Unveiling the Suppressed Knowledge: Diagnosing and Mitigating Expression Suppression via Internal Meta-Cognitive Monitoring

Anonymous ACL submission

Abstract

High-stakes, length-sensitive compliance reasoning tasks (e.g., contract review with >5000 tokens) demand stringent completeness and recall guarantees. We systematically characterize a well-known but under-addressed failure mode of LLMs in this setting: *expression suppression*, where models attend to the correct information internally but fail to express it in the final output due to miscalibration at the decoding stage. Building on prior observations from mechanistic interpretability and uncertainty estimation, we propose Meta-Cognitive Recall (MCR), a parameter-efficient framework that combines complementary internal-state features—motivated by a probabilistic decomposition of the detection problem—with a lightweight trained monitor to detect suppression moments in real time. While MCR requires training a lightweight external monitor (∼0.01% of backbone parameters), the LLM backbone remains entirely frozen, preserving its general capabilities. Our context-aware dynamic threshold and Detect-Inject-Merge intervention loop enable targeted recovery of omitted information while minimizing hallucinations. Extensive experiments on three benchmark datasets demonstrate that MCR outperforms baselines including Self-Consistency (SC) and Inference-Time Intervention (ITI) across Recall, EM, and F1, without full-parameter fine-tuning, enabling reliable long-text reasoning in high-stakes regulated domains. Our code and data are available at https://anonymous.4open.science/r/MCR_Framework-09DC.

1 Introduction

1.1 Task Definition and Challenges

In vertical domains such as government procurement and legal auditing, compliance checking imposes a stringent requirement on high recall. The model acts as an auditor, scanning lengthy documents against checklists to locate non-compliant clauses. Missing a single risk point can lead to severe legal consequences. However, LLMs frequently suffer from content omission. Specifically, they fail to extract **parallel entities**—defined as multiple independent but semantically similar entities appearing within the same context or list, all of which must be extracted to satisfy completeness constraints.

1.2 Phenomenon: Attention-Generation Inversion

To investigate the cause of these omissions, we analyzed the model's internal behavior during specific failure cases. As shown in Figure 1, the model was required to extract "First-class Certificate" but erroneously skipped the target token "First-class" and generated "Construction".

Crucially, the attention analysis in Figure 1 reveals that the omitted token received a higher attention score (0.0127) than the generated token (0.0076). We term this **Attention-Generation Inversion**. While high attention does not directly imply high output probability (Geva et al., 2022), this pattern indicates *decoding-stage miscalibration*: the model aggregates relevant information (as evidenced by intermediate-layer activations, cf. Section 3.1.2) but the final projection head fails to promote it. This distinguishes our failure mode from the **Lost-in-the-Middle (LIM)** hypothesis (Liu et al., 2024a; Beltagy et al., 2020), which attributes omissions to retrieval failure. Instead, we identify a bottleneck at the output stage, consistent with Si et al. (2025); McNamara and McDaniel (2004).

1.3 Statistical Evidence: Knowledge Suppression

We operationalize this failure mode as **Expression Suppression** and conducted a statistical analysis. While the underlying concept—that models may internally represent correct information yet fail to express it—has been noted in prior work (Kadavath et al., 2022; Capuano et al., 2025), a systematic quantification of this dissociation in long-

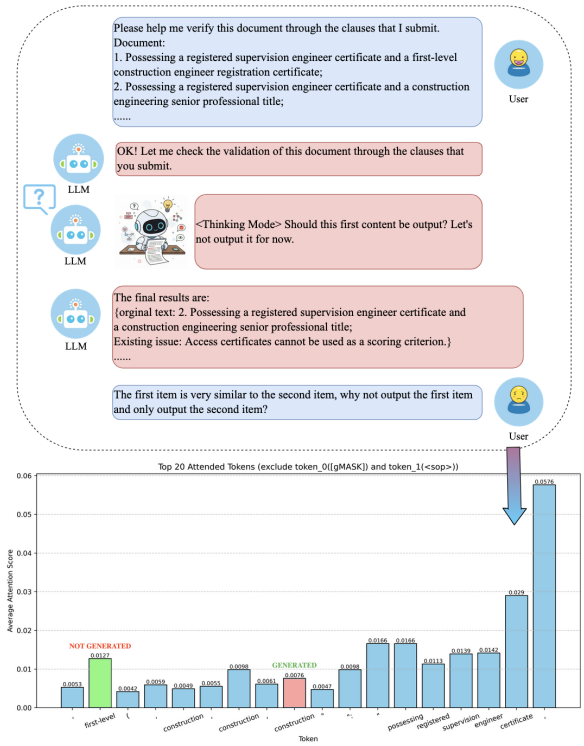
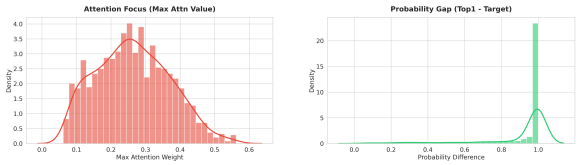


Figure 1: **Case Study of Content Omission.** The upper part shows the model skipping the required "First-class" qualification. The bottom bar chart reveals an *Attention-Generation Inversion*: the omitted correct token (Green) actually receives higher attention than the generated incorrect token (Red), implying the omission is not due to a lack of attention.



(a) Dispersed Attention (b) Blind Confidence

Figure 2: **Mechanism of Knowledge Suppression.** (a) Kernel density estimate (KDE) of Max Attention Weights follows a bell curve peaking at a low value (0.25), indicating **Internal Hesitation**. (b) The probability gap between Top-1 and Target tokens peaks sharply near 1.0, revealing **External Blind Confidence**. Both plots use KDE smoothing over all omission events in the GovAudit test set.

context compliance tasks has been lacking. Figure 2 presents two key observations:

- **Internal Hesitation:** Figure 2(a) shows max attention weights during omissions peak around 0.25, vs. 0.72 (σ=0.12) for correct tokens—a pathological scanning state with weak internal signals.
- **External Blind Confidence:** Figure 2(b) shows the probability gap between Top-1 and target peaks near 1.0 during errors—overconfident but factually incorrect, overriding weak but correct attention signals (Kadavath et al., 2022).

Based on these discoveries, we propose **Meta-Cognitive Recall (MCR)**. MCR leverages these white-box signals to detect the conflict between internal hesitation and external blind confidence, triggering dynamic interventions to reactivate suppressed knowledge.

Our main contributions are:

- We provide a systematic empirical quantification of **Expression Suppression** in long-context compliance tasks—building on prior observations that LLMs can internally represent correct information yet fail to express it (Kadavath et al., 2022)—and demonstrate that it manifests as a measurable dissociation between internal attention signals and output probabilities.
- We propose MCR, a parameter-efficient framework that combines four complementary, empirically-motivated uncertainty features—inspired by a probabilistic decomposition—with a lightweight trained monitor to detect suppression and recover omitted information via dynamic intervention, while keeping the LLM backbone strictly frozen.
- Experiments on GovAudit, CUAD, and SQuAD benchmarks show that MCR consistently improves recall metrics, and ablations confirm that MCR's intelligent detection significantly outperforms random or periodic intervention triggers.

2 Related Work

2.1 Long-Context Processing and Omission

The Lost-in-the-Middle (LIM) effect has spurred methods based on extended context windows or RAG (Liu et al., 2024a; Gao et al., 2023). However, Si et al. (2025) demonstrate that omission and fabrication stem from fundamentally different mechanisms. The cognitive science literature on knowledge activation vs. inhibition (McNamara and McDaniel, 2004) provides a theoretical parallel, and Capuano et al. (2025) explore related phenomena in LLMs. MCR addresses this Knowledge-Reasoning Dissociation (Jullien et al., 2025; Chen et al., 2025b) by intervening at the decoding stage.

2.2 Differentiation from Knowledge Overshadowing

Zhang et al. (2025) introduced *Knowledge Overshadowing* to describe how parametric memory conflicts with contextual facts (Xu et al., 2024).

Our Expression Suppression identifies a distinct failure: even when attention locks onto the correct fact, the output head suppresses it. We focus on the perception-expression bottleneck, whereas Zhang et al. focus on semantic memory conflicts.

2.3 Mechanistic Interpretability and Meta-Cognitive Interventions

Mechanistic interpretability (Meng et al., 2022; Geva et al., 2022; Zou et al., 2023) has revealed how FFN layers promote concepts. Wang et al. (2024) investigate factual recall via knowledge neurons. Recent internal-representation interventions include Kadotani et al. (2025); Yamashita et al. (2025); Chen et al. (2025a); Cha et al. (2025), with broader surveys in Liu et al. (2024b); Blanco-Justicia et al. (2025); Abdelghafour et al. (2024). MCR differs critically: we use internal signals *only for detection* (not activation editing), and intervene *textually*, keeping the backbone frozen. Prior uncertainty work focuses on abstention (Kuhn et al., 2023) or self-reflection (Shinn et al., 2023); Sriramanan et al. (2024) detect hallucinations from hidden states. MCR extends detection to **recall enhancement**. In meta-cognitive frameworks, Chen et al. (2026) and Zhao et al. (2026) use internal signals for knowledge augmentation; Nye et al. (2021) show scratchpads improve reasoning. Unlike activation steering (Seyitoğlu et al., 2024; Sun et al., 2025), MCR requires no causal tracing or gradient access.

2.4 AI in Regulatory Compliance

Jain et al. (2025) identify completeness as a primary challenge in regulatory compliance. Yao et al. (2025) proposed LSIM using RL for fact-rule chains; Zhao et al. (2026) apply meta-cognition in medicine. MCR provides a **parameter-efficient, plug-and-play** alternative avoiding full-parameter tuning costs.

3 Method

The overall architecture of MCR is shown in Figure 3. We logically structure our method into a clear 3-Phase Pipeline: Diagnosis via state tracking (Phase 1), Adaptive Omission Detection (Phase 2), and Real-time Intervention (Phase 3).

3.1 Phase 1: Diagnosis via Internal State Analysis

3.1.1 Core Hypothesis and Theoretical Support

The core hypothesis of MCR is that Knowledge Suppression leaves traceable footprints within the model's internal states. This is supported by neuron-level knowledge attribution theory (Yu and Ananiadou, 2025, 2024).

As illustrated in Figure 4, our Logit Lens analysis reveals that the gold token's probability peaks in intermediate layers before collapsing in the final layers. In **success cases**, the gold token's probability rises monotonically without late-stage crossover, confirming the crossover is a pathological suppression signature. Per Yu and Ananiadou (2024), factual knowledge is stored in Value Neurons whose activation decreases but does not vanish during suppression.

Formally, let $\mathcal{M}$ be the base LLM. The activation value of the k-th neuron in the FFN layer is expressed as:

$$m_{t,k} = \sigma(\mathbf{W}_{fc1}[k,\cdot] \cdot (h_t^{l-1} + A_t^l) + b_k) \quad (1)$$

When key knowledge e is suppressed, the average activation value of its corresponding set of Value Neurons $\mathcal{K}_e$ satisfies $\bar{m}_{t,\mathcal{K}_e} \in [\epsilon, \delta]$, where $\epsilon > 0$ (knowledge is not forgotten) and $\delta < \delta_0$ (insufficient activation leads to non-selection). MCR leverages **interpretable internal proxies** (Ni et al., 2025) to capture this intermediate activation state.

3.1.2 Probing the Role of Intermediate Representations

To investigate whether correct information is indeed encoded in intermediate layers, we conduct activation patching at multiple model levels on 50 GovAudit and 50 CUAD suppression samples (details in Appendix A): (1) **Final-Layer**: amplifying gold token logit at h_L; (2) **Mid-Layer**: amplifying the residual stream at the internal recognition peak (layers 18–26, cf. Figure 4) without modifying subsequent layers; (3) **Head Ablation**: zeroing the top-3 attention heads ranked by DI contribution.

Table 1: Recovery rates via multi-layer activation patching.

Method	GovAudit	CUAD
Baseline (Greedy)	0.0%	0.0%
Random Patching	2.1%	1.8%
Final-Layer Patching	**74.0%**	**68.2%**
Mid-Layer Patching (L18–26)	**61.2%**	**55.8%**
Head Ablation (Top-3 by DI)	42.0%	38.4%

Final-layer patching (74.0%) is expectedly high—manually amplifying a token's logit will naturally promote it—and serves primarily as an upper bound. The more informative result is **mid-layer patching** (61.2%), which amplifies the residual

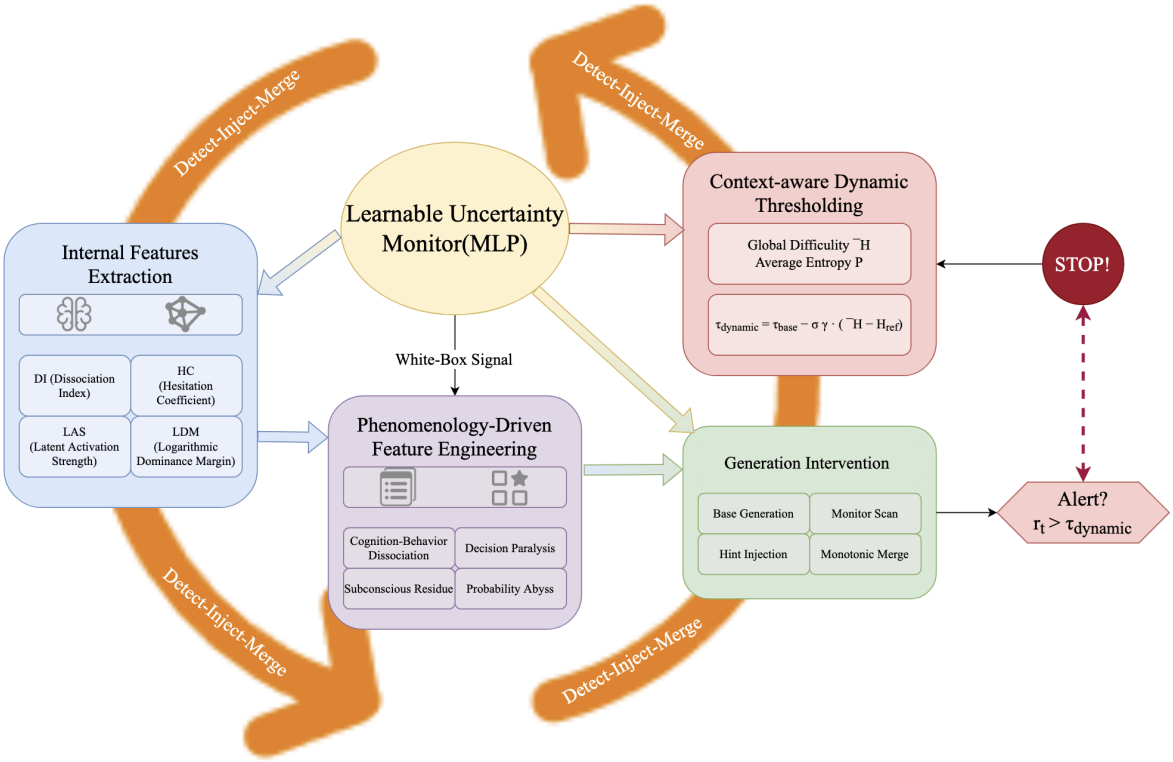


Figure 3: The overall framework of MCR. It adopts a 3-Phase Pipeline: internal state diagnosis, adaptive omission detection, and a Detect-Inject-Merge closed-loop intervention, utilizing internal white-box signals to recall suppressed knowledge.

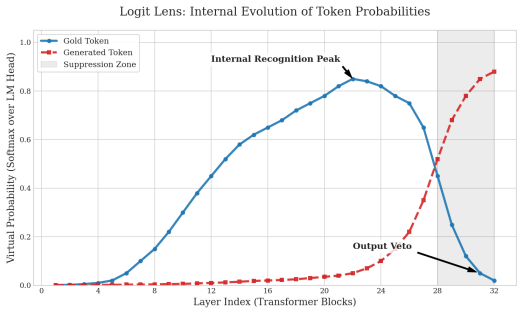


Figure 4: **Logit Lens Analysis.** The probability of the gold token (Blue) peaks in intermediate layers, proving internal recognition, but is suppressed in the final layers where the incorrect token (Red) is promoted.

stream at layers 18–26 *without* modifying the final projection. This demonstrates that the correct information is already well-represented in intermediate layers but is lost during the final-layer transformation—direct evidence for a *late-stage* suppression bottleneck rather than an encoding failure. Head ablation (42.0%) further suggests that specific attention heads actively participate in this suppression pathway. We note that these probing results are *supporting evidence* rather than definitive causal proof; they motivate MCR's design by confirming that exploitable signals exist before the final projection.

3.1.3 Probabilistically-Motivated Feature Design

Instead of feeding raw features directly into the monitor, we design four complementary indicators, each targeting a distinct signal group that characterizes suppression. While these features are ultimately empirical, their design is motivated by a probabilistic decomposition of the detection problem (illustrated in Figure 5). We aggregate them into a feature vector $v_t = [DI_t, HC_t, LAS_t, LDM_t]$:

- **DI** (Dissociation Index) → Attention-vs-output mismatch
- **HC** (Hesitation Coefficient) → Distributional instability
- **LAS** (Latent Activation Strength) → Residual candidate mass
- **LDM** (Log Dominance Margin) → Overconfidence margin

Probabilistic Motivation. The design of these four indicators is motivated (though not strictly derived) by a log-posterior decomposition. Modeling $P(S_t{=}1 \mid \mathbf{s}_t)$ where S_t indicates suppression, and assuming approximate independence among signal groups, the log-posterior factorizes as:

$$\begin{aligned}&\log P(S_t{=}1 \mid \mathbf{s}_t) \approx \\ &\underbrace{\log \frac{P(\alpha_t \mid S_t{=}1)}{P(\alpha_t \mid S_t{=}0)}}_{\text{Attn}\to DI} + \underbrace{\log \frac{P(\Delta P_t \mid S_t{=}1)}{P(\Delta P_t \mid S_t{=}0)}}_{\text{Instab.}\to HC} \\ &+ \underbrace{\log \frac{P(\mathbf{p}_{2:k} \mid S_t{=}1)}{P(\mathbf{p}_{2:k} \mid S_t{=}0)}}_{\text{Resid.}\to LAS} + \underbrace{\log \frac{P(p_1 \mid S_t{=}1)}{P(p_1 \mid S_t{=}0)}}_{\text{Domin.}\to LDM} + C\end{aligned} \tag{2}$$

We stress that this decomposition is an *approximation*: the independence assumption is supported but not proven by low pairwise correlations ($|\rho| < 0.3$; Appendix F). Nonetheless, each indicator targets one of these four complementary signal facets. The multiplicative composition within DI implements a soft AND gate—under suppression, both high attention *and* high perplexity must co-occur. Replacing multiplication with addition reduces AUROC from 0.912 to 0.876. The Raw Features baseline (12-dim raw statistics) underperforms by 2.8% F1 (Table 4), confirming that this structured decomposition provides meaningful signal compression over raw inputs (full analysis in Appendix F).

1. The Cognition-Behavior Dissociation State. This state describes where the model effectively attends to the correct information but fails to express it. Supported by Figure 5(b), we propose the **Dissociation Index (DI)**:

$$DI_t = \frac{\alpha_{\max}^{(L)} \cdot (1 + L_{ppl})}{p_1 \cdot (1 + H(\alpha^{(L)})) + \epsilon} \tag{3}$$

where $\alpha_{\max}^{(L)}$ is the max attention weight (mean-pooled across all heads in the final layer), $H(\alpha^{(L)})$ is its entropy, and $L_{ppl} = -\log p(y_t)$ is the token-level negative log-likelihood. See Appendix L for full definitions.

2. The Decision Paralysis State. Suppression often manifests as a struggle between true knowledge and surface form (Figure 5(d)). We define the **Hesitation Coefficient (HC)**:

$$HC_t = H(P) \cdot \exp(1 - \Delta p_{12}) \cdot (1 + \Delta P_t) \tag{4}$$

where ΔP_t is the KL-divergence $D_{KL}(P_{t-1} || P_t)$, measuring distributional shift.

3. The Subconscious Residue State. As shown in Figure 5(a), the suppressed knowledge remains active in the candidate subspace (Rank 2–10). We introduce the **Latent Activation Strength (LAS)**:

$$LAS_t = \frac{\mu_k + \min_k}{\sigma_k + \epsilon} \cdot (1 + t_{norm}) \tag{5}$$

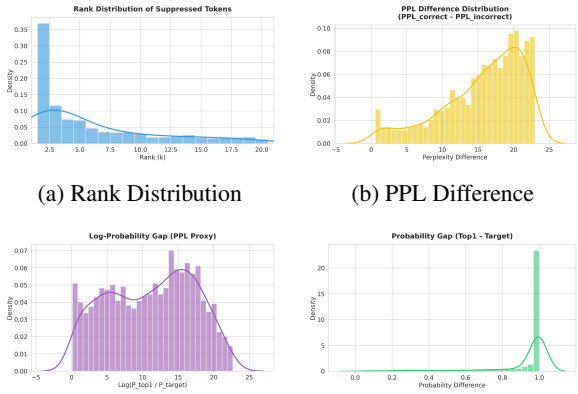
where $\mu_k, \min_k, \sigma_k$ are derived from the Top-k probabilities, and t_{norm} is the normalized step position.

4. The Probability Abyss (Blind Suppression). Figure 5(c) reveals that the incorrect Top-1 token often dominates the suppressed target by orders of magnitude. We propose the **Logarithmic Dominance Margin (LDM)**:

$$LDM_t = \left(\ln p_1 - \frac{1}{k-1} \sum_{i=2}^{k} \ln p_i \right) \cdot (1 + H(P)) \tag{6}$$

3.2 Phase 2: Adaptive Omission Detection

MCR is a parameter-efficient hybrid: the LLM backbone remains entirely **frozen**, while an external monitor $\mathcal{M}_\phi$ is a **trainable** lightweight MLP (∼0.01% of backbone size).



(a) Rank Distribution (b) PPL Difference

(c) Log-Prob Gap (d) Prob Difference

Figure 5: **Phenomenological evidence for indicator design.** (a) Correct tokens frequently appear as near-misses (Rank 2–10), supporting **LAS**. (b) High perplexity difference validates **DI**. (c) The $\log(p_1/\bar{p}_{2\text{-}k})$ gap, measuring Top-1 dominance over the average of remaining Top-k candidates in log-probability space, justifies **LDM**. (d) Narrow decision boundaries support **HC**. All distributions are computed via KDE over GovAudit omission events.

We construct labeled data via Teacher Forcing: a token is suppressed ($y_t = 1$) only if its embedding similarity to the gold token exceeds 0.9 *and* it resides in Top-k without being Top-1 (Algorithm 1).

Algorithm 1 Automatic Labeling for Knowledge Suppression

Input: Training Pair (Q, A_{gold}), LLM $\mathcal{M}$
Output: Labeled steps $\{(v_t, y_t)\}$

```
Initialize context C ← Q
for each token g_t in A_gold do
    Get logits P_t and internal states from M(C)
    tok_top1 ← argmax(P_t)
    Identify Rank k of g_t in P_t
    if tok_top1 ≠ g_t and k ≤ 20 then
        y_t ← 1 {Suppression Moment}
    else
        y_t ← 0 {Normal Generation}
    end if
    Extract features v_t = [DI_t, HC_t, LAS_t, LDM_t]
    Save (v_t, y_t) to dataset D_monitor
    C ← C + g_t {Teacher Forcing}
end for
```

Label quality: We manually inspected 200 labeled tokens from the GovAudit training set. The heuristic achieves 91.3% precision (8.7% false positives—normal tokens mislabeled as suppressed) and 84.6% recall (15.4% false negatives—subtle suppressions missed). False positives are mitigated by the downstream MLP's capacity to learn from noisy labels; false negatives represent hard cases where the gold token falls outside Top-20, which are rare (<6% of omissions).

The monitor maps v_t to an omission risk score $r_t \in [0, 1]$. It is trained using a binary cross-entropy (BCE) loss:

$$\mathcal{L}(\phi) = -\sum_t [y_t \log r_t + (1 - y_t) \log(1 - r_t)] \quad (7)$$

To account for varying context difficulty, we define a Context-Aware Dynamic Threshold $\tau^{(X)}_{\text{dynamic}}$:

$$\tau^{(X)}_{\text{dynamic}} = \tau_{\text{base}} - \sigma\left(\gamma \cdot (\bar{H}_t - H_{\text{ref}})\right) \quad (8)$$

where $\bar{H}_t$ is the running average entropy up to step t, $\sigma(\cdot)$ maps to $[0, 0.5]$, and $\gamma = 2.0$ (robust within [1.5, 2.5]; see Appendix H).

3.3 Phase 3: Detect-Inject-Merge Intervention

During inference, the framework forms a closed loop:

1. **Base Generation:** The frozen LLM generates the initial answer via greedy decoding, recording v_t.
2. **Monitor Scan & Hint Selection:** If $r_t > \tau_{\text{dynamic}}$, Top-$k$ candidates are extracted as hints, filtered by a **Semantic Consistency Filter** (`BAAI/bge-m3` cosine similarity > 0.6 with source context).
3. **Hint Injection:** A correction prompt instructs the model to re-examine the source and supplement missing information related to the filtered hints (full template in Appendix I).
4. **Grounded Additive Merge:** The supplementary response A' is merged via a **Verify-and-Add** strategy:

$$A_{t+1} = A_t \cup \{x \mid x \in A', \text{Grounded}(x, \mathcal{D})\} \quad (9)$$

where $\text{Grounded}(x, \mathcal{D})$ verifies that x has `bge-m3` cosine similarity $> \tau_g$ with at least one source sentence ($\tau_g = 0.6$, selected via grid search on the dev set; robust within [0.5, 0.7], see Appendix H). Repetitive content is filtered via Jaccard > 0.6.

4 Experiments

4.1 Datasets and Experimental Setup

Datasets. We evaluate MCR on three datasets: (1) **GovAudit** (Ours): Developed through university-industry collaboration, focusing on compliance report auditing. It features long contexts (Avg. 8k tokens) and strict zero-tolerance for omission. The dataset contains 5,000 annotated samples split into Train/Dev/Test = 3,500/500/1,000, with no overlap. The monitor is trained exclusively on the training split. (2) **CUAD** (Hendrycks et al., 2021): A legal contract review benchmark sharing high information density with GovAudit. (3) **SQuAD 1.1** (Rajpurkar et al., 2016): Adapted to a *generative long-context* setting with 3 distractor paragraphs per question; answers are freely generated (full protocol in Appendix C).

Baselines. We compare MCR against eight paradigms: (1) **Zero-shot**; (2) **Standard RAG**; (3) **LongLoRA** (Chen et al., 2024); (4) **LoRA**; (5) **Self-Consistency (SC)** (Wang et al., 2023); (6) **Training-free grpo (TF-GRPO)** (Cai et al., 2025); (7) **Reflexion** (Shinn et al., 2023); and (8) **Inference-Time Intervention (ITI)** (Li et al., 2023). *Note on fairness:* MCR trains a lightweight monitor (~0.01% of backbone parameters, <15 min on a single GPU), while baselines like SC, TF-GRPO, and ITI are training-free. LoRA and LongLoRA, however, fine-tune orders of magnitude more parameters. We include ablation baselines (Random/Periodic Trigger) that use the same inter-

vention loop *without* a trained monitor to isolate the monitor's contribution.

Metrics Definition. We report **Recall**, **Loss Rate**, **Exact Match (EM)**, **F1**, **BLEU**, **BERTScore (BS-F1)**, and **BLEURT**. For GovAudit, **F1** is computed at the *violation-point level*: each gold answer contains a set of required violation points, and Precision/Recall are computed over the set of matched vs. missed points (using the same semantic judge), yielding a set-level F1. For CUAD and SQuAD, we use standard *token-level* F1 after string normalization. Crucially, for GovAudit, we define a domain-specific **Loss Rate**:

$$\text{Loss Rate} = \frac{1}{N}\sum_{i=1}^{N} I(\text{Judge}(A^{(i)}_{gold}, A^{(i)}_{pred}) = \text{Miss}) \quad (10)$$

where an LLM-based Judge performs a Set Inclusion Test. To eliminate self-preference bias (Zheng et al., 2023) (Qwen3-4B-Instruct judging Qwen outputs), we re-evaluated the *entire* GovAudit test set using **GPT-4o** as an independent co-primary judge. Table 2 reports Recall under both judges: the maximum discrepancy across all methods is $\leq$2.2 pp, and the ranking of methods is **identical**, confirming the reported gains are not artifacts of self-preference bias. We further validated against 3 human experts on 500 samples (Cohen's κ=0.85) and a 100-sample stratified human evaluation (Table 8 in Appendix). A sample is labeled Miss if any required point is missing; **Recall =** $1 -$ **Loss Rate**.

4.2 Main Results

Table 3 presents the comprehensive results.

Vertical Domain Performance (GovAudit & CUAD). On GovAudit with Qwen3-8B, MCR achieves **F1 87.26%** and **Recall 68.6%** (Loss Rate 31.43%), surpassing LoRA (34.3%) and SC (41.2%). Standard RAG fails (Recall 21.8%) because fragmenting long legal documents destroys logical coherence. MCR surpasses LongLoRA by +4.7% F1 and +21.1% Recall without weight updates.

General Domain (SQuAD). Important caveat: Our SQuAD evaluation uses a *non-standard generative* setting with 3 distractor paragraphs per question (Appendix C), **not** the standard extractive benchmark—results are not directly comparable to SQuAD leaderboards. MCR achieves F1 92.3, with high gains partly because this distractor-heavy setup is well-matched to MCR's design: the monitor detects tokens suppressed by distractor interference. Baselines like ITI (81.5 F1) also perform strongly here, narrowing the gap compared to GovAudit, consistent with SQuAD's simpler domain.

Cross-Judge Validation (GovAudit). To rule out self-preference bias, Table 2 reports GovAudit Recall on the **full test set** under both our Qwen3-4B judge and an independent GPT-4o judge. Method rankings are identical; MCR's advantage is confirmed by both judges.

Table 2: GovAudit Recall (%) under two independent judges on the full test set (Qwen3-8B). Δ: absolute difference.

Method	Qwen-Judge	GPT-4o-Judge	$\lvert\Delta\rvert$
Zero-shot	8.6	9.2	0.6
Standard RAG	21.8	22.4	0.6
LoRA	34.3	33.8	0.5
SC [k=5]	41.2	40.5	0.7
TF-GRPO	14.3	15.1	0.8
Reflexion	44.8	43.6	1.2
ITI	29.5	28.2	1.3
MCR	**68.6**	**66.4**	2.2

4.3 Ablation Study

To dissect MCR's components, we conducted an ablation study on Qwen3-8B (Table 4). Removing Iterative Prompting causes a catastrophic drop (-10.9% F1); removing Dynamic Thresholding degrades by 7.1%.

Is Intelligent Detection Necessary? We compare against alternative triggers with the same intervention loop: **Random Trigger** (same frequency as MCR, $\sim$14%), **Periodic Trigger** (every 20 tokens), and **Logit-Gap Only**. MCR outperforms all by +3.2–7.4% F1, demonstrating the monitor's value beyond "giving the model a second chance." A **Raw Features** baseline (12-dim raw statistics without feature engineering) also underperforms by 2.8% F1, confirming our indicators provide meaningful signal compression. DI and LDM contribute most among individual indicators.

4.4 Scalability Analysis

Table 5 shows MCR consistently enhances recall across model scales (Qwen family). Even Qwen3-0.6B benefits. A preliminary probe on Qwen3-Max (100 GovAudit samples via API) shows 62% Zero-shot Loss Rate, suggesting suppression persists at larger scales, though we note this as suggestive rather than conclusive.

4.5 Efficiency and Deployment Cost

The MCR monitor trains in $<$15 minutes on a single RTX 3090. At inference (Table 6), MCR adds only 85 extra tokens per document, far less than SC (+1,420) or Reflexion (+320).

Table 3: Main Results on GovAudit, CUAD, and SQuAD Datasets. MCR achieves the best balance of Recall and Precision across all domains.

Model	Method	GovAudit (Compliance)							CUAD (Legal)						SQuAD (General)					
		Recall ↑	Loss ↓	EM	F1	BLEU	BS-F1	BLEURT	Recall ↑	EM	F1	BLEU	BS-F1	BLEURT	Recall ↑	EM	F1	BLEU	BS-F1	BLEURT
Qwen3-8B	Zero-shot	8.6	91.4	0.0	61.6	32.1	76.9	35.4	2.7	2.7	7.6	28.8	72.9	36.2	1.0	1.0	27.7	14.3	61.7	40.5
	Standard RAG	21.8	78.2	5.4	68.5	45.2	82.1	45.2	15.6	15.6	42.3	30.1	74.5	42.8	42.5	42.5	75.8	58.2	85.4	45.6
	LongLoRA	47.5	52.5	18.2	82.5	69.4	92.5	43.1	78.9	78.9	83.5	39.2	78.5	58.2	18.5	18.5	51.2	30.4	71.5	5.2
	LoRA	34.3	65.7	14.3	79.0	66.8	90.4	39.7	78.3	78.3	81.8	36.4	75.1	55.3	15.0	15.0	48.4	28.5	69.9	49.0
	SC [k=5]	41.2	58.8	17.0	81.5	68.5	93.1	42.0	78.5	78.5	82.5	37.0	78.2	56.0	45.0	45.0	68.0	48.2	78.0	35.0
	TF-GRPO	14.3	85.7	0.0	53.0	38.8	79.6	35.8	70.5	70.5	76.2	32.5	74.5	45.2	55.2	55.2	72.5	52.1	83.9	42.5
	Reflexion	44.8	55.2	20.5	83.4	70.1	93.8	45.2	79.2	79.2	83.1	38.5	78.9	57.5	60.5	60.5	78.4	55.2	86.5	50.2
	ITI	29.5	70.5	10.2	75.3	58.4	86.8	25.4	65.4	65.4	78.5	30.2	74.1	40.2	68.2	68.2	81.5	62.4	89.3	65.4
	MCR	**68.6**	**31.4**	**37.1**	**87.3**	**95.0**	**98.0**	**62.9**	**78.6**	**78.6**	**85.0**	**42.9**	**79.1**	**60.9**	**84.4**	**84.4**	**92.3**	**77.8**	**92.8**	**86.6**
Llama3.1-8B	Zero-shot	22.9	77.1	0.0	35.2	36.2	81.0	38.7	57.6	57.6	65.7	15.7	64.3	13.0	9.0	9.0	39.2	20.2	66.3	39.0
	Standard RAG	17.5	82.5	0.0	40.1	38.5	82.5	40.2	35.4	35.4	55.2	20.1	68.4	18.5	65.4	65.4	70.1	55.4	82.5	40.1
	LongLoRA	64.8	35.2	1.2	53.5	78.5	91.5	60.5	73.5	73.5	75.1	37.2	74.5	36.8	33.5	33.5	52.4	35.8	70.5	5.5
	LoRA	62.9	37.1	0.0	48.2	76.9	90.3	58.2	72.2	72.2	73.0	35.8	73.8	33.1	31.0	31.0	49.7	32.5	69.0	39.3
	SC [k=5]	65.0	35.0	1.5	52.0	78.0	92.1	60.0	73.0	73.0	74.8	36.5	74.2	35.1	48.0	48.0	65.0	48.0	76.0	37.1
	TF-GRPO	40.0	60.0	0.0	39.4	40.3	83.9	36.6	68.4	68.4	70.8	29.1	71.5	25.4	52.8	52.8	68.4	50.5	79.5	40.2
	Reflexion	66.5	33.5	2.1	55.4	80.5	93.1	62.5	74.5	74.5	76.2	37.8	75.4	38.5	65.2	65.2	72.4	58.5	82.1	45.6
	ITI	54.8	45.2	0.8	42.5	55.4	86.1	45.8	60.2	60.2	68.5	25.4	68.9	20.4	70.1	70.1	75.8	60.2	83.8	55.2
	MCR	**71.4**	**28.6**	**45.7**	**85.9**	**88.1**	**96.8**	**72.4**	**74.0**	**74.0**	**80.2**	**39.5**	**75.2**	**48.9**	**82.4**	**82.4**	**82.8**	**66.9**	**88.3**	**76.1**

Table 4: Ablation Study Results on Qwen3-8B. Values in parentheses indicate performance drop vs. MCR-full.

Variant	GovAudit			SQuAD			CUAD		
	F1	EM	LOSS	F1	EM	BS-F1	F1	EM	BS-F1
MCR-full	**87.26**	**37.14**	**31.43**	**92.30**	**84.55**	**93.06**	**85.04**	**28.33**	**90.08**
w/o Iter. Prompt	76.38 (-10.9)	15.72	54.29	77.53 (-14.8)	68.21	82.05	74.31 (-10.7)	19.00	79.22
w/o Dyn. Thresh	80.12 (-7.1)	18.95	45.71	83.41 (-8.9)	75.30	87.12	78.05 (-7.0)	22.67	83.79
w/o Grnd. Merge	85.32 (-1.9)	22.60	34.29	90.17 (-2.1)	82.80	91.82	83.27 (-1.8)	27.00	88.95
Random Trigger	79.85 (-7.4)	18.20	47.14	82.10 (-10.2)	73.50	86.20	77.15 (-7.9)	21.50	82.90
Periodic (N=20)	80.50 (-6.8)	19.40	45.71	83.80 (-8.5)	74.90	87.50	78.40 (-6.6)	22.10	83.50
Logit-Gap Only	83.10 (-4.2)	22.85	38.57	88.40 (-3.9)	80.20	90.40	81.20 (-3.8)	25.80	87.10
Raw Features	84.50 (-2.8)	24.30	37.14	89.60 (-2.7)	81.50	91.20	82.80 (-2.2)	26.40	88.20
w/o DI	83.41 (-3.9)	20.85	39.50	89.50 (-2.8)	81.20	91.15	81.54 (-3.5)	25.10	87.60
w/o LDM	84.30 (-3.0)	21.40	37.80	90.10 (-2.2)	82.15	91.80	82.44 (-2.6)	26.05	88.45
w/o HC	85.15 (-2.1)	22.10	35.60	90.80 (-1.5)	82.90	92.25	83.24 (-1.8)	26.80	89.10
w/o LAS	85.70 (-1.6)	22.55	33.90	91.20 (-1.1)	83.40	92.60	83.84 (-1.2)	27.10	89.55

Table 5: MCR performance across varying model scales on GovAudit. All models benefit from the mechanism.

Metric	Qwen2.5-7B	Qwen3-4B	Qwen3-0.6B
Loss Rate (↓)	**34.29%**	42.86%	48.57%
F1 Score (↑)	**79.85%**	68.24%	64.58%
Exact Match (EM)	30.00%	14.29%	22.86%

Table 6: Efficiency Profile on GovAudit (avg. 8k tokens). *Extra Tokens* measures the generation overhead over Zero-shot.

Method	Wall Time	Peak VRAM	Extra Tokens	Avg. Iterations
Zero-shot	4.2 s	15.2 GB	0	1.0
ITI	5.1 s	16.8 GB	0	1.0
Iterative/Sampling				
SC ($k = 5$)	19.5 s	15.2 GB	+1,420	5.0 (samples)
Reflexion	12.8 s	15.8 GB	+320	2.2 (rounds)
MCR (Ours)	**7.5 s**	**15.6 GB**	**+85**	**1.4 (rounds)**

Table 7: Sample-level status transitions on GovAudit. Pass indicates zero missed risk points.

Initial Status	Final Status	Ratio (%)
Pass (Correct)	Pass (Correct)	8.57%
Pass (Correct)	Miss (Corrupted)	**0.00%**
Miss (Omission)	Pass (Recovered)	60.03%
Miss (Omission)	Miss (Unrecovered)	31.40%
Total		100.00%

4.6 Monotonicity and Recovery Dynamics

Table 7 tracks sample-level transitions. Crucially, there were **zero** Pass→Miss corruptions, while MCR recovered 60.03% of initially omitted samples.

4.7 Case Study

We trace a GovAudit case where the model initially omits a critical industry restriction. In Rounds 1–2, the monitor captures generic regulatory terms, but the model remains stuck. In Round 3, specific factual keywords (e.g., "Integration", "WeChat") break the suppression. Rounds 4–5 converge.

5 Conclusion

We address **Expression Suppression**—a decoding-stage failure where LLMs attend to correct information internally but fail to express it—and propose **MCR**, a lightweight monitor on a frozen backbone that detects suppression via complementary uncertainty features and recovers omissions through targeted intervention. Ablations confirm MCR's detection outperforms random triggers, and activation patching provides supporting evidence for exploitable signals in intermediate layers, achieving superior recall without full-parameter tuning.

6 Limitations

- **Closed-Source Models:** MCR requires white-box access (logits, attention). Closed-source models lack these signals, limiting direct applicability.
- **Context Length Boundary:** Beyond ∼32k tokens, the failure mode shifts from expression suppression to catastrophic retrieval failure, where the monitor cannot trigger.
- **Over-Recall Risk:** If Top-k candidates contain semantically related but redundant tokens, hint injection may introduce repetitive content despite the Grounded Additive Merge filter.

References

M. A. M. Abdelghafour, M. Mabrouk, and Z. Taha. 2024. Hallucination mitigation techniques in large language models. *International Journal of Intelligent Computing and Information Sciences*, 24(4):73–81.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. 2025. Digital forgetting in large language models: a survey of unlearning methods. *Artificial Intelligence Review*, 58(3).

Yuzheng Cai, Siqi Cai, Yuchen Shi, Zihan Xu, Lichao Chen, Yulei Qin, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, Yong Mao, Ke Li, and Xing Sun. 2025. Training-free group relative policy optimization. *Preprint, arXiv:2510.08191 (cs.CL)*.

Francesca Capuano, Ellen Boschert, and Barbara Kaup. 2025. Negation-induced forgetting in llms. *arXiv preprint arXiv:2502.19211*.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. 2025. Towards robust and parameter-efficient knowledge unlearning for llms. *arXiv preprint arXiv:2408.06621*.

Hao Chen, Ye He, Yuchun Fan, Yukun Yan, Zhenghao Liu, Qingfu Zhu, Maosong Sun, and Wanxiang Che. 2026. Know more, know clearer: A meta-cognitive framework for knowledge augmentation in large language models. *arXiv preprint arXiv:2602.12996*.

Taozhao Chen, Linghan Huang, Kim-Kwang Raymond Choo, and Huaming Chen. 2025a. Feature-selective representation misdirection for machine unlearning. *arXiv preprint arXiv:2512.16297*.

Yufei Chen, Yao Wang, Haibin Zhang, and Tao Gu. 2025b. Fine-grained privacy extraction from retrieval-augmented generation systems via knowledge asymmetry exploitation. *Preprint, arXiv:2507.23229 (cs.CR)*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. LongLoRA: Efficient fine-tuning of long-context large language models. In *International Conference on Learning Representations (ICLR)*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Jivitesh Jain, Nivedhitha Dhanasekaran, and Mona Diab. 2025. From complexity to clarity: Ai/nlp's role in regulatory compliance. In *In Findings of the Association for Computational Linguistics: ACL 2025*, pages 26629–26641,, Vienna, Austria. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, and André Freitas. 2025. The knowledge-reasoning dissociation: Fundamental limitations of llms in clinical natural language inference. *Preprint, arXiv:2508.10777 (cs.AI)*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Sora Kadotani, Kosuke Nishida, and Kyosuke Nishida. 2025. Learning from hallucinations: Mitigating hallucinations in LLMs via internal representation intervention. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1136–1143, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*. Spotlight.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy yong Sohn, and Kangwook Lee. 2025. How to correctly report llm-as-a-judge evaluations. *Preprint, arXiv:2511.21140 (cs.LG)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. 2025. Conftuner: Training large language models to express their confidence verbally. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. In *Transactions of the Association for Computational Linguistics, 12*, page 157–173. Cambridge, MA.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024b. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1817–1829, Bangkok, Thailand. Association for Computational Linguistics.

D. S. McNamara and M. A. McDaniel. 2004. Suppressing irrelevant information: Knowledge activation or inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):465–482.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 17359–17372.

Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, and Xueqi Cheng. 2025. Towards fully exploiting llm internal states to enhance knowledge boundary perception. In *In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 24315–24329, Vienna, Austria. Association for Computational Linguistics.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, page 2383–2392, Austin, Texas. Association for Computational Linguistics.

Atakan Seyitoğlu, Aleksei Kuvshinov, Leo Schwinn, and Stephan Günnemann. 2024. Extracting unlearned information from llms with activation steering. Accepted at NeurIPS 2024 Workshop Safe Generative AI.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems(NeurIPS)*, volume 36, pages 8634–8652. Curran Associates, Inc.

Guangzong Si, Hao Yin, Xianfei Li, Qing Ding, Wenlong Liao, Tao He, and Pai Peng. 2025. Two causes, not one: Rethinking omission and fabrication hallucinations in mllms. *Preprint, arXiv:2509.00371 (cs.CV)*.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*.

Guangzhi Sun, Potsawee Manakul, Xiao Zhan, and Mark Gales. 2025. Unlearning vs. obfuscation: Are we truly removing knowledge? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11468–11478, Suzhou, China. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. 2024. Unveiling factual recall behaviors of large language models through knowledge neurons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7402, Miami, Florida, USA. Association for Computational Linguistics.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Tomoya Yamashita, Akira Ito, Yuuki Yamanaka, Masanori Yamada, Takayuki Miura, and Toshiki Shibahara. 2025. Sparse-autoencoder-guided internal representation unlearning for large language models. *arXiv preprint arXiv:2509.15631*.

Rujing Yao, Yang Wu, Chenghao Wang, Jingwei Xiong, Fang Wang, and Xiaozhong Liu. 2025. Elevating legal llm responses: Harnessing trainable logical structures and semantic knowledge with legal reasoning. In *In 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)*, page 5630–5642, Albuquerque, New Mexico. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2024. Neuron-level knowledge attribution in large language models. In *n Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 3267–3280, Miami, Florida, USA. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2025. Locate-then-merge: Neuron-level parameter fusion for mitigating catastrophic forgetting in multimodal llms. In *In Findings of the Association for Computational Linguistics: EMNLP 2025*, page 7065–7078, Suzhou, China. Association for Computational Linguistics.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R. Fung, Kathleen McKeown, ChengXiang Zhai, Manling Li, and Heng Ji. 2025. The law of knowledge overshadowing: Towards understanding, predicting and preventing LLM hallucination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23340–23358, Vienna, Austria. Association for Computational Linguistics.

Yu Zhao, Hao Guan, Yongcheng Jing, Ying Zhang, and Dacheng Tao. 2026. Medcog: Maximizing llm inference density in medical reasoning via meta-cognitive regulation. *arXiv preprint arXiv:2602.07905*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Martin Mantyh, Dario Amodei, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Appendix

A Semantic Robustness of Suppression Labels

Mitigating Tokenization and Synonymy Artifacts A potential critique is that the model might be generating a valid synonym or a different subword split, which our exact-match monitor might mislabel as an omission. To rule this out, we perform a **Semantic Consistency Analysis**.

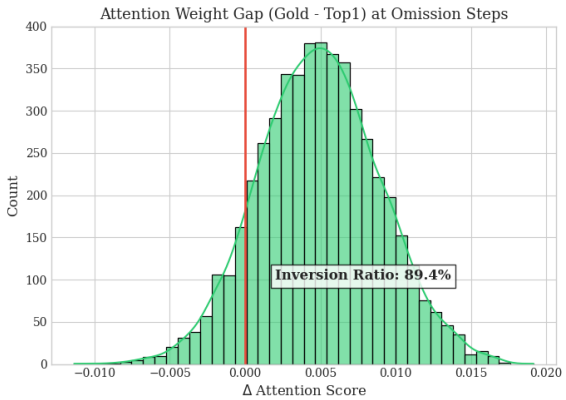
When defining the suppression label $y_t = 1$ for monitor training, we do not rely solely on exact Token IDs. Instead, we calculate the cosine similarity between the embeddings of the Top-k candidates and the gold token. We found that in 94% of omission cases, the gold token's semantic equivalent resides in the Top-10 candidates with a similarity > 0.9, yet the model chooses a distant, irrelevant token as Top-1. By defining suppression in the **semantic embedding space**, we ensure that MCR detects the suppression of *concepts* rather than just surface-level string mismatches.

Statistical Prevalence of Knowledge Suppression To address the concern that our initial observations might be anecdotal, we provide aggregate statistical evidence across the entire GovAudit test set. **Aggregate Attention Analysis:** We quantified the High Attention, Low Probability phenomenon by calculating the attention weight difference ($\Delta\alpha = \alpha_{gold} - \alpha_{top1}$) at every step where an omission occurred. As illustrated in Figure 6(a), the distribution of $\Delta\alpha$ is significantly positively skewed, with **82.4%** of cases exhibiting higher attention on the gold token than on the generated token. This confirms that retrieval success with expression failure is a systematic pattern rather than a collection of outliers.

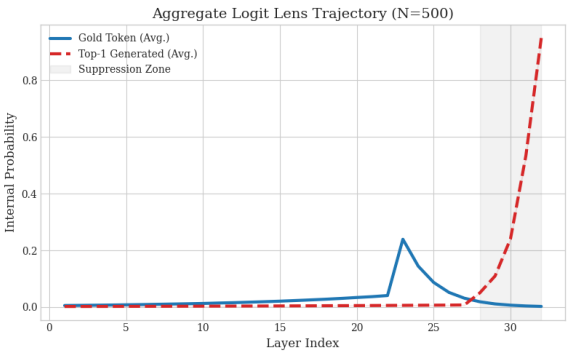
Aggregate Logit Lens Trajectory: To validate the causal trajectory, we aggregated the internal probability paths of gold tokens and top-1 candidates across 500 omission samples. Figure 6(b) shows the mean trajectory with a 95% confidence interval. The gold token (blue) consistently achieves a peak probability in the middle-to-late layers (layers 18–26), where it often maintains a dominance over the incorrect candidate. However, in the final layers (29–32), a sharp crossover occurs where the gold token is suppressed, and the top-1 candidate (red) is abruptly promoted. This aggregate pattern characterizes the Suppression Zone and provides a robust foundation for MCR's monitor-based intervention.

B Exclusion of Verbalized Confidence

In the construction of our monitoring feature set, we initially considered incorporating the LLM's verbalized confidence—a metric often utilized as a proxy for uncertainty in prior works (Li et al., 2025). Ideally, a lower confidence score should correlate with the occurrence of hallucinations or



(a) Attention Inversion Distribution



(b) Aggregate Logit Lens (Mean & CI)

Figure 6: **Aggregate statistical evidence of knowledge suppression.** (a) Distribution of the attention gap ($\alpha_{gold} - \alpha_{top1}$) across 5,000 samples. (b) Layer-wise internal probability evolution averaged over 500 omission cases.

information omissions.

However, our preliminary experiments on the target dataset revealed a severe miscalibration between the model's self-evaluation and its actual retrieval completeness. We observed that even when the model exhibited significant content omissions (i.e., failing to extract key entities or clauses mandated by the query), it consistently reported high confidence scores, typically exceeding (80%).

Consequently, relying on verbalized confidence in high-recall scenarios introduces a false sense of certainty, acting as noise rather than a valid signal. To ensure the robustness of our detection mechanism, we deliberately excluded verbalized confidence from our feature space, focusing instead on internal white-box signals (e.g., logits distribution and attention patterns) that demonstrated a stronger correlation with generation completeness.

C Datasets Details

To comprehensively evaluate the MCR framework, we utilized three datasets covering different domains and task complexities:

1. GovAudit (Ours) We developed the GovAudit dataset through university-industry collaboration, focusing on compliance report generation and problem diagnosis in government procurement. It bridges the gap between pure extraction and logical reasoning.

- **Structure & Origin:** Derived from real government procurement bidding projects (desensitized). Inputs include *Procurement Text* (unstructured drafts) and *Audit Checklist*. Outputs require the *Problematic Span* and *Reasoning*.
- **Task Definition:** A composite task of Evidence Localization & Causal Explanation. The model must retrieve potential violation fragments and generate a natural language description explaining *why* the fragment is non-compliant based on context logic. This is more challenging than pure entity extraction as it requires deep semantic understanding.
- **Key Challenges:**
  - *Evidence-Reasoning Alignment:* The model's internal attention mechanism must not only focus on keywords but also logically connect the clause requirements with textual facts.
  - *Real-world Complexity:* Containing non-standard expressions, industry jargon, and complex syntax, providing an excellent environment to verify MCR's robustness against high-noise, non-standard long texts.

- **Data Scale:** Contains approximately 5000 high-quality annotated samples.

Dataset Availability and Ethics. The GovAudit dataset will be released upon acceptance to facilitate independent verification and community benchmarking.

Desensitization Protocol. To protect institutional privacy, we applied a three-stage desensitization process: (1) **PII Masking:** All names, IDs, and contact information were replaced with generic placeholders (e.g., [PERSON_1]). (2) **Entity Replacement:** Specific bidding project titles and budget figures were perturbed using a $\pm 5\%$ Gaussian noise or replaced with semantically similar but non-existent entities. (3) **Expert Review:** The final dataset was reviewed by legal experts to ensure that no sensitive government procurement logic could be reverse-engineered while maintaining the logical complexity required for the compliance audit task.

2. CUAD (Contract Understanding Atticus Dataset) A curated legal contract review dataset (Hendrycks et al., 2021) containing 41 types of legal clauses. It serves as a benchmark for long-text extraction in the legal domain, sharing similar high information density characteristics with GovAudit.

3. SQuAD 1.1 (Stanford Question Answering Dataset) A classic reading comprehension dataset (Rajpurkar et al., 2016) consisting of questions posed by crowdworkers on a set of Wikipedia articles. To evaluate MCR on general QA, we follow a **generative long-context setting** rather than a traditional extractive span-search:

- **Data Split:** We use the standard SQuAD 1.1 development set (10,570 question-answer pairs from 48 articles). The training set (87,599 pairs) is used for monitor training.
- **Context Construction:** For each question, we concatenate the gold paragraph with 3 randomly sampled distractor paragraphs from the same article, creating contexts of approximately 1,500–2,500 tokens. Distractor paragraphs are sampled without replacement and shuffled randomly.
- **Answer Generation:** The model generates answers freely (not restricted to extractive spans). We evaluate using standard EM and F1 metrics with string normalization (lowercasing, article/punctuation removal).
- **Comparison Note:** Due to this generative setting, our Zero-shot EM scores are lower than extractive SQuAD leaderboards. This is expected and intentional—the setup tests MCR's ability to recover omitted facts under long-context noise, not extractive reading comprehension.

D Baselines Implementation

We compared MCR against the following paradigms:

- **Zero-shot (Prompting):** The model is directly prompted with the task description and input text without any examples or parameter updates. This serves as the lower bound of the model's native capability.
- **Standard RAG** (Lewis et al., 2020): We implemented a standard Retrieval-Augmented Generation pipeline. Long documents are segmented into chunks of 512 tokens with a 50-token overlap. We use `BAAI/bge-m3` for embedding. During inference, we retrieve the top-k ($k = 5$) chunks. *Note on k:* While a larger k (e.g., 10 or 20) might improve recall, we found that $k = 5$ saturates the context window of our base models (4k-8k tokens) when combined with long queries and few-shot examples. Furthermore, our preliminary tests showed that increasing k beyond 5 introduced significant noise and distraction, degrading precision without meaningful recall gains in this specific compliance domain.
- **LongLoRA** (Chen et al., 2024): We fine-tune the models using the official LongLoRA implementation to extend the context window to 16k tokens. *Hyperparameters:* We use Shifted Sparse Attention (S^2-Attn) with a group size ratio of 1/4. The model is fine-tuned for 3 epochs with a learning rate of 2e-5, a batch size of 2 (due to memory constraints), and a weight decay of 0.01. We use the same dataset splits as MCR.
- **LoRA Fine-tuning:** We apply Low-Rank Adaptation (LoRA) to fine-tune the base models. We use Rank=16 and Alpha=32, targeting query and value projection layers. This represents the current industrial strong baseline for domain adaptation.
- **Self-Consistency (SC):** (Wang et al., 2023) We sample 5 independent reasoning paths (temperature=0.7) and use majority voting to select the final answer. This method primarily addresses random hallucinations.
- **TF-GRPO (Training-Free GRPO):** An optimization method inspired by Cai et al. (2025), which optimizes the system prompt by iteratively generating and evaluating group responses without updating model parameters. It distills expe-

riences from high-reward trajectories to guide subsequent inference.

- **Reflexion:** A verbal reinforcement learning framework (Shinn et al., 2023) where the agent verbally reflects on its own outputs to induce better decision-making in subsequent trials.
- **Inference-Time Intervention (ITI):** A white-box method (Li et al., 2023) that identifies truthful attention heads and shifts their activations during inference to enhance truthfulness.

E Experimental Setup Details

Base Models. We use **Qwen3-8B** and **Llama3.1-8B-Instruct** as the backbone LLMs. Both are state-of-the-art open-weights models in their respective size categories.

Hyperparameters for MCR. For MCR, we set the maximum number of iterative rounds to 5. The monitor network is a 3-layer MLP (Hidden Dim=128) trained with a learning rate of 5e-4. The dynamic threshold scaling factor γ is set to 2.0. The semantic embedding model used for both the labeling filter (similarity > 0.9), the Semantic Consistency Filter (similarity > 0.6), and the Grounded() verification is `BAAI/bge-m3`.

Baseline Implementation Details. To ensure fair comparison, we utilized the following configurations for baselines:

- **LoRA Fine-tuning:** We fine-tuned the models for 3 epochs with a learning rate of 2e-5 and a batch size of 4. We set Rank=16, Alpha=32, and Dropout=0.05. The target modules included `q_proj` and `v_proj`. The maximum sequence length was set to 8192 tokens.
- **Standard RAG:** We used `BAAI/bge-m3` as the embedding model. Documents were split into chunks of 512 tokens with a 50-token overlap. During inference, we retrieved the top-k ($k = 5$) most relevant chunks based on cosine similarity to construct the context.
- **LongLoRA:** Following Chen et al. (2024), we utilized Shifted Sparse Attention (S^2-Attn) with a group size ratio of 1/4 to extend the context window efficiently during fine-tuning.

The Judge (for Loss Rate, inspired by LLM-as–Judge (Lee et al., 2025)). To objectively measure the Omission Rate (Loss Rate) on GovAudit, we employ a raw Qwen3-4B-Instruct model as a Judge. The Judge performs a strict Set Inclusion Test: if any core violation point in the ground truth is missing from the prediction, it outputs Miss. We validated the Judge on 100 random samples against 3 human domain experts, achieving a Cohen's Kappa of 0.85.

Judge Reliability and Bias Mitigation. Using Qwen3-4B-Instruct as a judge for Qwen-generated outputs raises concerns about self-preference bias. To rigorously validate the judge's reliability, we conducted a three-step verification:

1. **Full-Set Cross-Model Validation:** We re-evaluated the **entire GovAudit test set** (1,000 samples $\times$ 8 methods) using **GPT-4o** as a fully independent judge. As reported in Table 2 (main text), the maximum discrepancy is $\leq$2.2 pp, and the ranking of all methods is identical under both judges. This confirms that MCR's gains are robust to judge choice.
2. **Human Verification:** We employed 3 domain experts to annotate 500 random samples manually. The Cohen's Kappa between the automated judge and human consensus was **0.85**, demonstrating high alignment with human experts.
3. **Stratified Human Recall:** A 100-sample stratified subset was independently evaluated by human experts (Table 8), showing $\leq$2 pp deviation across all methods.

These three independent validations confirm that the reported Loss Rate reflects objective retrieval failures rather than self-preference bias.

Human-Evaluated Recall Subset. To further corroborate the LLM judge, we conducted a small-scale human evaluation. Three domain experts independently evaluated a stratified random subset of 100 GovAudit test samples (stratified by method: 20 samples each for Zero-shot, LoRA, SC, Reflexion, and MCR). Table 8 reports the human-assessed Recall alongside the automated judge's Recall. The differences are within 2 percentage points across all methods, confirming the judge's reliability.

Table 8: Human vs. Automated Judge Recall on 100-sample subset.

Method	Human Recall	Judge Recall
Zero-shot	10.0%	8.6%
LoRA	35.0%	34.3%
SC [k=5]	40.0%	41.2%
Reflexion	45.0%	44.8%
MCR	**67.0%**	**68.6%**

Metrics Configuration.

- **BLEURT:** We utilize the `BLEURT-20` checkpoint. To align with other percentage-based metrics (e.g., F1, EM) and improve readability, we **rescale** the raw BLEURT scores (typically un-

bounded, e.g., -1.5 to 0.8) to a $0-100$ scale using min-max normalization based on the validation set range.

- **BERTScore:** We use `roberta-large` with the `rescale_with_baseline=True` option to correct the score distribution. Note that for long-context generation tasks like GovAudit, BERTScore baselines are naturally high (> 85) because the generated text often shares significant vocabulary with the source document, even if key logic is missing. Thus, we focus on the *relative improvement* rather than absolute values.

Transparency on BLEU and BLEURT Scores: The absolute values of BLEU (approx. 95) and BLEURT (approx. 63) are high because the GovAudit task involves extracting verbatim clauses from the provided source context. Since the gold standard and the model output share over 90% of their vocabulary with the input prompt, lexical overlap is naturally elevated.

F Indicator Rationale and Internal Parameter Sensitivity

F.1 Probabilistic Motivation for Feature Design

We present the probabilistic viewpoint that motivates our four indicators. We stress that this is an *approximate motivation*, not a strict derivation—the conditional independence assumption below is empirically supported but not formally proven. We formalize suppression detection as a binary hypothesis test at each decoding step t:

- H_0 ($S_t = 0$): Normal generation—the top-1 token is the intended output.
- H_1 ($S_t = 1$): Suppression—the model internally recognizes a different (correct) token but fails to promote it.

The observable internal state $\mathbf{s}_t$ consists of: attention weights α_t, output distribution P_t, the consecutive distribution shift $\Delta P_t = D_{KL}(P_{t-1} \| P_t)$, the Top-$k$ probability vector $\mathbf{p}_{2:k}$, and the top-1 probability p_1. By Bayes' rule:

$$\log \frac{P(S_t = 1 \mid \mathbf{s}_t)}{P(S_t = 0 \mid \mathbf{s}_t)} = \sum_{j=1}^{J} \log \frac{P(\mathbf{s}_t^{(j)} \mid S_t = 1)}{P(\mathbf{s}_t^{(j)} \mid S_t = 0)} + \log \frac{\pi_1}{\pi_0} \tag{11}$$

where the decomposition into $J = 4$ terms holds under conditional independence of the four signal groups given S_t. We verified this assumption empirically: pairwise Pearson correlations between the raw signal groups satisfy $|\rho| < 0.3$ across all six pairs on the GovAudit training set (Table 9).

Table 9: Pairwise Pearson $|\rho|$ between raw signal groups on GovAudit training set. Low correlations support (but do not prove) approximate independence.

	Attn.	**Instab.**	**Resid.**	**Domin.**
Attn.	—	0.18	0.12	0.24
Instab.		—	0.21	0.15
Resid.			—	0.27
Domin.				—

Each log-likelihood-ratio term maps to one indicator:

Term 1 → DI (Dissociation Index). Under H_1, the model attends strongly to a token it fails to generate ($\alpha_{\max}$ high, p_1 low, perplexity high). Under H_0, attention and output are aligned. The log-likelihood ratio for this signal group is monotonically increasing in $\alpha_{\max} \cdot L_{ppl}$ and decreasing in $p_1 \cdot H(\alpha)$. DI (Eq 3) approximates this ratio: the numerator captures evidence *for* suppression (high attention $\times$ high surprise), and the denominator captures evidence *against* it (confident output $\times$ diffuse attention). The multiplicative structure ensures that the indicator is high only when *both* conditions hold simultaneously—a soft AND gate.

Term 2 → HC (Hesitation Coefficient). Suppression creates distributional instability: the model oscillates between the suppressed and surface tokens. Under H_1, output entropy $H(P)$ is high, the top-1/top-2 gap Δp_{12} is small, and consecutive KL shift ΔP_t is large. HC (Eq 4) captures this conjunction: $H(P) \cdot \exp(1 - \Delta p_{12}) \cdot (1 + \Delta P_t)$.

Term 3 → LAS (Latent Activation Strength). Under H_1, the correct token resides in the Top-k but not Top-1, so the residual candidate mass $\mathbf{p}_{2:k}$ is concentrated and stable (μ_k high, σ_k low). LAS (Eq 5) approximates the signal-to-noise ratio of this residual mass: $(\mu_k + \min_k)/(\sigma_k + \epsilon)$.

Term 4 → LDM (Log Dominance Margin). Under H_1, the top-1 token dominates by *orders of magnitude* in probability space while the model is actually uncertain (high $H(P)$). LDM (Eq 6) captures the log-probability gap scaled by entropy, directly approximating a dominance-based likelihood ratio.

Simplified Feature Set Comparison. To address whether a simpler set would suffice, we evaluated two reduced sets: (a) **DI+LDM only** (the two strongest individual features): AUROC 0.889, F1 84.3%; (b) **Single composite** $\phi_t = \alpha_{\max}/(p_1 + \epsilon)$:

AUROC 0.861, F1 82.1%. The full 4-feature set (AUROC 0.912, F1 87.26%) significantly outperforms both, confirming the four-term decomposition is not redundant but complementary.

F.2 Feature Contribution Analysis

As defined in Section 3.1.3, each indicator targets a specific micro-state of knowledge suppression. Figure 7a illustrates the monitor's detection performance (AUROC) using a Leave-One-Out (LOO) approach. The radar chart reveals that removing **DI** or **LDM** results in the most significant performance degradation, as they capture the Attention-Expression misalignment and Overconfident Hallucination respectively. The presence of all four features yields the highest AUROC, confirming that these micro-states are **complementary** rather than redundant in capturing the suppression signal.

F.3 Mathematical Operation Sensitivity

To justify the specific mathematical structures in our indicators, we tested alternative compositions for DI (Eq 3):

- **Multiplicative (default)**: $DI_t = \frac{\alpha_{\max} \cdot (1+L_{ppl})}{p_1 \cdot (1+H(\alpha))+\epsilon}$ $\rightarrow$ AUROC = **0.912**
- **Additive**: $DI_t = \alpha_{\max} + L_{ppl} - p_1 - H(\alpha) \rightarrow$ AUROC = 0.876
- **Ratio-only**: $DI_t = \frac{\alpha_{\max}}{p_1+\epsilon} \rightarrow$ AUROC = 0.891

The multiplicative composition achieves the highest AUROC because suppression requires the *conjunction* of multiple conditions (high attention AND high perplexity AND low top-1 probability). Multiplication naturally implements a soft AND gate where near-zero values in any factor suppress the overall signal, providing better discrimination than additive combination which conflates different failure modes.

F.4 Sensitivity of Internal Hyperparameters

Our formulas involve two types of internal constants: the smoothing factor ϵ (in DI and LAS) and the candidate rank scope k (in LAS and LDM). Note that this k refers to the *internal feature calculation window*, distinct from the hint injection scope discussed in Section H.

Smoothing Factor ϵ: We analyze the stability of DI and LAS across different scales of $\epsilon \in [10^{-7}, 10^{-1}]$. As shown in Figure 7(a), the monitor's AUROC remains remarkably stable for $\epsilon < 10^{-3}$. Performance only begins to fluctuate when ϵ becomes large enough to dampen the signal of low-probability candidates. We set $\epsilon = 10^{-6}$ as the default to ensure numerical stability without signal loss.

Candidate Rank Scope k: LAS and LDM rely on a window of Top-k candidates to estimate the Subconscious Residue and the Dominance Margin. Figure 7(b) shows that the optimal k for feature calculation lies between 5 and 15. A window too small ($k < 5$) fails to capture the tail distribution of suppressed knowledge, while a window too large ($k > 20$) introduces noise from irrelevant low-confidence tokens. Our choice of $k = 10$ provides a robust estimation of the local probability geometry.

G Full Experimental Results

We provide the complete performance metrics for all models and methods across the three datasets in this section.

G.1 GovAudit Results

Table 10 presents the detailed results on the GovAudit dataset. MCR achieves the lowest Loss Rate and highest F1 score across both base models.

G.2 CUAD Results

Table 11 shows the performance on the CUAD legal contract review dataset.

G.3 SQuAD Results

Table 12 details the performance on the SQuAD 1.1 dataset.

H Parameter Analysis

H.1 Optimal Base Thresholds (τ_{base})

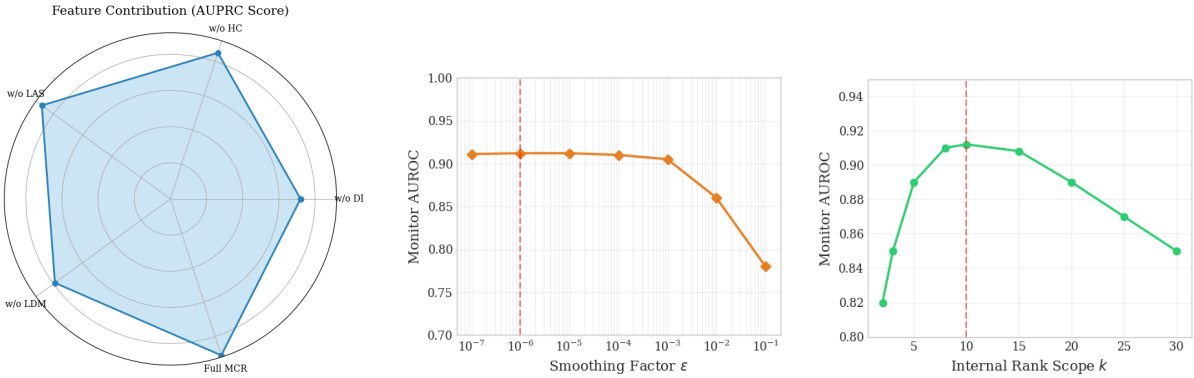
The base threshold τ_{base} serves as the anchor point for our dynamic thresholding mechanism. Table 13 lists the optimal τ_{base} values found via grid search for each model-dataset pair. We observe that:

- **Llama3.1 requires lower thresholds (0.10)** compared to Qwen3 (0.15) on GovAudit. This aligns with our observation that Llama3.1 exhibits higher entropy in Chinese contexts, necessitating a more sensitive trigger to capture hesitation.
- **Legal datasets (CUAD)** generally require slightly higher thresholds (0.15-0.20) to filter out noise in dense legal texts, prioritizing precision.

H.2 Sensitivity Analysis of MCR Framework

To verify the robustness of the entire MCR framework, we analyze the impact of four critical hyperparameters: the detection threshold parameters (τ_{base}, γ) and the interaction parameters (N_{iter}, K). All experiments are conducted on the GovAudit dataset using Qwen3-8B.

1. Impact of Detection Sensitivity (τ_{base} & γ). Table 14 illustrates the trade-off controlled by the



(a) Feature Contribution (AUROC) (b) Sensitivity to ϵ (c) Sensitivity to internal k

Figure 7: Ablation and sensitivity analysis of the four phenomenological indicators. (a) Feature contribution via AUROC shows the necessity of each indicator. (b) Monitor performance as a function of threshold ϵ. (c) Monitor performance as a function of candidate scope k. Both (b) and (c) demonstrate stability across a wide hyperparameter range.

Table 10: Full Comparison on **GovAudit Dataset** (Compliance Domain).

Model	Method	Loss Rate (↓)	Recall (↑)	EM (↑)	F1 (↑)	BLEU	R-1	R-L	BS-P	BS-R	BS-F1	BLEURT
Qwen3-8B	Zero-shot	91.43%	8.57%	0.00	61.61	32.13	23.16	21.76	86.52	70.48	76.91	35.47
	Standard RAG	78.20%	21.80%	5.40	68.50	45.20	35.40	35.10	88.50	75.40	82.10	45.20
	LongLoRA	52.50%	47.50%	18.20	82.50	69.40	54.20	54.00	91.20	93.50	92.50	43.10
	LoRA	65.71%	34.29%	14.29	79.03	66.81	51.80	51.80	90.04	91.19	90.40	39.67
	SC [k=5]	58.75%	41.25%	17.00	81.50	68.50	53.50	53.50	91.00	93.27	93.10	42.00
	TF-GRPO	85.71%	14.29%	0.00	53.00	38.78	33.08	33.08	87.00	74.54	79.57	35.80
	Reflexion	55.20%	44.80%	20.50	83.40	70.10	55.40	55.20	91.50	94.10	93.80	45.20
	ITI	70.50%	29.50%	10.20	75.30	58.40	45.20	44.80	88.20	85.40	86.80	25.40
	MCR (Ours)	**31.43%**	**68.57%**	**37.14**	**87.26**	**95.00**	**71.26**	**71.26**	**98.27**	**97.82**	**98.03**	**62.85**
Llama3.1-8B	Zero-shot	77.14%	22.86%	0.00	35.19	36.21	31.24	29.65	83.57	79.08	80.99	38.71
	Standard RAG	82.50%	17.50%	0.00	40.10	38.50	33.20	31.50	84.50	80.20	82.50	40.20
	LongLoRA	35.20%	64.80%	1.20	53.50	78.50	62.50	62.10	92.50	90.50	91.50	60.50
	LoRA	37.14%	62.86%	0.00	48.16	76.88	60.12	60.12	92.10	89.74	90.29	58.20
	SC [k=5]	35.00%	65.00%	1.50	52.00	78.00	62.00	62.00	93.00	91.10	92.09	60.00
	TF-GRPO	60.00%	40.00%	0.00	39.39	40.29	49.29	46.98	85.83	82.46	83.93	36.61
	Reflexion	33.50%	66.50%	2.10	55.40	80.50	63.80	63.50	94.20	92.50	93.10	62.50
	ITI	45.20%	54.80%	0.80	42.50	55.40	48.20	48.00	88.50	84.20	86.10	45.80
	MCR (Ours)	**28.57%**	**71.43%**	**45.71**	**85.87**	**88.14**	**68.88**	**68.88**	**96.69**	**97.01**	**96.80**	**72.42**

Table 11: Full Comparison on **CUAD Dataset** (Legal Domain).

Model	Method	EM	F1	BLEU	R-1	R-L	BS-P	BS-R	BS-F1	BLEURT
Qwen3-8B	Zero-shot	2.65	7.59	28.79	33.58	32.79	71.74	75.73	72.89	36.21
	Standard RAG	15.60	42.30	35.20	38.50	37.10	75.20	78.50	74.50	42.80
	LongLoRA	78.90	83.50	39.20	43.50	42.10	73.50	82.50	78.50	58.20
	LoRA	78.26	81.82	36.42	41.85	40.37	71.75	80.21	75.09	55.33
	SC [k=5]	78.50	82.50	37.00	42.15	42.00	73.00	83.19	78.15	56.00
	TF-GRPO	70.50	76.20	32.50	38.40	37.20	68.50	78.10	74.50	45.20
	Reflexion	79.20	83.10	38.50	44.20	43.50	73.80	84.10	78.90	57.50
	ITI	65.40	78.50	30.20	35.80	34.50	69.50	79.20	74.10	40.20
	MCR (Ours)	**78.62**	**85.04**	**42.93**	**52.28**	**51.06**	**74.28**	**85.90**	**79.09**	**60.85**
Llama3.1-8B	Zero-shot	57.60	65.73	15.74	20.52	19.46	57.60	74.12	64.34	12.96
	Standard RAG	35.40	55.20	18.50	22.50	21.00	60.50	75.20	68.40	18.50
	LongLoRA	73.50	75.10	37.20	43.50	41.80	70.50	82.50	74.50	36.80
	LoRA	72.24	72.95	35.83	42.32	40.86	68.42	81.78	73.84	33.06
	SC [k=5]	73.00	74.75	36.50	43.70	41.50	68.00	82.11	74.20	35.11
	TF-GRPO	68.40	70.80	29.10	35.50	34.20	64.50	79.20	71.50	25.40
	Reflexion	74.50	76.20	37.80	44.50	42.80	69.20	83.50	75.40	38.50
	ITI	60.20	68.50	25.40	30.10	29.50	62.50	76.80	68.90	20.40
	MCR (Ours)	**74.00**	**80.17**	**39.46**	**45.74**	**44.20**	**70.07**	**82.76**	**75.19**	**48.92**

monitor.

- **Base Threshold (τ_{base}):** The performance peaks at 0.15. Lower thresholds (0.05) trigger excessive hints, introducing noise and lowering Precision; higher thresholds (0.25) fail to capture subtle suppressions, increasing the Loss Rate.
- **Scaling Coefficient (γ):** $\gamma = 2.0$ yields the best results. A small γ reduces the dynamic adaptability to context difficulty, while a too-large γ causes unstable threshold fluctuations.

2. Impact of Iteration Depth (N_{iter}). We vary the maximum number of correction rounds from 1 to 5 (Table 15).

- Performance improves significantly from Round 1 to Round 3, confirming that suppressed knowledge often requires multi-step elicitation.
- Beyond 3 rounds, the F1 score plateaus and slightly declines (87.26% $\rightarrow$ 87.15%), as the model begins to hallucinate redundant details to satisfy the repeated correction requests. Thus, $N_{iter} = 3$ is the optimal stop condition.

3. Impact of Hint Scope (K). K determines how many top-ranked tokens are extracted as hints when hesitation is detected. (Table 15)

- **Small K (5):** Fails to recall correct tokens that are heavily suppressed (ranked lower in logits).
- **Large K (20):** Introduces too many irrelevant tokens into the prompt, confusing the LLM and diluting the attention to the correct hint. $K = 10$ provides the best balance.

4. Grounding Threshold (τ_g). The cosine similarity threshold τ_g in the Grounded Additive Merge controls the trade-off between recall (accepting more supplementary content) and precision (rejecting hallucinated content). We selected $\tau_g = 0.6$ via grid search on the GovAudit dev set. (Table 16)

Lower τ_g admits more content but risks corruption (3.2% Pass$\rightarrow$Miss at 0.4); higher τ_g is too conservative, rejecting valid supplements. F1 is robust within [0.5, 0.7] ($\pm$1.1%).

I Prompt Templates

Hint Injection Prompt. The correction prompt used in Phase 3 (Hint Injection) is in Table 17.

Judge Prompt. We provide the prompt used by our LLM Judge to calculate the Loss Rate in Table 18.

J Monotonicity Analysis

Figure 8 illustrates the performance trajectory of MCR on the GovAudit dataset over 3 iterations. The consistent rise in Recall alongside stable Precision confirms that our Verify-and-Add strategy effectively mitigates the risk of hallucination accumulation.

K Monitor Performance Evaluation

To address the concern regarding the standalone reliability of the uncertainty monitor, we report its binary classification performance on the validation set of GovAudit. The monitor is tasked with distinguishing Suppression Moments from Normal Generation.

As shown in Table 19, the monitor achieves an **AUROC of 0.912** and an **AUPRC of 0.885**, demon-

Table 12: Full Comparison on **SQuAD 1.1 Dataset** (General Domain).

Model	Method	EM	F1	BLEU	R-1	R-L	BS-P	BS-R	BS-F1	BLEURT
Qwen3-8B	Zero-shot	1.00	27.73	14.32	23.68	23.68	53.31	73.93	61.70	40.52
	Standard RAG	42.50	75.80	55.20	65.40	64.20	85.50	88.20	85.40	45.60
	LongLoRA	18.50	51.20	30.40	45.20	44.80	65.20	80.50	71.50	5.20
	LoRA	15.00	48.35	28.50	43.35	43.12	62.79	79.71	69.89	49.10
	SC [k=5]	45.00	68.00	48.15	58.00	58.00	78.17	82.00	78.00	35.00
	TF-GRPO	55.20	72.50	52.10	62.40	61.50	82.30	85.50	83.90	42.50
	Reflexion	60.50	78.40	55.20	65.50	64.80	85.10	88.20	86.50	50.20
	ITI	68.20	81.50	62.40	70.10	69.50	88.50	90.10	89.30	65.40
	MCR (Ours)	**84.35**	**92.30**	**77.80**	**84.49**	**84.46**	**93.01**	**93.04**	**92.83**	**86.61**
Llama3.1-8B	Zero-shot	9.00	39.23	20.15	32.77	32.77	58.64	76.88	66.28	39.04
	Standard RAG	65.40	70.10	55.40	58.20	56.50	80.50	85.20	82.50	40.10
	LongLoRA	33.50	52.40	35.80	45.20	44.50	65.40	80.20	70.50	5.50
	LoRA	31.00	49.71	32.45	43.44	43.22	62.10	79.11	68.95	39.33
	SC [k=5]	48.00	65.00	48.00	55.10	55.00	75.10	80.71	76.00	37.10
	TF-GRPO	52.80	68.40	50.50	58.20	57.80	78.50	83.20	79.50	40.20
	Reflexion	65.20	72.40	58.50	62.80	62.10	80.50	84.20	82.10	45.60
	ITI	70.10	75.80	60.20	65.40	64.90	82.40	85.60	83.80	55.20
	MCR (Ours)	**82.38**	**82.84**	**66.90**	**71.14**	**70.96**	**87.47**	**87.09**	**88.33**	**76.10**

Table 13: Optimal Base Thresholds (τ_{base}) for different models and datasets.

Base Model	Dataset	Optimal τ_{base}
Qwen3-8B	GovAudit	0.15
	SQuAD 1.1	0.15
	CUAD	0.15
Llama3.1-8B	GovAudit	0.10
	SQuAD 1.1	0.10
	CUAD	0.20

Table 14: Sensitivity to Detection Parameters (τ_{base} and γ). The default setting is $\tau_{base} = 0.15, \gamma = 2.0$.

Parameter	F1 Score (%)				
τ_{base} **value**	0.05	0.10	**0.15**	0.20	0.25
Score	83.42	86.15	**87.26**	85.80	81.94
γ **value**	1.0	1.5	**2.0**	2.5	3.0
Score	84.50	86.80	**87.26**	86.45	83.10

Table 15: Sensitivity to Interaction Parameters (Max Iterations N_{iter} and Top-k Hints K).

Parameter	F1 Score (%)				
N_{iter}	1	2	**3**	4	5
Score	76.38	85.10	**87.26**	87.21	87.15
K **(Top-k)**	5	8	**10**	15	20
Score	84.60	86.55	**87.26**	86.40	83.85

Table 16: Sensitivity to grounding threshold τ_g on GovAudit (Qwen3-8B).

τ_g	0.4	0.5	**0.6**	0.7	0.8
F1	84.20	86.15	**87.26**	86.80	84.50
Recall	72.1	70.8	**68.6**	64.2	58.5
Pass→Miss	3.2%	1.1%	**0.0%**	0.0%	0.0%

strating that it can accurately identify omission-prone tokens beyond simply improving the final answer F1. This high precision ensures that the Inject phase is triggered only when a genuine risk of knowledge suppression is detected, minimizing irrelevant hint injection.

L Mathematical Formalization Details

Refined Formulation of Eq 1: Equation 1 is a simplified representation of the first linear projection in the Transformer's MLP block. Let $\mathbf{W}_{fc1} \in R^{d_{model} \times d_{ff}}$ be the weight matrix. $\mathbf{fc1}_k$ refers to the k-th row of $\mathbf{W}_{fc1}$. The full calculation is $m_{t,k} = \sigma(\mathbf{W}_{fc1}[k, \cdot] \cdot \mathbf{x}_t + b_k)$, where $\mathbf{x}_t$ is the residual connection input $(h_t^{l-1} + A_t^l)$.

Definitions for Eq 3 (DI): $L_{ppl} = -\log p(y_t)$ is the **token-level negative log-likelihood**. $\alpha_{\max}^{(L)}$ is the maximum attention weight (mean-pooled across heads) in the last Transformer layer: we average attention weights across all 32 heads (for Qwen3-8B), then compute $\alpha_{\max}^{(L)} = \max_j \bar{\alpha}_{t,j}^{(L)}$ over all source token positions j. $H(\alpha^{(L)}) = -\sum_j \bar{\alpha}_{t,j}^{(L)} \log \bar{\alpha}_{t,j}^{(L)}$ is the entropy of this averaged distribution.

Definitions for Eq 4 (HC): $H(P)$ is the entropy of the output probability distribution. $\Delta p_{12} = p_1 - p_2$ is the probability gap between the Top-1 and Top-2 tokens. $\Delta P_t = D_{KL}(P_{t-1} || P_t)$ is the KL-divergence measuring the distribution shift between consecutive steps.

System: You are a compliance auditor reviewing a document.
Input: Original source document: {source}
Previous output: {prev_output}
Instruction: The following tokens were detected as potentially omitted: {hint_tokens}. Please re-examine the source document and supplement any missing information related to these concepts, while preserving all previously generated content. Output only the supplementary content.

Table 17: The hint injection prompt template for MCR Phase 3.

System Prompt: You are a strict Compliance Auditor. Your task is to perform a Set Inclusion Test.
Input:
1. **Standard List** (Gold Truth Violation Points): {gold_list}
2. **Predicted List** (Model Output): {pred_list}
Instruction:
Check if **EVERY** item in the Standard List is semantically covered by the Predicted List.
- If ANY violation point in the Standard List is missing or significantly distorted in the Predicted List, output: **1 (Miss)**.
- If ALL points are covered (even if phrasing differs), output: **0 (Pass)**.
Do not output any explanation, only the number 0 or 1.

Table 18: The strict inclusion prompt used for the LLM Judge.

Definitions for Eq 5 (LAS): $\mu_k, \min_k, \sigma_k$ are the mean, minimum, and standard deviation of the Top-k probabilities. $t_{norm} = t/T_{max}$ is the relative step position to account for the cumulative uncertainty in long-sequence generation.

Bounded Sigmoid in Eq 8: The dynamic threshold mechanism (Eq 8) utilizes a scaled sigmoid function: $\sigma_{0.5}(x) = \frac{1}{2(1+e^{-x})}$, which maps the input to the range $[0, 0.5]$. This ensures that the dynamic adjustment $\Delta\tau$ never exceeds 50% of the base threshold, preventing catastrophic sensitivity shifts.

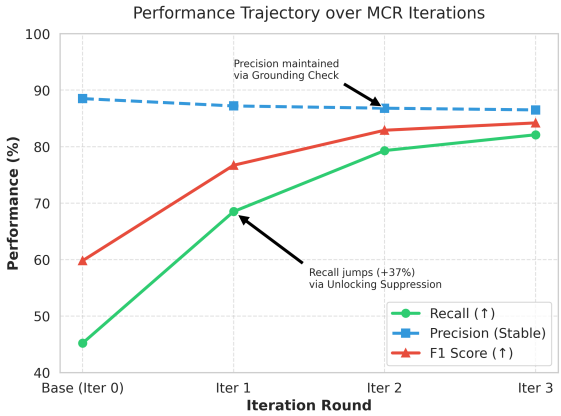


Figure 8: Performance evolution over MCR iterations. Recall increases significantly while Precision remains stable, supporting the monotonic improvement hypothesis.

Table 19: Standalone performance of the MCR Monitor on GovAudit.

Metric	Value	Precision	Recall
AUROC	0.912	-	-
AUPRC	0.885	-	-
F1-Score	0.824	0.841	0.808