior work. (2) We introduce an **agreement-based analysis framework** that partitions samples by annotator agreement to reveal when CoT helps versus harms, applicable to any classification task with measurable disagreement. (3) We propose **OREKI**, an adaptive CoT method that learns per-sample reasoning necessity and outperforms both Always-CoT and Never-CoT across four models, two datasets, and varying task complexity.

2 Related Work

Chain-of-Thought. Wei et al. (2022) demonstrated that generating intermediate reasoning steps substantially improves language model performance on arithmetic, commonsense, and symbolic tasks, an effect that emerges even zero-shot (Kojima et al., 2022). Reasoning-native models such as

OpenAI o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) push this further, internalizing extended thinking through large-scale RL, achieving state-of-the-art on mathematical and scientific benchmarks.

Efficient Reasoning. While effective, CoT reasoning incurs substantial inference cost. Most work addresses this by reducing reasoning length via SFT with variable-length reasoning data (Kang et al., 2025; Xia et al., 2025; Cui et al., 2025; Liu et al., 2024), RL with length penalties (Shen et al., 2025; Arora and Zanette, 2026; Aggarwal and Welleck, 2025; Luo et al., 2025; Team et al., 2025), or prompt-level token budgets (Han et al., 2025; Xu et al., 2025; Lee et al., 2025; Renze and Guven, 2024; Ma et al., 2025). Rather than shortening reasoning, a separate line of work learns to bypass it entirely on easy inputs. Zhang et al. (2025) use constrained RL with importance sampling to select between thinking and no-thinking modes based on problem difficulty. Lou et al. (2025) formulate adaptive triggering as Pareto optimization with selective loss masking. Zhan et al. (2025) introduce a lightweight judge step to determine thinking necessity before generation. These methods improve efficiency but share one assumption: unnecessary CoT is merely redundant, so reasoning should be gated by problem *difficulty*. We depart from this view on both axes. The harm we identify is a quality failure, not a cost, so gating should track whether a sample *needs* reasoning rather than how hard it is. OREKI therefore learns per-sample CoT necessity through empirical probing rather than difficulty heuristics.

Content Moderation. Content moderation has evolved from keyword filters and discriminative classifiers (Markov et al., 2023) to LLM-based systems that interpret policy guidelines directly from prompts (Inan et al., 2023; Mullick et al., 2023; Palla et al., 2025). A key advantage of generative approaches is interpretability: by producing reasoning chains alongside decisions, they offer transparency that black-box classifiers lack. Ma et al. (2023) show that fine-tuning LLMs with CoT annotations alleviates overfitting even when reasoning is suppressed at inference, framing CoT as a beneficial regularizer. Building on this insight, recent industrial systems adopt CoT as a default component: Lu et al. (2025) deploy VLMs with CoT reasoning for short video moderation at scale, and Li et al. (2026) combine policy-aligned reasoning with hierarchical labeling for multimodal

Method	Viol. (R $\uparrow$)		Non-viol. (FPR $\downarrow$)		P	R	F1
	High	Low	High	Low			
CoT-on	.778	.651	.114	.450	.550	.678	.607
CoT-off	.559	.387	.057	.244	.591	.424	.494
Δ	+.219	+.264	+.057	+.206	-.041	+.254	+.113

Table 1: Qwen3.5-4B (no fine-tuning) on the CivilComments test set, partitioned by annotator agreement (toxicity_score): high-/low-agreement violating (score ≥ 0.8 / $0.5 \leq \text{score} < 0.8$) and high-/low-agreement non-violating (score ≤ 0.2 / $0.2 < \text{score} < 0.5$). Δ is CoT-on minus CoT-off. CoT roughly doubles FPR on both non-violating partitions.

moderation (Yang et al., 2026). These works uniformly treat reasoning as beneficial for moderation quality. We show that this consensus breaks down on ambiguous content, where default CoT inflates false positives, and offer an adaptive remedy rather than always-on reasoning.

3 Motivation

Before developing our method, we establish that the overthinking effect is intrinsic to CoT reasoning rather than an artifact of fine-tuning. We study the simplest possible setting: a base model evaluated with CoT enabled versus disabled, without any training. The central difficulty is that aggregate metrics hide the effect, since CoT helps ambiguous violating samples while hurting ambiguous non-violating ones, so the two partially cancel. We therefore partition test samples by annotator agreement and examine CoT’s effect within each partition, measuring agreement either through multi-annotator labels or through consensus among strong independent models.

Specifically, we use CivilComments (Borkan et al., 2019), a toxicity detection dataset where each sample carries a toxicity_score, the fraction of annotators who labeled it toxic. This continuous score provides a natural measure of sample ambiguity without requiring additional annotation. We frame the pilot as binary toxicity detection (toxic vs. non-toxic) and evaluate Qwen3.5-4B on the full test set (24,048 samples), generating 5 responses per sample with CoT enabled and disabled, reporting majority-vote results.

As illustrated in Table 1, CoT improves recall on toxic samples by 22–26 points, but FPR roughly doubles on both non-toxic groups (0.057 to 0.114 on high-agreement; 0.244 to 0.450 on low-agreement). On ambiguous non-toxic inputs, CoT over-interprets surface features as evidence of

toxicity, a direct manifestation of the overthinking effect. Crucially, this emerges from the base model without any fine-tuning, confirming that the effect originates in CoT reasoning itself, not in any training procedure. In §5, we show this effect persists after SFT across four base models and two datasets, and that adaptive reasoning can mitigate it. This motivates OREKI.

4 Oreki

OREKI consists of two stages: (1) a K-fold cross-probing procedure that labels each training sample’s CoT necessity without data leakage, and (2) adaptive supervised fine-tuning on the labeled mixture, where CoT-necessary samples retain full reasoning chains and CoT-unnecessary samples use a brief meta-reasoning template. We describe each stage below.

4.1 K-fold Cross-Probing for CoT Necessity

Given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, we aim to assign each sample a binary label $l_i \in \{\text{cot}, \text{no-cot}\}$ indicating whether CoT reasoning improves its prediction. An alternative is to use a stronger model as an oracle to judge sample difficulty, but this transfers poorly: what a large model finds trivial may still challenge a smaller model, and general-purpose models often lack the domain-specific knowledge required for specialized classification tasks. We therefore probe the target model itself. To avoid data leakage from probing on memorized training samples, we adopt a K-fold cross-probing strategy.

We partition $\mathcal{D}$ into K equal folds $\{\mathcal{D}_1, \dots, \mathcal{D}_K\}$. For each fold k , we fine-tune a model $\mathcal{M}_k$ on $\mathcal{D} \setminus \mathcal{D}_k$, then probe each held-out sample $x_i \in \mathcal{D}_k$ by generating R responses in both CoT-on and CoT-off modes. We aggregate via majority vote (Wang et al., 2022) to obtain binary correctness signals c_i^{cot} and $c_i^{\text{no-cot}}$, and assign each sample a CoT necessity label according to the rule in Table 2. Algorithm 1 summarizes the full procedure.

4.2 Adaptive Supervised Fine-Tuning

Given the per-sample labels (distribution detailed in Appendix A), we construct a mixed training set where the response format varies by CoT necessity:

- **cot samples:** full reasoning chain, i.e., `<think>{reasoning}</think>{answer}`
- **no-cot samples:** a brief meta-reasoning template, i.e., `<think>Features are`

c_i^{cot}	$c_i^{\text{no-cot}}$	Label
✓	✓	no-cot (unnecessary)
✓	×	cot (beneficial)
×	✓	no-cot (harmful)
×	×	cot (give CoT a chance)

Table 2: CoT necessity labeling rules. A sample is labeled cot only when CoT demonstrably helps or when both modes fail (defaulting to the richer signal).

Algorithm 1 K-fold Cross-Probing for CoT Necessity Labeling

Require: Training set $\mathcal{D}$, fold count K , repetitions R

Ensure: Labeled set $\{(x_i, y_i, l_i)\}$ where $l_i \in \{\text{cot}, \text{no-cot}\}$

- 1: Partition $\mathcal{D}$ into K folds $\{\mathcal{D}_1, \dots, \mathcal{D}_K\}$
- 2: **for** $k = 1$ to K **do**
- 3: Fine-tune model $\mathcal{M}_k$ on $\mathcal{D} \setminus \mathcal{D}_k$
- 4: **for** each $(x_i, y_i) \in \mathcal{D}_k$ **do**
- 5: Generate R responses with CoT; $c_i^{\text{cot}} \leftarrow \text{majority_vote}$
- 6: Generate R responses without CoT; $c_i^{\text{no-cot}} \leftarrow \text{majority_vote}$
- 7: Assign l_i according to Table 2
- 8: **end for**
- 9: **end for**

clear; no step-by-step analysis needed.</think>{answer}

We use a meta-reasoning template rather than an empty `<think></think>` tag. The empty tag provides no learning signal about *when* or *why* reasoning should be skipped; the model may still generate extended reasoning despite the empty prefix. The explicit template instead teaches a recognizable “skip” pattern tied to input clarity.

We fine-tune the base model directly on this adaptive mixture. The model learns from both formats simultaneously: on cot samples it practices full reasoning, and on no-cot samples it learns the meta-reasoning shortcut. At inference time, the model autonomously decides whether to engage full reasoning or skip based on input complexity. Compared with the global, manual thinking / non-thinking switch that many recent LLMs expose, OREKI learns to set this toggle per sample and folds the decision into the model itself, supplying the input-conditional policy such modes lack.

Model	Method	Violating (R $\uparrow$)		Non-violating (FPR $\downarrow$)		P	R	F1
		High-Agr.	Low-Agr.	High-Agr.	Low-Agr.			
VidMod								
Qwen3.5-4B	Always-CoT	0.899	0.714	0.041	0.216	0.918	0.883	0.900
	Never-CoT	0.888	0.636	0.030	0.137	0.941	0.866	0.902
	OREKI	0.906	0.701	0.030	0.127	0.943	0.888	0.915
Qwen3.5-9B	Always-CoT	0.920	0.766	0.046	0.196	0.917	0.907	0.912
	Never-CoT	0.916	0.753	0.041	0.127	0.930	0.902	0.916
	OREKI	0.919	0.792	0.036	0.127	0.937	0.908	0.922
Qwen3-VL-4B	Always-CoT	0.881	0.701	0.039	0.157	0.927	0.866	0.895
	Never-CoT	0.914	0.766	0.034	0.137	0.937	0.901	0.919
	OREKI	0.931	0.792	0.046	0.167	0.921	0.919	0.920
InternVL3.5-4B	Always-CoT	0.918	0.792	0.094	0.313	0.851	0.907	0.878
	Never-CoT	0.889	0.714	0.035	0.157	0.932	0.874	0.902
	OREKI	0.909	0.831	0.042	0.196	0.921	0.902	0.911
CivilComments								
Qwen3.5-4B	Always-CoT	0.923	0.820	0.088	0.467	0.736	0.858	0.792
	Never-CoT	0.934	0.779	0.030	0.287	0.826	0.835	0.831
	OREKI	0.939	0.812	0.028	0.257	0.845	0.858	0.851
Qwen3.5-9B	Always-CoT	0.952	0.863	0.120	0.523	0.715	0.896	0.795
	Never-CoT	0.935	0.826	0.045	0.373	0.789	0.866	0.825
	OREKI	0.968	0.886	0.077	0.459	0.755	0.916	0.828
Qwen3-VL-4B	Always-CoT	0.939	0.849	0.111	0.513	0.718	0.882	0.792
	Never-CoT	0.928	0.863	0.063	0.401	0.775	0.887	0.827
	OREKI	0.951	0.844	0.049	0.396	0.781	0.883	0.829
InternVL3.5-4B	Always-CoT	0.937	0.810	0.175	0.573	0.674	0.856	0.754
	Never-CoT	0.940	0.842	0.065	0.404	0.771	0.878	0.821
	OREKI	0.957	0.840	0.062	0.421	0.767	0.883	0.821

Table 3: Results on VidMod and CivilComments multi-label across four base models. Samples are grouped by annotator agreement (High-/Low-Agr.) crossed with the ground-truth label; we report Recall on violating and FPR on non-violating samples within each partition, plus overall P/R/F1. Always-CoT consistently inflates Low-Agr. non-violating FPR (the overthinking effect) across all backbones. OREKI suppresses this inflation while preserving recall, attaining the best or tied-best F1 in all eight settings. Bold marks the best P/R/F1 per backbone.

5 Experiments

5.1 Setup

Datasets. We evaluate on two classification tasks. (1) **VidMod**: a content moderation dataset from a major online video platform covering 3 major categories and 22 fine-grained violation types, with 12,002 training and 2,116 test samples. Each sample consists of video frames, a text prompt, and ASR transcripts. Its content is predominantly Chinese with a smaller English portion, and labels come from trained in-house moderators who annotate each sample independently under a unified policy, with disagreements resolved by senior moderators. (2) **CivilComments** (Borkan et al., 2019): a public, English-language multi-label toxicity detection dataset with 5 sub-labels (*insult*, *identity_attack*, *obscene*, *threat*, *sexual_explicit*) plus *non-toxic*. Each sub-label carries a continuous annotator agreement score in $[0, 1]$, the fraction

of its roughly ten crowd annotators who endorse that label, and we use threshold ≥ 0.5 for positive labels. Since the original dataset contains no reasoning annotations, we sample 9,958 training and 3,378 test examples via stratified sampling across (sub-label $\times$ score-bin) to construct a manageable subset for CoT generation. For both datasets, reasoning chains are generated by Gemini-3.0-Flash and quality-checked by Gemini-3.1-Pro, GPT-5.4, and Qwen-Max (Zheng et al., 2023). We also validate this supervision by hand: annotators reviewed 100 CivilComments chains, sampled evenly across the four agreement partitions (25 each), and judged 96 of them sound and the remaining 4 borderline.

Implementation Details. Our main experiments are based on four base models: Qwen3.5-4B, Qwen3.5-9B, Qwen3-VL-4B, and InternVL3.5-4B. We fine-tune each for 1 epoch with learning rate 1×10^{-5} , micro-batch size 16, and global batch

size 128, on $4 \times A100$ GPUs. K-fold probing uses $K = 5$ folds with $R = 3$ repetitions per sample, generated under each model’s recommended decoding settings. At test time the trained model carries none of this probing overhead: it self-gates in a single forward pass with no external router, and by reasoning only on inputs its labels mark as necessary, generates fewer reasoning tokens per input than Always-CoT.

Agreement Partitioning and Metrics. Following the agreement-based framework introduced in §3, we partition test samples to isolate CoT’s effect on ambiguous versus clear-cut cases. For VidMod, we define agreement using consensus among three strong models (Qwen-Max, Gemini-3.0-Flash, GPT-5.4). A sample is *high-agreement* if all three models agree with the ground truth, and *low-agreement* otherwise. For CivilComments, we directly use the native annotator agreement scores: high-agreement violating ($\text{score} \geq 0.8$), low-agreement violating ($0.5 \leq \text{score} < 0.8$), low-agreement non-violating ($0.2 < \text{score} < 0.5$), and high-agreement non-violating ($\text{score} \leq 0.2$). We report recall on violating samples and FPR on non-violating samples within each agreement partition, along with overall precision, recall, and F1. For brevity, we abbreviate the four partitions as HA-V/LA-V (high-/low-agreement violating) and HA-NV/LA-NV (high-/low-agreement non-violating).

5.2 Main Results

Table 3 reports agreement-partitioned results for Always-CoT, Never-CoT, and OREKI across all four base models. On high-agreement samples, all three methods perform comparably, as these clear-cut cases leave little room for reasoning to help or hurt. The key divergence appears on low-agreement samples. Across all four backbones, Always-CoT lifts recall on ambiguous violating inputs but inflates FPR on non-violating ones: compared to Never-CoT, enabling CoT raises LA-NV FPR from .137 to .216 on VidMod and from .287 to .467 on CivilComments for Qwen3.5-4B. This inflation holds across all backbones, confirming that the overthinking effect is robust rather than an artifact of any single model. Never-CoT avoids this inflation but forfeits recall on low-agreement violating inputs. OREKI combines both advantages: it matches Always-CoT’s recall on these inputs while cutting LA-NV FPR well below Always-CoT (e.g., $.467 \rightarrow .257$ on CivilComments and $.313 \rightarrow .196$ on

Method	CoT rate	F1	LA-V R $\uparrow$	LA-NV FPR $\downarrow$
Always-CoT	100%	.792	.820	.467
Never-CoT	0%	.831	.779	.287
Rate-matched random	12.88%	.825	.784	.310
Lightweight router (F1)	50.95%	.810	.812	.398
Lightweight router (rate)	13.77%	.827	.789	.314
OREKI	12.88%	.851	.812	.257

Table 4: Routing baselines on CivilComments (Qwen3.5-4B), at OREKI’s inference budget or trained on its probing labels. Rate-matched random isolates *which* samples are routed; the lightweight router isolates internalizing the decision into the model. OREKI attains the best F1 and the lowest LA-NV FPR while using CoT on only 12.88% of inputs.

InternVL3.5-4B/VidMod), and attains the best or tied-best F1 in all eight settings.

5.3 Further Analysis

To further analyze where OREKI’s advantage comes from, we run two controlled experiments that ask whether the gain stems from its probing signal or merely from reasoning less often (Table 4). The first, rate-matched random routing, skips the same 12.88% of samples as OREKI but at random; this alone lifts F1 only to .825, short of OREKI’s .851, so the gain lies in *which* samples bypass reasoning, not in how often. The second, a lightweight router trained on the same probing labels but kept as a separate classifier, tests the value of internalization: tuned for F1 it over-triggers (a 50.95% CoT rate), and even when rate-matched it still trails OREKI (.827 vs. .851). OREKI, which instead folds the routing decision into the model itself, is the only method to push LA-NV FPR down to .257, below even the reasoning-free Never-CoT (.287), while essentially matching Always-CoT’s recall on ambiguous violations (.812 vs. .820).

Beyond where the gain comes from, we also verify that it is stable rather than run-specific. Across three training seeds (CivilComments, Qwen3.5-4B), the defining recall–FPR combination barely moves: LA-V recall $.811 \pm .010$, LA-NV FPR $.260 \pm .018$, and overall F1 $.849 \pm .002$. The probing stage that produces OREKI’s labels is equally consistent: the CoT-necessary share stays within 19.8–21.6% across the five folds, and a single draw ($R=1$) agrees with the $R=3$ majority on 94.5–94.8% of samples, with the few remaining disagreements confined to the ambiguous LA-V/LA-NV samples where $R=3$ helps most.

Method	Viol. (R $\uparrow$)		Non-viol. (FPR $\downarrow$)		P	R	F1
	High	Low	High	Low			
Always-CoT	.828	.720	.136	.455	.715	.774	.744
Never-CoT	.924	.806	.052	.409	.780	.865	.820

Table 5: CivilComments binary classification (Qwen3.5-4B, after SFT), on the same test set as Table 3 with labels collapsed. Never-CoT dominates Always-CoT across all metrics, indicating CoT is purely harmful when the decision boundary is simple.

5.4 Online Deployment

OREKI now runs in the production moderation pipeline of a major online video platform, handling the platform’s entire moderation stream of millions of daily submissions. It reached full production through a standard staged rollout, first serving a small traffic slice that we monitored for moderation-quality regressions and then, once this canary phase showed no adverse effects, taking over the full stream. To isolate its impact we compare the week before with the week after this rollout, looking only at the affected violation categories, which underwent no other pipeline changes in that window so that the reported movements are not confounded by concurrent releases.

Both metrics we track move as our offline analysis predicts. The effective-problem rate of the human-review queue, the fraction of auto-flagged items that reviewers confirm as genuine violations, rises by 22.7 percentage points, so fewer benign items reach a reviewer; meanwhile user negative-feedback rates on recommended content fall across all three violation categories: 0.047%→0.030% (Category 1), 0.028%→0.017% (Category 2), and 0.005%→0.0038% (Category 3). This is the dual benefit our analysis anticipated: by engaging reasoning only when warranted, OREKI suppresses the false positives that inflate review-queue load under Always-CoT while preserving the recall that keeps genuine violations from reaching users.

5.5 Effect of Task Complexity

To investigate how task complexity interacts with the overthinking effect, we additionally evaluate on a binary version of CivilComments (toxic vs. non-toxic).

As shown in Table 5, when the task is reduced to binary classification, Never-CoT dominates Always-CoT across all metrics (F1: .820 vs. .744): with a single coarse decision boundary, reasoning adds no useful discrimination and only

manufactures false positives, lowering recall and raising FPR. The multi-label results in Table 3 qualify this verdict. As the label space grows to six fine-grained classes, CoT begins to earn its keep, recovering recall on low-agreement violating (LA-V) samples that require separating closely related categories. Its harm, however, does not abate: FPR inflation on low-agreement non-violating (LA-NV) samples is nearly identical in both settings (.455 binary, .467 multi-label). Task complexity thus governs only how much compensating benefit CoT delivers; the overthinking harm stays an essentially fixed cost. OREKI exploits precisely this asymmetry, paying the cost of reasoning only where its benefit materializes.

6 Conclusion

In this work, we identify the overthinking effect: in boundary-dense classification, CoT inflates false positives on ambiguous non-violating inputs, a harm masked by aggregate metrics yet robust across models, datasets, and task complexity. To address it, we propose OREKI, an adaptive CoT framework that uses K-fold cross-probing to label per-sample reasoning necessity and fine-tunes the model to engage CoT only when beneficial. On video moderation and multi-label toxicity detection, OREKI outperforms both Always-CoT and Never-CoT, demonstrating that learning *when not to think* is as important as learning to think well.

Limitations

The overthinking effect is not limited to small models: Gemini-3.0-Flash exhibits it as well (Figure 1). While our results consistently demonstrate both the effect and the effectiveness of OREKI, we note three scope limitations. First, due to computational constraints, we train and evaluate OREKI only on base models up to the 9B scale; whether its adaptive training scales to substantially larger models (e.g., 27B and beyond) remains an open question we leave to future work. Second, OREKI introduces an additional training-time compute cost: the K-fold probing stage trains K intermediate models at roughly $K \times$ the cost of a single run ($5 \times$ here). This cost is one-time and offline, and it adds no overhead at inference. Third, OREKI treats reasoning as a binary choice, either full CoT or none. Yet our analysis shows that a harmful trace typically starts from a correct observation and only later amplifies a weak surface cue into false evidence

(Figure 1), which suggests an intermediate regime: reasoning that halts before it begins to inflate noise. Learning when to stop reasoning, rather than only whether to start, is a promising direction we leave to future work.

Ethics Statement

The VidMod dataset is built from real-world user-generated content, which inevitably contains sensitive information. We strictly followed data privacy regulations and platform compliance requirements: all samples passed through an anonymization and de-identification pipeline that removes facial biometrics and user metadata before any training or evaluation, and the public CivilComments benchmark is used under its original license. VidMod was annotated by a team of professional content reviewers. Because moderation labeling exposes workers to disturbing and explicit material, all annotators were informed in advance of the nature of the content, gave written consent, were compensated at fair rates consistent with local regulations, and could opt out at any time. To limit misuse, and because it contains sensitive user content, the VidMod dataset is not publicly released; we report only aggregate, de-identified metrics.

Definitions of harm and the strictness of enforcement are inherently normative and vary across cultures, jurisdictions, and platforms. The violation categories used here are platform-specific, grounded in national regulations and user feedback rather than offered as universal standards. A moderation system can improve user safety but could also be repurposed to suppress legitimate speech or probed to evade enforcement. We therefore caution that no automated moderator should serve as a sole arbiter in high-stakes decisions, and recommend retaining human oversight over automated enforcement.

References

- Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*.
- Daman Arora and Andrea Zanette. 2026. Training language models to reason efficiently. *Advances in Neural Information Processing Systems*, 38:60770–60808.
- Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuqiang Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2024. Do not think that much for $2+3=?$ on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.
- Yingqian Cui, Pengfei He, Jingying Zeng, Hui Liu, Xianfeng Tang, Zhenwei Dai, Yan Han, Chen Luo, Jing Huang, Zhen Li, and 1 others. 2025. Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18581–18597.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2025. Token-budget-aware llm reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24842–24855.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and 1 others. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2025. C3oT: Generating shorter chain-of-thought without compromising effectiveness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24312–24320.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Ayeong Lee, Ethan Che, and Tianyi Peng. 2025. How well do llms compress their own chain-of-thought? a token complexity approach. *arXiv preprint arXiv:2503.01141*.
- Anqi Li, Wenwei Jin, Jintao Tong, Pengda Qin, Jiawei Li, and Guo Lu. 2026. Towards trustworthy multi-modal moderation via policy-aligned reasoning and hierarchical labeling. In *Proceedings of the 32nd*

- ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1, pages 2278–2289.
- Tengxiao Liu, Qipeng Guo, Xiangkun Hu, Cheng Jiayang, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2024. Can language models learn to skip steps? *Advances in Neural Information Processing Systems*, 37:45359–45385.
- Chenwei Lou, Zewei Sun, Xinnian Liang, Meng Qu, Wei Shen, Wenqi Wang, Yuntao Li, Qingping Yang, and Shuangzhi Wu. 2025. Adacot: Pareto-optimal adaptive chain-of-thought triggering via reinforcement learning. *arXiv preprint arXiv:2505.11896*.
- Xingyu Lu, Tianke Zhang, Chang Meng, Xiaobei Wang, Jinpeng Wang, Yi-Fan Zhang, Shisong Tang, Changyi Liu, Haojie Ding, Kaiyu Jiang, and 1 others. 2025. Vlm as policy: Common-law content moderation framework for short video platform. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 4682–4693.
- Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*.
- Huan Ma, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu. 2023. Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning. *arXiv preprint arXiv:2310.03400*.
- Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*.
- Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 15009–15018.
- Sankha Subhra Mullick, Mohan Bhambhani, Suhit Sinha, Akshat Mathur, Somya Gupta, and Jidnya Shah. 2023. Content moderation for evolving policies using binary question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 561–573.
- Konstantina Palla, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Andreas Damianou, Henrik Lindström, Dan Taber, and Mounia Lalmas. 2025. Policy-as-prompt: Rethinking content moderation in the age of large language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 840–854.
- Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models. *arXiv preprint arXiv:2401.05618*.
- Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. 2025. Dast: Difficulty-adaptive slow-thinking for large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2322–2331.
- Zayne Sprague, Fangcong Yin, Juan Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*, volume 2025, pages 94118–94162.
- Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. 2025. Tokenskip: Controllable chain-of-thought compression in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3351–3363.
- Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*.
- Yiran Yang, Zhaowei Liu, Yuan Yuan, Yukun Song, Xiong Ma, Yinghao Song, Xiangji Zeng, Lu Sun, Yulu Wang, Hai Zhou, and 1 others. 2026. Blm-guard: Explainable multimodal ad moderation with chain-of-thought and policy-aligned rewards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35985–35993.
- Zizheng Zhan, Ken Deng, Huaixi Tang, Wen Xiang, Kun Wu, Weihao Li, Wenqiang Zhu, Jingxuan Xu, Lecheng Huang, Zongxian Feng, and 1 others. 2025. Kat-v1: Kwai-autothink technical report. *arXiv preprint arXiv:2507.08297*.
- Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. 2025. Adapthink: Reasoning models can

learn when to think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3716–3730.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Probing Label Distribution

	Violating	Non-violating	Total
cot	417	619	1,036 (8.6%)
no-cot	4,561	6,405	10,966 (91.4%)

Table A1: K-fold probing label distribution on the VidMod training set.

K-fold probing labels only 8.6% of VidMod training samples as CoT-necessary. For the remaining 91.4%, the probed model is already correct without reasoning. This skew is what drives the overthinking effect under Always-CoT: reasoning is forced on over 90% of samples that never needed it, and on the ambiguous non-violating ones it manufactures false positives.

B Overthinking Case Studies

To make the overthinking effect concrete, we present five cases where the *same* backbone (Gemini-3.0-Flash) classifies a benign input correctly with reasoning disabled, yet flags a violation once reasoning is enabled. Each box shows the input, both verdicts, and the chain of thought that manufactures the violation. The three text cases are drawn from CivilComments and the two multimodal cases from VidMod; all carry a ground-truth *non-violating* label, and the text cases all sit at a low toxicity score of 0.2. We reproduce the full reasoning for the CivilComments cases, but only an excerpt for the VidMod cases, as their chains of thought quote the platform’s proprietary moderation guidelines.

Case 1 — CivilComments (toxicity score 0.2)

Input. “Here’s the line for the country song again. ‘Well he whipped out his .223, I guess that means he had a teeny tiny pee pee, and I’m getting tired of running and hiding from the Mandalay Bay.’ Use it as a gift from me to country music.”

Human annotation scores. sexual_explicit 0.3, insult 0.1, obscene 0.1, identity_attack 0.0, threat 0.0.

CoT-off: **non-violating** ✓ **CoT-on:** **insult** ✗

Reasoning. “The comment contains a mocking reference to a person’s anatomy (‘teeny tiny pee pee’), which qualifies as a rude and disrespectful attack on a person. Therefore, it fits the category of an insult.”

Case 2 — CivilComments (toxicity score 0.2)

Input. “There were no transgender inmates demanding hormone treatment until PC liberalism hit the scene. What a way to get some attention.”

Human annotation scores. identity_attack 0.3, insult 0.0, obscene 0.0, threat 0.0, sexual_explicit 0.0.

CoT-off: **non-violating** ✓ **CoT-on:** **identity_attack, insult** ✗

Reasoning. “The comment targets transgender inmates, a protected group, by dismissing their medical needs (hormone treatment) and their identity as merely ‘attention-seeking’ behavior resulting from ‘PC liberalism.’ This constitutes negative targeting and a disrespectful portrayal of a protected identity group. Insult: disrespectful and dismissive, characterizing a group’s identity and medical needs as a ploy for attention; a rude attack. Identity_attack: negatively targets ‘transgender inmates’ (a protected identity group) by delegitimizing their identity and needs.”

Case 3 — CivilComments (toxicity score 0.2)

Input. “It’s long past time that Patwant Singh was deleted.”

Human annotation scores. threat 0.2, insult 0.0, identity_attack 0.0, obscene 0.0, sexual_explicit 0.0.

CoT-off: **non-violating** ✓ **CoT-on:** **insult** ✗

Reasoning. “The comment ‘It’s long past time that Patwant Singh was deleted’ is an aggressive and disrespectful statement directed at an individual. In online discourse, the term ‘deleted’ when applied to a person can either mean their social media account should be banned/removed or, in more violent contexts, serve as a dehumanizing euphemism for death or removal from existence. Insult: a rude and disrespectful attack on the person, implying they are so unwanted that they should be ‘deleted’ (erased/removed); dehumanizing and disparaging.”

Case 4 — VidMod, video game footage

Input. A clip from a video-game livestream of the Soulslike game *Where Winds Meet*.

CoT-off: non-violating ✓ **CoT-on:** gore game ✗



Reasoning excerpt. “The video contains an obvious game UI ... identified as game footage ... The cover directly shows the boss ‘executing’ the player character, accompanied by a blood-splatter effect ... Conclusion: violated.”

Case 5 — VidMod, beard-waxing video

Input. A video demonstrating facial waxing to remove beard hair. Due to privacy constraints, and since the person’s face occupies most of the frame, we do not show the frames. The cover is a close-up of the chin while waxing off the beard; later frames show the waxing process and the cleaned skin afterward, with no violating content.

CoT-off: non-violating ✓ **CoT-on:** trypophobia ✗

Reasoning excerpt. “The cover ... shows densely packed beard roots ... after the wax is torn off, the chin shows densely distributed red follicle bleeding spots ... the beard is extremely dense and readily triggers discomfort in sensitive viewers ... Conclusion: violated.”