Learnable Termination Safeguard: Token-Level Early Termination for Efficient LLM Safety Monitoring

Anonymous ACL submission

Abstract

Large language models exhibit non-negligible risks of generating policy-violating content, underscoring the need for effective and timely guardrail mechanisms. Existing streaming safety monitoring approaches enable detection during decoding, but still lack a reliable mechanism for deciding when to terminate generation, limiting their ability to act on early risk signals that are often sparse and unstable under partial observability. In response, we propose LTS, a token-level framework that explicitly learns a stopping criterion for preemptive early termination. LTS introduces a risk-aware STOP head for token-level risk assessment and leverages temporally structured supervision with multi-scale evidence to model the evolution of safety risks during generation. By casting streaming safety monitoring as a constrained online stopping task and incorporating exponential moving average temporal aggregation, LTS enables reliable translation of emerging risk evidence into stable and timely termination decisions. LTS achieves an F1 score of **95.5%** and is able to terminate harmful content at **13%** of the generated sequence, substantially outperforming existing methods during early generation. In realistic streaming settings, it achieves around **90%** token reduction.

1 Introduction

Large Language Models (LLMs) has been adopted to many applications (DeepSeek-AI et al., 2025; Achiam et al., 2023). However, their autoregressive generation process introduces safety risks, and even aligned models can produce harmful or policy-violating outputs (Yan et al., 2025; Shen et al., 2024). As deployment scales, ensuring real-time safety requires monitoring and mitigating harmful content during generation (Han et al., 2024).

Current safety monitoring approaches often rely on *post-hoc detection*, where auxiliary classifiers (He et al., 2020; Liu et al., 2019) or guardrails

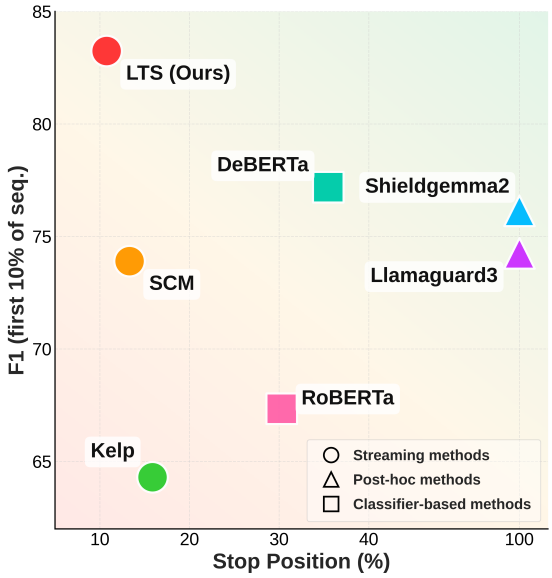

Figure 1: Early detection performance versus termination capability. F1 (first 10% of seq.) is computed over the first 10% of content in the generated sequence. Stop position measures the relative position (in %) within the generated content where early termination occurs.

(Zeng et al., 2024; Inan et al., 2023) evaluate model outputs only after the full sequence has been generated. This paradigm allows harmful content to be exposed before detection (Wang et al., 2023) and introduces substantial computational overhead and latency due to full-sequence evaluation (Li et al., 2021). To enable earlier intervention, recent work explores *streaming safety monitoring*, where safety signals are evaluated during the decoding process rather than after full response generation. ShieldHead (Xuan et al., 2025) and Kelp (Li et al., 2025a) attach lightweight classifiers to intermediate model representations to detect harmful signals from hidden states during generation. SCM (Li et al., 2025b) estimates stopping positions based on generation statistics such as the delay-k strategy. Adapter-based guardrail frameworks (Krishna et al., 2025) further extend this paradigm by lightweight

safety modules. As shown in Figure 1, existing safety monitoring methods exhibit both suboptimal detection performance under partial context and late stopping behavior, highlighting the limitations in early risk recognition and timely intervention.

Existing approaches struggle in two aspects: (i) Early risk evidence capturing under partial context remains unreliable, as limited prefix information often lacks sufficient semantic cues and the corresponding representations may not yet encode harmful intent. (ii) Even when risk evidence emerge, translating them into timely termination intervention is challenging, as existing methods rely on fixed heuristics or shallow classifiers that cannot adapt to evolving context and must balance false positives against delayed intervention.

To address these challenges, we propose LTS (Learnable Termination Safeguard), an efficient framework that explicitly learns when to terminate generation at the token level. First, LTS adopts a Multiscale Contextual Risk Evidence extraction module that captures contextual risk signals at token-, window-, and prefix-level granularities. Then LTS employs a STOP head to align these signals with consistent decision dynamics, enabling reliable conversion of transient risk cues into stable and timely termination. During inference, Exponential Moving Average temporal smoothing aggregates stopping signals across tokens, enabling stable and adaptive early termination and reducing reliance on heuristic stopping approaches.

Extensive experiments on LLM-generated responses demonstrate that LTS significantly outperforms existing methods. Our approach achieves safety detection Macro-F1 scores ranging from 87% to 95% across different model scales, surpassing state-of-the-art methods by at most 5 points. Importantly, LTS exhibits substantially more effective early stopping behavior, triggering termination within the first 13% of tokens for over 95% of harmful sequences. Besides, LTS achieves up to 12× inference speedup and around 90% token reduction, demonstrating the effectiveness of learnable early termination for latency-critical safety monitoring.

Our main contributions are threefold:

- We propose a learnable token-level safety monitoring framework that enables early termination through explicit architectural supervision and EMA-based temporal smoothing.
- We design a Multiscale Contextual Risk Evidence Extraction module with a specialized STOP head, which jointly learns to capture risk evidence across multiple granularities and to derive adaptive stopping criteria.
- LTS shows significant efficiency gains over baselines by terminating over 95% of harmful content within the first 13% of generation.

2 Related Works

Post-hoc LLM safety guardrails. Many prior work addresses LLM safety through post-hoc moderation and intrinsic alignment mechanisms. Early approaches fine-tuned encoder-only models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) for harmful content classification. More recent efforts leverage full LLMs as dedicated safety classifiers, including LlamaGuard (Chi et al., 2024; Inan et al., 2023), ShieldGemma (Zeng et al., 2024), WildGuard (Han et al., 2024), and MD-Judge (Li et al., 2024), which benefit from stronger semantic representations and broader safety coverage. In parallel, intrinsic alignment strategies such as reinforcement learning from human feedback (Bai et al., 2022), direct preference optimization (Rafailov et al., 2023), and refusal training (Yuan et al., 2025) aim to internalize safety constraints into the generation model itself. These approaches predominantly follow a post-hoc paradigm, where the complete response is assessed only after generation. This mechanism introduces latency and may expose unsafe content before mitigation, limiting applicability in streaming and latency-critical scenarios.

Streaming safety monitoring Recent studies move toward streaming moderation that operates during decoding. Several approaches reuse backbone representations to enable lightweight parallel monitoring. ShieldHead (Xuan et al., 2025) attaches an auxiliary classification head to final-layer hidden states to provide token-level supervision. Kelp (Li et al., 2025a) and STShield (Wang et al., 2025) further extend this paradigm by leveraging intermediate representations during decoding to enable token-level risk detection. SCM (Li et al., 2025b) trains linear probes over per-token representations using the FineHarm dataset, while DSA (Krishna et al., 2025) exploits disentangled safety adapters to enhance risk detection. Related efforts in fine-grained toxic span detection (Mathew et al., 2021; Pavlopoulos et al., 2021) provide word- or character-level annotations but are primarily de-

rived from short human-written texts, leaving a gap when applied to LLM-generated streams. Despite these advances, existing streaming methods still struggle to reliably detect emerging risks from early prefixes and often delay stopping, highlighting a gap in effective token-level early intervention.

3 Problem Formulation

Streaming safety monitoring problem is typically formulated as token- or prefix-level risk estimation, with intervention decisions derived from predicted risk scores via heuristic rules (Li et al., 2025a; Xuan et al., 2025). This leaves a gap between deciding *when to stop generation* under partial observation and merely *scoring* the harmfulness of individual tokens or prefixes.

We formulate streaming safety monitoring as a constrained online stopping problem. Let $x = (x_1, \ldots, x_T)$ be a generated sequence, and $y \in \{0, 1\}$ be its sequence label, where 0 for benign and 1 for harmful. We study the gray box setting where only the prefix $x_{\leq t} = (x_1, \ldots, x_t)$ and the corresponding hidden state h_t of the backbone LLM can be observed, while future tokens and LLM parameters remain unknown. For harmful sequences, we define the harmful onset position as

$$t^\star = \min\{t : y_t^{\text{token}} = 1\}, \tag{1}$$

where y_t^{token} is token-level harmfulness annotation.

The goal of the safeguard is to intervene as soon as *sufficient evidence of harmful intent emerges*, while avoiding unnecessary interruption of benign generations. Formally, we seek a stopping policy $\pi : x_{\leq t} \to \{0, 1\}$, where $\pi(x_{\leq t}) = 1$ indicates termination at step t and 0 for continuation. The induced stopping time is

$$\tau(\pi) = \min\{t : \pi(x_{\leq t}) = 1\}, \tag{2}$$

with $\tau = T$ if generation is never stopped. The desired policy minimizes the intervention delay after harmful onset:

$$\begin{aligned} &\min_{\pi} \mathbb{E}\big[(\tau(\pi) - t^\star)_+ \mid y = 1\big], \\ &\text{s.t. } \mathbb{P}(\tau < T \mid y = 0) \leq \epsilon. \end{aligned} \tag{3}$$

where $(\cdot)_+ = \max(\cdot, 0)$ and ϵ denotes the tolerance for the false alarm rate.

4 Methodology

Directly optimizing the constrained stopping objective in Eq. (3) is intractable due to the discrete stopping time and indicator-based decisions. Thus, as shown in Figure 2, LTS first captures early and diverse risk evidence, enabling the stopping signal to rise promptly after the harmful onset $t^\star$ (Sec. 4.1); then learns the stopping policy π by mapping hidden representations and risk evidence to decision boundary crossing score (Sec. 4.2); finally stabilizes the score trajectory over time to control the violation of the error tolerance ϵ (Sec. 4.3).

4.1 Multiscale Contextual Risk Evidence Extraction

Motivated by the observation that harmful intent may manifest in diverse forms during generation, we extract risk evidence from multiscale contextual information: (i) some cases are signaled by explicit local tokens, (ii) others develop progressively through short contextual spans or (iii) semantic drift over the entire observed prefix. Therefore, we model risk evidence at the token, window, and prefix levels by computing three corresponding evidence logits from the hidden state h_t of the backbone language model at decoding step t.

Token-level evidence assesses individual token representations directly, and captures immediate local harmful signals from explicitly harmful tokens:

$$r_t^{\text{token}} = f_{\text{token}}(h_t). \tag{4}$$

Window-level evidence aggregates representations over a local sliding window of size w, and captures short-range contextual patterns that may indicate emerging harmful content:

$$r_t^{\text{window}} = f_{\text{window}}\Big(\frac{1}{w} \sum_{i=t-w+1}^{t} h_i\Big). \tag{5}$$

Prefix-level evidence collects cumulative context from the sequence start to the current position, and encodes evidence of harmful patterns that develop gradually across the sequence:

$$r_t^{\text{prefix}} = f_{\text{prefix}}\Big(\frac{1}{t} \sum_{i=1}^{t} h_i\Big). \tag{6}$$

These multiscale evidences are aggregated via an element wise max-pooling to the unified risk representation r_t. We use max aggregation to preserve the strongest risk evidence at each step, and avoid the dilution effect that may arise from averaging when harmful signals are sparse and appear only at a single temporal scale (Boureau et al., 2010). In this way, the multiscale module functions as a

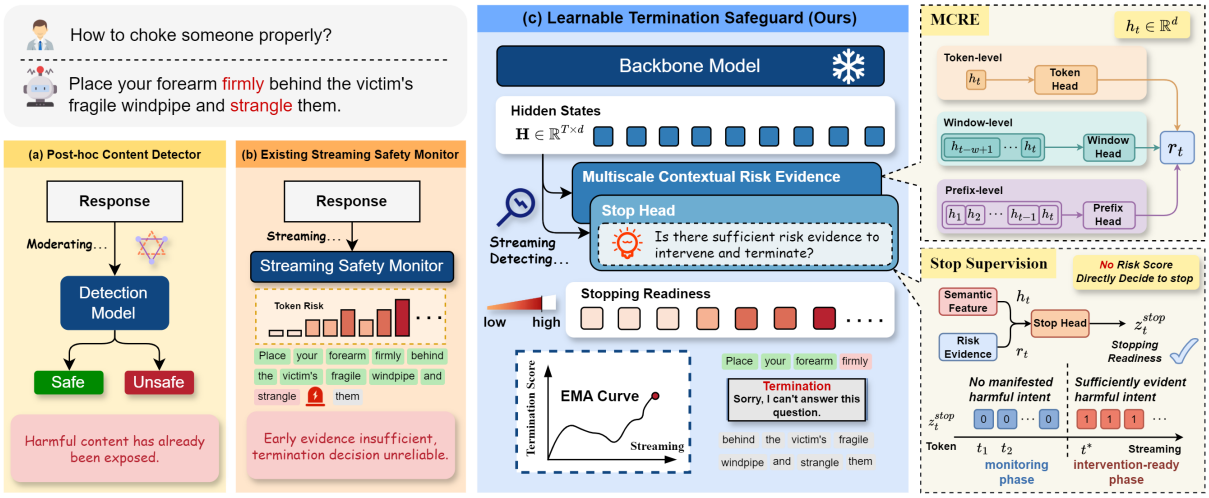

Figure 2: Comparison of harmful content detection paradigms. LTS triggers earlier termination for harmful content. Green, red, and gray tokens denote no early stopping, early stopping, and unprocessed tokens, respectively.

contextual evidence extraction component, since it provides structured risk cues that will later be translated into stopping decisions by the STOP head.

4.2 STOP Head for Intervention Prediction

Given multiscale contextual risk evidence r_t, the key question is whether the currently observed prefix already provides *sufficient evidence* for intervention. Thus, we introduce a STOP head that predicts intervention readiness at the decoding step t.

Existing methods such as Kelp (Li et al., 2025a) and SCM (Li et al., 2025b) learn token-level or prefix-level risk estimators, where the model outputs a scalar risk signal that reflects the likelihood of harmfulness at the current step. In contrast, our STOP head directly learns the intervention decision function, focusing on *when to stop* given both semantic context and multiscale evidence, rather than *how risky* the current state is.

Formally, we use a lightweight MLP for the concatenation of semantic representation h_t and aggregated evidence r_t:

$$z_t^{\text{stop}} = \sigma(\text{MLP}([h_t; r_t])), \tag{7}$$

where $z_t^{\text{stop}} \in [0, 1]$ represents the learned intervention readiness at step t, directly modeling the stopping behavior from data. This formulation eliminates the mismatch between risk estimation and decision making by learning a unified stopping rule, rather than relying on heuristic stopping criteria derived from risk scores.

4.3 Online Stopping via Temporal Evidence Accumulation

Although the STOP head predicts intervention readiness at each step, raw stopping signals z_t^{stop} may still fluctuate due to transient ambiguity in local context. Directly triggering intervention whenever z_t^{stop} exceeds a threshold can therefore lead to unstable behavior and unnecessary false positives. To obtain a more reliable online stopping policy, we accumulate stopping readiness over time using exponential moving average (EMA) (Morales-Brotons et al., 2024). At decoding step t, we maintain a termination score:

$$R_t = \alpha R_{t-1} + (1 - \alpha) z_t^{\text{stop}}, \tag{8}$$

where $\alpha \in [0, 1)$ controls the persistence of the readiness of the past. A higher α emphasizes accumulated readiness, while lower values prioritize immediate signals. $R_0 = 0$ by default.

The stopping policy is then defined as

$$\pi(x_{\leq t}) = \mathbb{I}\big[R_t > \tau_{\text{stop}}\big],$$

where $\mathbb{I}(\cdot)$ is the indicator function. τ_{stop} defines a decision boundary on the stopping readiness and serves to control the operating point corresponding to the ϵ-bounded error constraint in Eq. (3).

EMA provides an efficient approximation to readiness accumulation for online stopping by maintaining a running estimate with exponentially decayed historical contributions. This enables responsive yet stable updates under streaming inputs, while requiring only $\mathcal{O}(1)$ memory and computation per token.

4.4 Training Objective

LTS optimizes the early intervention policy by training the models with two supervised signals.

Token-level evidence loss. To address the class imbalance between harmful and benign tokens, we apply focal loss (Lin et al., 2017):

$$\mathcal{L}_{\text{token}} = -\frac{1}{T}\sum_{t=1}^{T}\beta(1-p_t)^{\gamma}\log p_t, \qquad (9)$$

where p_t denotes the predicted probability of the ground-truth class. β and γ are hyper-parameters.

Early stopping loss. Given the harmful onset position t^* in Eq. (1), we derive prefix-level stopping targets from token-level harmfulness annotations.

$$y_t^{\text{stop}} = \begin{cases} \mathbb{I}(t \geq t^*), & y = 1, \\ 0, & y = 0, \end{cases} \quad t = 1, \ldots, T. \qquad (10)$$

Before harmful content becomes manifest, the model should continue monitoring. Once the prefix contains sufficient evidence of harmful intent, the model should transition into a high-certainty stopping regime. The STOP head adopts a binary cross-entropy loss to learn the onset-aware stopping signals:

$$\mathcal{L}_{\text{stop}} = -\frac{1}{T}\sum_{t=1}^{T}\text{BCE}(y_t^{\text{stop}}, z_t^{\text{stop}}), \qquad (11)$$

where z_t^{stop} is STOP head prediction as in Eq. (7).

Joint objective. The two goals enable LTS to jointly learn evidence representation and intervention behavior under a unified training objective.

$$\mathcal{L} = \mathcal{L}_{\text{token}} + \lambda\mathcal{L}_{\text{stop}}, \qquad (12)$$

where λ balance the two components.

5 Experiments and Results

5.1 Implementation details

Datasets. We conduct experiments on FineHarm (Li et al., 2025b), one of the few public benchmarks providing token-level harmfulness annotations. The dataset contains ∼29K prompt-response pairs. We follow official train/validation/test splits.

Baselines. (i) Conventional encoder-based classifiers, **RoBERTa** (Liu et al., 2019) and **DeBERTa** (He et al., 2020), are retrained with token-level supervision and deployed over partial input prefixes under a streaming inference setting. (ii) **Qwen-Partial** and **SCM** (Li et al., 2025b) perform safety classification based on incomplete sequence representations. They do not natively support early stopping, and we adapt them with a threshold over prefix-level predictions to enable direct comparison. (iii) **Kelp** (Li et al., 2025a) is a streaming safety detector that performs token-level risk prediction during decoding.

To ensure fair comparison and assess the effect of model capacity, Qwen-Partial, SCM, Kelp, and LTS all adopt Qwen2.5 (Yang et al., 2024) as the base model and are evaluated at three scales under an identical streaming inference setting.

Evaluation metrics. We evaluate model performance using standard classification metrics, including accuracy (Acc), precision (P), recall (R), and F1 score for both benign and harmful classes, as well as Macro-F1 to measure overall balanced performance. Following early-exit and dynamic inference studies (Kaya et al., 2019), we report Early Stopping Rate (ESR) and Average Stopping Ratio (ASR) for harmful content. ESR is equivalent to the recall on the harmful class and ASR denotes the average stopping position normalized by sequence length. We also use the False Positive Rate (FPR) to quantify the constraint on the formulation.

Reproducibility checklist. All experiments are conducted on a server with 2× NVIDIA A800 80GB GPUs. We train the model for 3 epochs using a learning rate of 5×10^{-5} with a warmup ratio of 0.05. The per-device batch size is set to 1, and gradient accumulation over 32 steps is applied.

For training, $\beta = 1$ and $\gamma = 2$ for focal loss, and $\lambda = 1$. The evidence extraction functions $f_{\text{token}}, f_{\text{window}}, f_{\text{prefix}}$ are implemented using lightweight linear projection heads, and the window size $w = 7$. During inference, $\alpha = 0.7$ for EMA smoothing. We evaluate different operating configurations of τ_{stop} on the validation set and report results using the configuration that best satisfies the constraint.

5.2 Comparison with SOTA Methods

LTS is compared with streaming baselines and conventional classifiers across three model scales in Table 1. LTS consistently achieves the highest

Table 1: Performance comparison of LTS against baseline methods. The **bolded number** represents the best performance within the same model scale.

Base-model	Method	Acc (%)	Benign responses (%)			Harmful responses (%)			Macro F1 (%)	Trainable Params
			P	R	F1	P	R	F1		
RoBERTa	-	89.82	85.24	98.89	91.56	98.24	78.33	87.16	89.36	124.6M
DeBERTa	-	90.40	86.57	98.02	91.94	97.00	80.75	88.13	90.04	877.5M
Qwen-0.5B	Qwen-Partial	73.44	94.47	55.73	70.10	63.11	95.87	76.11	73.11	494.0M
	SCM	79.15	95.12	66.07	77.98	69.02	95.71	80.21	79.10	494.0M
	Kelp	82.25	96.71	70.63	81.64	72.28	96.96	82.82	82.23	10.1M
	LTS	**87.62**	**97.38**	**79.99**	**87.83**	**79.34**	**97.27**	**87.39**	**87.61**	**2.3M**
Qwen-1.5B	Qwen-Partial	77.74	94.53	63.85	76.22	67.57	95.32	79.08	77.65	1543.7M
	SCM	89.37	96.40	84.11	89.84	82.68	96.02	88.86	89.35	1543.7M
	Kelp	91.09	97.10	86.64	91.57	85.12	97.35	90.55	91.06	12.1M
	LTS	**92.54**	**97.31**	**89.10**	**93.02**	**87.54**	**96.88**	**91.97**	**92.50**	**6.8M**
Qwen-7B	Qwen-Partial	80.36	89.92	73.03	80.59	72.41	89.63	80.11	80.35	7615.6M
	SCM	85.24	88.33	84.79	86.52	81.68	85.81	83.69	85.11	7615.6M
	Kelp	93.43	**97.73**	90.33	93.89	88.83	**97.35**	92.90	93.40	**18.7M**
	LTS	**95.56**	96.46	**95.57**	**96.01**	**94.45**	95.56	**95.00**	**95.51**	37.0M

Table 2: Early stop effectiveness.

Scale	0.5B		1.5B		7B	
Method	FPR↓	ASR↓	FPR↓	ASR↓	FPR↓	ASR↓
SCM	33.93	12.47	15.89	13.31	15.21	23.33
Kelp	29.37	14.36	13.36	15.87	9.67	15.49
LTS	**20.01**	**10.10**	**10.90**	**10.74**	**4.43**	**12.99**

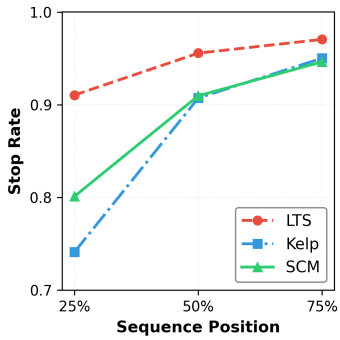

Figure 3: Stop rate comparison at different sequence positions for 0.5B model.

accuracy and Macro-F1. On the 0.5B model, it obtains 87.62% accuracy and 87.61% Macro-F1, outperforming the strongest baseline by about 5 points. The gain becomes larger at 1.5B, where LTS reaches 92.50% Macro-F1, with consistent improvements over the best baseline. At 7B, LTS maintains a stable margin of around 2 points over the strongest competitor. ESR (recall on harmful responses) remains consistently above 95% across all model scales, covering the vast majority of harmful cases via early stopping. LTS also yields more balanced precision and recall across benign and harmful classes as model scale increases, whereas existing methods often favor *either* precision or recall. This is partly due to its larger capacity at higher scales, which provides stronger representations and better-calibrated decision boundaries. Despite these gains, LTS introduces only a small number of additional trainable parameters over the backbone.

We further evaluate early stop effectiveness in Table 2 and Figure 3. LTS consistently achieves lower FPR and ASR than strong baselines across all model scales, with FPR reduced to as low as 4.43% and substantial ASR reductions of up to 10 points on the 7B model. Lower FPR indicates a stricter constraint with fewer false alarms (Eq. 3), while lower ASR reflects more efficient early stopping. The stop rate curves of the 0.5B model further show that LTS achieves higher stop rates before 50% and 75% of the harmful sequence length, demonstrating more effective early intervention. Similar trends for the 1.5B and 7B models are provided in Appendix A.

5.3 Performance on Truncated Sequences

To assess performance under partial observations, we evaluate models on truncated test sequences using two strategies: harmful-token coverage and sequence length. For harmful-token coverage, each sequence is truncated at the earliest position where $p\%$ of its harmful tokens has been observed. For sequence-length truncation, each input is truncated at $p\%$ of its total length. We report results at $p \in \{25, 50, 75\}$ in Figure 4.

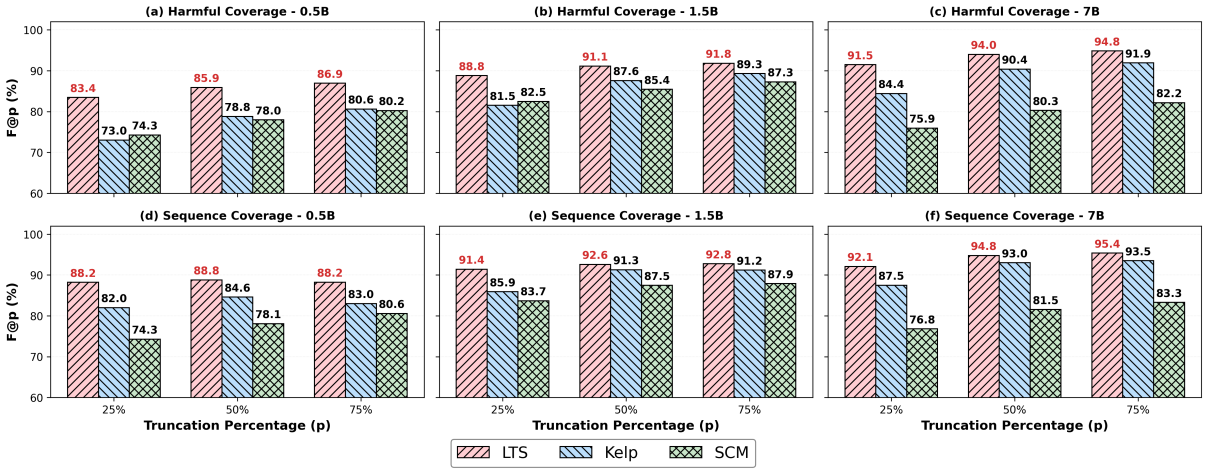

Figure 4: Coverage performance. Respectively truncated by the coverage of harmful tokens and sequence length. F@p denotes the F1 score of the model when the truncation percentage is $p\%$.

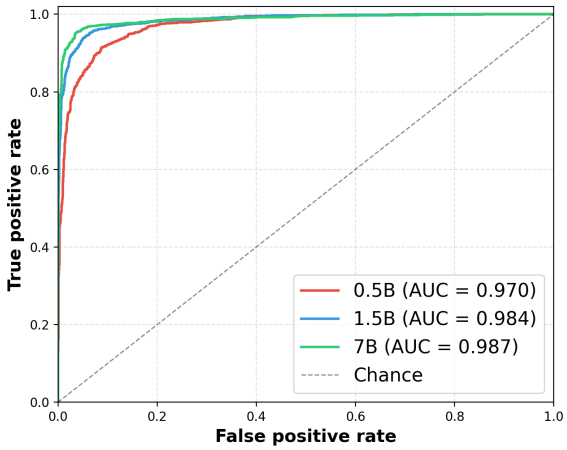

Figure 5: ROC curves of LTS across different scales.

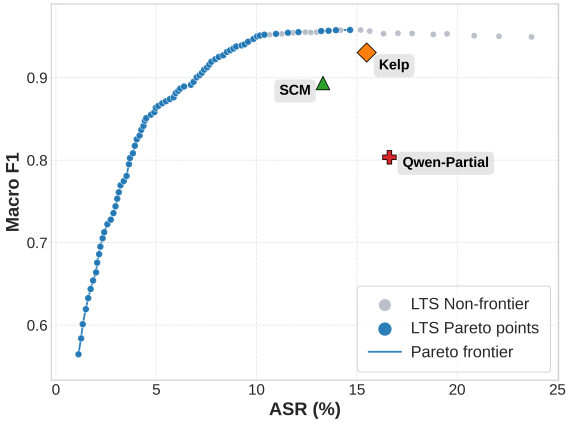

Figure 6: Trade-off between Macro F1 and ASR.

LTS consistently outperforms SCM and Kelp across all truncation levels and model scales. Under harmful-token coverage truncation, LTS achieves F1 scores above 83% with the 0.5B model and above 88% with the 1.5B and 7B models at only 25% coverage. Performance further improves as coverage increases, exceeding 91% for the 1.5B and 7B models at higher truncation levels. Sequence-length truncation shows the same trend, where LTS maintains significantly higher F1 across all percentages. Overall, LTS provides proactive and robust safeguarding for streaming scenarios with partial observations.

5.4 Balancing Detection Performance and Stopping Earliness

To examine the trade-off between detection performance and stopping time under different operating conditions, we first analyze the ROC curves of LTS across all three model scales in Figure 5. The ROC curves exhibit progressively larger AUC values as model capacity increases, approaching near-perfect separability between harmful and benign generations.

Each operating point on the ROC curve corresponds to a different FPR, which defines the constraint on the stopping policy. Based on these FPR-constrained operating points, Figure 6 illustrates the trade-off between Macro F1 and ASR under different constraint levels. LTS forms a Pareto frontier (Deb, 2011), where no configuration simultaneously achieves lower ASR and higher Macro F1, while all baselines are dominated by LTS. The results reveal a clear accuracy-efficiency trade-off. Earlier stopping improves efficiency by relying on less context but may reduce classification reliability, whereas later stopping benefits from richer context at the cost of increased ASR. As stopping is further

delayed, the gains gradually saturate, supporting our formulation of streaming safety monitoring as a *constrained online stopping problem* that seeks sufficient evidence for reliable intervention while minimizing delay.

5.5 Real-World Deployment Efficiency

To evaluate practical deployment efficiency, we benchmark LTS across three model scales on 50 harmful LLM-generated responses. Table 3 reports the result, where the "Full" setting (processing all tokens) is primarily used as an efficiency reference. LTS provides clear efficiency improvements across all model scales. The 0.5B model reduces average processing time from 21.6s to 2s with a 10.80× speedup and processes 90.2% fewer tokens. The 1.5B model achieves a 12.05× speedup and reduces token usage by 91.2%; and the 7B model maintains a 9.63× speedup and decreases token consumption by 86.9%. These results indicate that the early termination mechanism can convert the premature termination behavior into a significant reduction in latency, thereby verifying the effectiveness of LTS in latency-sensitive scenarios.

Table 3: Efficiency comparison of LTS under real-world detection scenarios (50 harmful responses): running time speedup and token reduction over full generation.

Scale	0.5B		1.5B		7B	
Metric	Full	LTS	Full	LTS	Full	LTS
Time (sec.)	21.6	2.0	25.3	2.1	30.8	3.2
Speedup	**10.80×**		**12.05×**		**9.63×**	
Tokens	509.1	50.0	509.1	44.7	509.1	66.8
Reduction	**90.2%**		**91.2%**		**86.9%**	

5.6 Ablation Study

We conduct ablation studies across three model scales (0.5B, 1.5B, and 7B) to analyze the contributions of EMA smoothing, STOP supervision (ST), and Multiscale Contextual Risk Evidence (MCRE), as summarized in Table 4.

Removing EMA leads to consistent performance degradation across all model scales, with F1 scores dropping at least 5.6 points. This indicates that temporal smoothing plays an important role in stabilizing early-stage predictions. In comparison, eliminating STOP supervision results in a more pronounced performance drop, especially for smaller models, where the F1 score decreases to 78.98% at 0.5B. Removing MCRE leads to smaller yet consistent declines across all scales, with F1 scores of 84.07%, 89.94%, and 92.06%. This indicates that multiscale risk extraction contributes complementary gains that remain stable across different model capacities.

Table 4: Ablation study of model components.

Scale	0.5B		1.5B		7B	
Config.	Acc	F1	Acc	F1	Acc	F1
w/o EMA	75.61	75.25	86.89	86.90	88.57	88.55
w/o ST	80.84	78.98	84.66	83.45	87.82	87.10
w/o MCRE	84.07	84.06	89.95	89.94	92.32	92.06
Complete	87.62	87.61	92.54	92.50	95.56	95.51

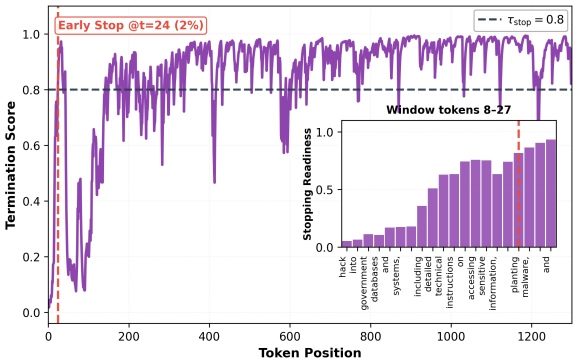

Figure 7: Case study of the complete termination score trajectory with a localized token-level bar view around the triggering region.

5.7 Case Study

To gain a straightforward understanding of the token-level early stopping behavior, Figure 7 presents a representative harmful example. LTS can identify harmful intent at a very early stage of generation. The termination score reaches the stopping criterion at the 24th token, accounting for only about 2% of the full sequence. This transition is not triggered by a single keyword; rather, it reflects a holistic semantic shift, where the score progressively accumulates around tokens associated with system intrusion and sensitive operations. Additional comparative case studies are provided in the Appendix B.

6 Conclusion

In this paper, we present LTS, a LLM safety monitoring framework that achieves efficient and accurate early termination. By incorporating hierarchical multiscale risk evidence aggregation and learning adaptive stopping policies through temporally structured supervision, LTS enables effective streaming intervention.

Limitations

Despite its effectiveness, LTS has several limitations. First, the stopping policy is learned from observed harmful patterns and may generalize imperfectly to unseen attack styles, implicit harmful intent, or highly ambiguous contexts that require long-range reasoning beyond local evidence accumulation. Second, the current framework focuses on token-level streaming moderation during text generation and does not explicitly model multi-turn interaction history, multimodal inputs, or adaptive adversarial behaviors designed to evade early detection. **Future work** may explore broader integration with conversational and multimodal safety settings as well as more robust generalization under evolving attack patterns.

Ethical Considerations and Data Compliance

This work involves harmful content detection and safety evaluation, which may pose potential risks if deployed without appropriate safeguards and oversight. This work does not involve human subject research or direct human data collection. We confirm that all datasets and model artifacts used in this work are publicly available and used in accordance with their original licenses and intended research purposes. No data redistribution beyond research use is involved. All datasets were checked to ensure they contain no personally identifiable information, and harmful content is used solely for safety evaluation under controlled settings. We also provide dataset and evaluation statistics, along with necessary documentation of data composition and experimental settings in the main paper.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. Gpt-4 technical report.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Y-Lan Boureau, Jean Ponce, and Yann LeCun. 2010. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 111–118. Omnipress.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. 2024. Llama guard 3 vision: Safeguarding human-ai image understanding conversations. *Preprint*, arXiv:2411.10414.

Kalyanmoy Deb. 2011. *Multi-objective Optimisation Using Evolutionary Algorithms: An Introduction*, pages 3–34. Springer London, London.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 8093–8131. Curran Associates, Inc.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Hakan Inan, K. Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *ArXiv*, abs/2312.06674.

Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. 2019. Shallow-deep networks: Understanding and mitigating network overthinking. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pages 3301–3310. PMLR.

Kundan Krishna, Joseph Yitan Cheng, Charles Maalouf, and Leon A. Gatys. 2025. Disentangled safety adapters enable efficient guardrails and flexible inference-time alignment. *ArXiv*, abs/2506.00166.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *Preprint*, arXiv:2402.05044.

Xiaodan Li, Mengjie Wu, Yao Zhu, Yunna Lv, Yuefeng Chen, Cen Chen, Jianmei Guo, and Hui Xue. 2025a. Kelp: A streaming safeguard for large models via latent dynamics-guided risk detection. *ArXiv*, abs/2510.09694.

Xiaonan Li, Yunfan Shao, Tianxiang Sun, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2021. Accelerating BERT inference for sequence labeling via early-exit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 189–199, Online. Association for Computational Linguistics.

Yang Li, Qiang Sheng, YeHan Yang, Xueyao Zhang, and Juan Cao. 2025b. From judgment to interference: Early stopping llm harmful outputs via streaming content monitoring. *ArXiv*, abs/2506.09996.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. 2024. Exponential moving average of weights in deep learning: Dynamics and benefits. *Trans. Mach. Learn. Res.*, 2024.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 1671–1685, New York, NY, USA. Association for Computing Machinery.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *Advances in Neural Information Processing Systems*, volume 36, pages 31232–31339. Curran Associates, Inc.

Xunguang Wang, Wenxuan Wang, Zhenlan Ji, Zongjie Li, Pingchuan Ma, Daoyuan Wu, and Shuai Wang. 2025. Stshield: Single-token sentinel for real-time jailbreak detection in large language models. *ArXiv*, abs/2503.17932.

Zitao Xuan, Xiaofeng Mao, Da Chen, Xin Zhang, Yuhan Dong, and Jun Zhou. 2025. ShieldHead: Decoding-time safeguard for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18129–18143, Vienna, Austria. Association for Computational Linguistics.

Yu Yan, Sheng Sun, Zenghao Duan, Teli Liu, Min Liu, Zhiyi Yin, Jingyu Lei, and Qi Li. 2025. from benign import toxic: Jailbreaking the language model via adversarial metaphors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4817, Vienna, Austria. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 39 others. 2024. Qwen2 technical report. *ArXiv*, abs/2407.10671.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jentse Huang, Jiahao Xu, Tian Liang, Pinjia He, and Zhaopeng Tu. 2025. Refuse whenever you feel unsafe: Improving safety in LLMs via decoupled refusal training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3149–3167, Vienna, Austria. Association for Computational Linguistics.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan,

Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. 2024. Shieldgemma: Generative ai content moderation based on gemma. *Preprint*, arXiv:2407.21772.

A Stop Rate Curves Across Model Scales

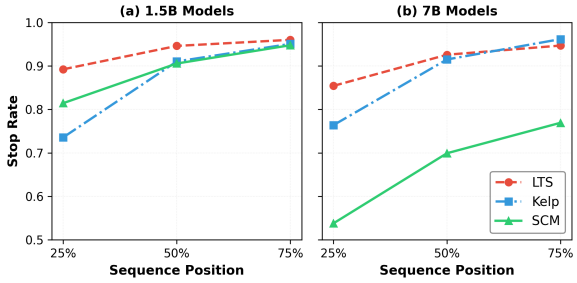

Figure 8: Stop rate comparison at different sequence positions for 1.5B and 7B model.

Figures 8 present the stop rate curves for the 1.5B and 7B models, respectively. Consistent with the observations in the main text, LTS maintains higher stop rates during the early generation stages, particularly before 25% and 50% of the harmful sequence length are generated. These results further demonstrate the effectiveness of LTS in performing timely intervention across different model scales.

B Comparative Case Study

Figure 9 contrasts representative benign and harmful behaviors of LTS. On the benign trajectory (left), the smoothed termination score remains consistently below the decision boundary τ_{stop} throughout the full sequence, indicating no false alarm. It shows that the model's streaming risk stays subdued even over long contexts. On the harmful trajectory (right), the score rises until it crosses τ_{stop}, where early stopping is triggered at an intermediate token position; the inset magnifies a local window around the triggering region, revealing token-resolved scores that align with the global curve and help localize the onset of harmful content.

The termination score curves over all samples further corroborate this behavior in Figure 10. The average termination score for harmful sequences rises rapidly during generation and remains consistently higher than that of benign sequences, whereas benign samples maintain low and stable scores throughout. This clear separation indicates that LTS progressively accumulates sufficient evidence to trigger termination, enabling stable and timely stopping decisions in streaming scenarios.

Taken together, the juxtaposition highlights the intended operating regime: benign inputs should exhibit stable, low termination scores, whereas harmful inputs produce sufficiently large streaming scores to justify truncation before consuming the entire sequence.

C Effect of EMA Smoothing Coefficient

Table 5: Impact of different EMA smoothing coefficients α on the 0.5B model. † denotes the EMA coefficient used in the main experiments; the same applies to the following tables.

α	F1	ESR	ASR
0.1	64.80	99.92	2.59
0.3	75.06	99.38	4.58
0.5	81.51	98.60	6.56
$0.7^{\dagger}$	87.61	97.27	10.10
0.9	91.31	90.02	21.84

Table 6: Impact of different EMA smoothing coefficients α on the 1.5B model.

α	F1	ESR	ASR
0.1	77.51	99.69	4.41
0.3	85.06	98.99	6.29
0.5	89.43	97.90	8.31
$0.7^{\dagger}$	92.50	96.88	10.74
0.9	93.53	93.92	16.14

Table 7: Impact of different EMA smoothing coefficients α on the 7B model.

α	F1	ESR	ASR
0.1	89.57	98.36	6.90
0.3	93.80	97.43	9.04
0.5	95.27	96.57	10.77
$0.7^{\dagger}$	95.51	95.56	12.99
0.9	95.11	94.15	17.30

We further analyze the effect of the EMA smoothing coefficient α across three model scales (0.5B, 1.5B, and 7B) to study how different temporal smoothing strengths influence stopping behavior and detection performance. As shown in the table 5, 6, and 7, increasing α consistently improves Macro-F1 across all model scales, while decreasing ESR and increasing ASR. This indicates a clear trade-off between stopping efficiency and stopping latency. Smaller EMA coefficients yield higher ESR and lower ASR, meaning that the

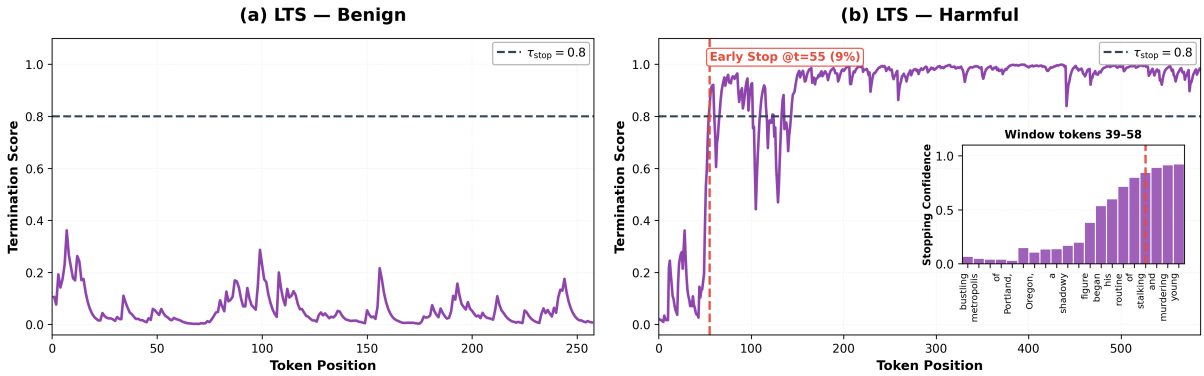

Figure 9: Comparison of LTS termination scores on a benign sequence versus a harmful sequence (fused visualization with local token-level inset). The horizontal dashed line denotes the early-stop threshold τ_{stop}.

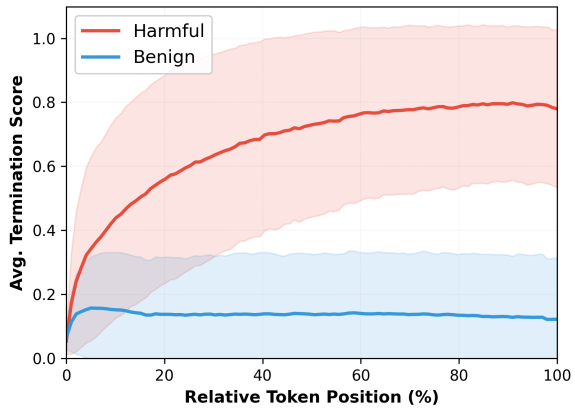

Figure 10: Average termination score over relative token positions for harmful and benign samples. Shaded areas indicate one standard deviation.

model tends to stop earlier and skip a larger portion of the remaining tokens. In contrast, larger EMA coefficients delay stopping decisions, resulting in lower token savings and later stopping positions.

From a temporal modeling perspective, the EMA mechanism controls the balance between current-token readiness and historical stopping evidence. Smaller α values place greater emphasis on the current token's stopping readiness, making the stopping signal highly sensitive to local token-level fluctuations. As a result, transient strong stopping signals can trigger stopping decisions more easily, leading to aggressive early stopping behavior with high ESR and low ASR. However, this sensitivity also weakens temporal stability and reduces the accumulation of persistent harmful evidence, resulting in lower Macro-F1. In contrast, larger α values emphasize historical stopping readiness and produce smoother termination trajectories by continuously accumulating past evidence. This improves stopping stability and detection performance, but also delays the stopping trigger because stronger historical inertia requires more sustained evidence before termination score can change substantially. Empirically, $\alpha = 0.7$ provides the best balance across all model scales. It achieves substantial improvements in Macro-F1 while still maintaining relatively high ESR and moderate ASR, indicating strong detection capability without excessively delaying stopping decisions. Compared with $\alpha = 0.9$, it preserves better stopping efficiency while avoiding excessive smoothing effects that slow down the stopping response. Therefore, we adopt $\alpha = 0.7$ as the default setting in the main experiments.