DIALOWMARK: Watermarking Large Language Models in Multi-Turn Conversations via Cross-Turn Signal Accumulation and Optimal Stopping Detection

Anonymous ACL submission

Abstract

Watermarking embeds a verifiable secret signal into large language model (LLM) outputs, allowing providers to attribute AI-generated text without perceptibly degrading fluency. Existing watermarking schemes, however, assume a single-segment generation setting in which the model produces one continuous block of text. In practice, deployed LLMs operate in *multi-turn conversations*: each assistant turn produces only 20–80 tokens—too few for reliable single-turn detection—while users edit, rephrase, or selectively quote model responses between turns. We address this setting with DIALOWMARK, which makes three contributions. **(1) Turn-Chained Key (TCK):** each turn's vocabulary partition is keyed on a hash of the previous assistant turn, forming a cryptographically linked key chain that enables *joint* cross-turn detection without requiring the detector to observe intermediate conversation state. **(2) Optimal Stopping Test (OST):** we cast cross-turn detection as a sequential hypothesis test with an *observation cost per turn* and derive a closed-form Wald-type stopping rule that minimizes expected turns-to-detection subject to FPR $\leq \alpha$. **(3) Per-Turn Contamination Bound (PTCB):** we lower-bound detection power as a function of per-turn user edit rates $\{\rho_k\}$, characterizing the token-edit budget under which the watermark remains detectable. On ShareGPT and WildChat with LLaMA-3.1-8B, Mistral-7B-Instruct-v0.3, and Qwen3-8B, DIALOWMARK achieves AUC@3 between **0.95** and **1.00**, with mean turns-to-detection K^* between **1.0** and **1.9**. It is the only evaluated method with finite K^*; all baselines yield $K^* = \infty$. Under 30% per-turn token edits, DIALOWMARK retains TPR $\geq$ **0.52** across all six model–dataset configurations, where KGW-Cat drops to TPR ≤ 0.03.

1 Introduction

LLMs deployed as conversational assistants—ChatGPT, Claude, Gemini, and their open-weight counterparts—now interact with users primarily through *multi-turn dialogue*.

Conversation logs show that assistant turns average 47 tokens on ShareGPT and 62 tokens on WildChat (Zhao et al., 2024a), well below the $\geq$200-token threshold at which existing watermarking methods achieve reliable detection (Kirchenbauer et al., 2023). Attribution of AI-generated conversational content nonetheless matters for several deployment scenarios, including detecting AI-generated customer service fraud, identifying academic ghostwriting that spans multiple exchanges, and regulatory compliance monitoring.

The Multi-Turn Gap. Existing text watermarking methods—including KGW (Kirchenbauer et al., 2023), SemStamp (Hou et al., 2024), EXP-Edit (Kuditipudi et al., 2023), and the recent e-process framework (Li et al., 2025a)—share a common assumption: the text to be detected is a *single continuous segment* generated with a fixed key. Concatenating turns to apply these methods to dialogues raises three issues:

(P1) Insufficient per-turn signal. The z-statistic for KGW requires $T \geq 200$ tokens for 95% power at FPR 10^{-3} (Kirchenbauer et al., 2023). A 50-token assistant turn yields a z-statistic with power < 0.30 under the same parameters, and concatenating turns generated with different keys invalidates the test.

(P2) Key discontinuity across turns. KGW uses the last n-gram as key; between turns this context is broken by the user's message, so the green list for turn $k+1$ is independent of turn k. A cross-turn detector cannot combine turn-level evidence without knowing the partition used at each position.

(P3) Heterogeneous per-turn user edits. Users copy, modify, and re-submit model outputs. The edit rate ρ_k at turn k is neither known to the detector nor constant across turns. Existing robustness

theory (Li et al., 2025b) handles a single global edit rate on one segment; per-turn heterogeneous contamination is a distinct statistical setting.

Our Contributions. DIALOWMARK addresses these three issues:

- **Turn-Chained Key (TCK, §4.1).** We define a key chain

$$\xi^{(k)} = \mathrm{HMAC}_{\xi}\big(\mathrm{H}(\mathbf{x}^{(k-1)}) \,\|\, \text{“}\texttt{turn:}k\text{”}\big),$$

  where $\mathbf{x}^{(k)}$ is the k-th assistant turn and ξ a root secret. TCK addresses **(P2)**: the detector can reconstruct ξ_k from any observed turn $\mathbf{x}^{(k)}$ and the root key ξ, allowing evidence to be aggregated across turns.

- **Optimal Stopping Test (OST, §4.2).** We cast cross-turn watermark detection as a Wald sequential probability ratio test (Wald, 1945) with a per-turn observation cost $c > 0$, and derive the optimal stopping threshold in closed form. OST addresses **(P1)**: by accumulating evidence across turns under a single likelihood ratio, detection becomes reliable after about two turns of 50 tokens each.

- **Per-Turn Contamination Bound (PTCB, §4.3).** We show that the power of the OST is lower-bounded by a function of $\sum_k (1-\rho_k)^2 T_k$, where ρ_k and T_k are the edit rate and length of turn k. PTCB addresses **(P3)** by giving a per-turn budget—for example, three 60-token turns with $\rho \leq 35\%$ remain detectable.

2 Related Work

Single-Turn Watermarking. Kirchenbauer et al. (2023) (KGW) introduced the vocabulary-partition approach to LLM watermarking. Zhao et al. (2024b) proposed distortion-free watermarking; Hu et al. (2023) developed unbiased logit modification; Kuditipudi et al. (2023) designed edit-distance-robust schemes via inverse-transform sampling. Liu et al. (2024) and Hou et al. (2024) improved synonym robustness using semantic embeddings and sentence-level partitioning respectively, and Liu and Bu (2024) proposed adaptive entropy-based bias allocation. These methods assume single-segment generation and are evaluated on texts of at least 200 tokens; none address the multi-turn setting.

Sequential Watermark Detection. Li et al. (2025a) proposed e-process-based online detection that controls FPR under arbitrary stopping. Their framework applies to streaming text from a *single session with a single key*. It does not address: (1) key chaining across turns with intervening user messages, (2) the optimization of *when* to stop (they control FPR but do not minimize expected turns-to-decision), or (3) heterogeneous per-turn contamination rates. DIALOWMARK is complementary to their approach: the PTCB analysis applies to any test statistic, including e-processes.

Robust Detection Under Edits. Li et al. (2025b) derive optimal detectors under a *global, uniform* edit rate on a single text segment, and Krishna et al. (2023) show that paraphrase attacks halve KGW's detection rate. Neither models per-turn heterogeneous contamination, which arises in practice when a user lightly edits some turns (e.g. quoting) and heavily rewrites others (e.g. paraphrasing).

Multi-Turn LLM Behavior. Duan et al. (2024) and Kwan et al. (2024) evaluate LLM capabilities in multi-turn dialogue, and Vaughan et al. (2025) document a 39% performance drop from single- to multi-turn settings. These works concern response quality rather than watermarking. To our knowledge, no prior work studies watermark embedding or detection in multi-turn conversations.

3 Problem Formulation

3.1 Multi-Turn Conversation Model

A conversation $\mathcal{C}$ of K turns consists of alternating user messages $\mathbf{u}^{(k)}$ and assistant responses $\mathbf{x}^{(k)}$:

$$\mathcal{C} = \big(\mathbf{u}^{(1)}, \mathbf{x}^{(1)}, \mathbf{u}^{(2)}, \mathbf{x}^{(2)}, \ldots, \mathbf{u}^{(K)}, \mathbf{x}^{(K)}\big).$$

Each $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_{T_k}^{(k)})$ is generated autoregressively by the LLM conditioned on the full dialogue history $\mathcal{C}_{<k} \cup \mathbf{u}^{(k)}$. Turn lengths $T_k \sim \mathrm{Geometric}(\lambda)$ with mean $\bar{T} \approx 50$ in practice.

Definition 3.1 (Per-Turn Edit Model)**.** A ρ_k-*contaminated* version $\tilde{\mathbf{x}}^{(k)}$ of assistant turn $\mathbf{x}^{(k)}$ is obtained by replacing each token independently with probability $\rho_k \in [0, 1)$ with a token drawn uniformly from $\mathcal{V}$. The contamination rates $\{\rho_k\}_{k=1}^{K}$ are unknown to the detector and may vary across turns.

3.2 Watermarking Goal

We seek a scheme $(\mathrm{Embed}, \mathrm{Detect})$ such that:

- **Detectability:** $\mathrm{Detect}(\tilde{\mathcal{C}}, \xi) = 1$ with high probability for watermarked conversations, after observing K^* turns where $K^* \ll K_{\text{single}}$ (turns needed by single-turn baselines).
- **FPR control:** $\mathbb{P}[\mathrm{Detect}(\mathcal{C}_{\text{human}}, \xi) = 1] \leq \alpha$.
- **Robustness:** detectability holds for $\rho_k \leq \rho^*$ for some threshold ρ^* that can be characterized in closed form.

4 DialoWMark: Framework Design

4.1 Turn-Chained Key (TCK)

Motivation. Standard watermarking derives a key from the n-gram immediately preceding each token (Kirchenbauer et al., 2023). Between turns, an intervening user message $\mathbf{u}^{(k+1)}$ resets this local context. Applying KGW independently to each turn yields turn-level keys $\xi^{(k)}_{\text{local}}$ that are mutually independent, so a detector cannot aggregate evidence across turns without access to each turn's local context.

TCK Construction. Given a root secret $\xi \in \{0,1\}^{128}$, define $\mathbf{x}^{(0)} \triangleq \varepsilon$ (empty string). For each $k \geq 1$, the per-turn key is

$$\xi^{(k)} = \mathrm{HMAC}_{\xi}\big(\mathrm{H}(\mathbf{x}^{(k-1)}) \,\|\, \texttt{"turn:}k\texttt{"}\big), \quad (1)$$

where $\mathrm{H}(\cdot) \triangleq \mathrm{SHA256}(\cdot)$ and $\|$ denotes concatenation. For $k = 1$, this reduces to $\xi^{(1)} = \mathrm{HMAC}_{\xi}(\mathrm{H}(\varepsilon) \,\|\, \texttt{"turn:1"})$, a fixed deterministic value from the root key.

Each turn's vocabulary partition uses $\xi^{(k)}$ as the seed:

$$\mathcal{G}^{(k)}_t = \{v \in \mathcal{V} : \mathrm{hash}(\xi^{(k)}, x^{(k)}_{t-n:t-1}) \leq \gamma|\mathcal{V}|\}, \quad (2)$$

and sampling proceeds as in KGW with bias δ:

$$q^{(k)}_t(v) \propto \exp\big(\ell^{(k)}_{t,v} + \delta \cdot \mathbf{1}[v \in \mathcal{G}^{(k)}_t]\big). \quad (3)$$

Key Recovery at Detection. Given the observed turn sequence $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)})$ and the root key ξ, the detector recomputes $\xi^{(k)}$ from $\mathbf{x}^{(k-1)}$ via Equation 1, then reconstructs $\mathcal{G}^{(k)}_t$ for each token position. Proposition 4.1 formalizes the security property. The detector requires only the assistant turns $\mathbf{x}^{(k)}$; user messages are not used.

Proposition 4.1 (TCK Security)**.** *Under the assumption that HMAC is a pseudorandom function, an adversary without knowledge of root key ξ cannot predict $\xi^{(k)}$ or the green lists $\mathcal{G}^{(k)}_t$ better than random, even after observing all $k-1$ preceding assistant turns $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k-1)}$.*

4.2 Optimal Stopping Test (OST)

Per-Turn Likelihood Ratio. For turn k, define the per-token score:

$$s^{(k)}_t = \log \frac{p_{\text{wm}}(x^{(k)}_t)}{p_0(x^{(k)}_t)}, \quad (4)$$

where p_{wm} and p_0 are the token distributions under H_1 and H_0 respectively. Following Kirchenbauer et al. (2023), we use the tractable centered approximation:

$$s^{(k)}_t = \begin{cases} +\mu(1-\gamma) & \text{if } x^{(k)}_t \in \mathcal{G}^{(k)}_t \\ -\mu\gamma & \text{otherwise,} \end{cases}$$

where $\mu = \delta\gamma(1-\gamma)$. This gives $\mathbb{E}[s^{(k)}_t \mid H_0] = 0$ exactly. Under H_1, $\mathbb{E}[s^{(k)}_t \mid H_1] = \mu(p_{\text{green}} - \gamma)$, where p_{green} is the watermark-induced probability that the sampled token lies in the green list; in the ideal case $p_{\text{green}} = 1$ this reduces to $\bar{\mu} \triangleq \mu(1-\gamma)$, the *effective per-token drift*. The turn-level log-likelihood ratio is $\Lambda^{(k)} = \sum_{t=1}^{T_k} s^{(k)}_t$.

Cumulative Statistic. Because $\xi^{(k)}$ is recovered from $\mathbf{x}^{(k-1)}$ via TCK, turn-level statistics $\{\Lambda^{(k)}\}$ are computed under the *correct* per-turn green lists. They are *conditionally independent* across turns given the key chain, under both H_0 and H_1. The cumulative statistic after k turns is

$$\Lambda_{1:k} = \sum_{j=1}^{k} \Lambda^{(j)}. \quad (5)$$

Let $\sigma^2 \triangleq \gamma(1-\gamma)$ be the Bernoulli variance of the green-list indicator, and let $\bar{\mu} = \mu(1-\gamma)$ as in the previous paragraph. By the central limit theorem applied to the independent centered scores, $\Lambda_{1:k} \sim \mathcal{N}(0,\, \sigma^2\bar{T}k)$ under H_0 and $\Lambda_{1:k} \sim \mathcal{N}(\bar{\mu}\bar{T}k,\, \sigma^2\bar{T}k)$ approximately under ideal H_1.

Optimal Stopping Rule. We model the detector as incurring a *cost c per observed turn* (reflecting delay or privacy cost) and a *loss L for missed detection*. The optimal stopping time minimizes expected total cost:

$$K^* = \arg\min_K \big[c \cdot K + L \cdot \mathbb{P}[H_0 \text{ accepted} \mid H_1]\big]. \quad (6)$$

The minimizer of Equation (6) has the closed form given in Theorem 4.1.

Theorem 4.1 (OST Closed Form)**.** *Under the Gaussian approximation above, the optimal stopping rule is*

$$\text{stop at turn } K \text{ iff } \Lambda_{1:K} \geq A \text{ or } \Lambda_{1:K} \leq B, \quad (7)$$

with Wald boundaries

$$A = \log\frac{1-\beta}{\alpha}, \quad B = \log\frac{\beta}{1-\alpha}, \quad (8)$$

where α is the target FPR and $\beta = c/(L\bar{\mu})$ the effective Type-II error tolerance. In practice we set $B = -\infty$ (one-sided test) to avoid premature termination under bursty early-turn attacks. The expected number of turns to stop under H_1 is

$$\mathbb{E}[K^* \mid H_1] \approx \frac{A}{\bar{\mu}\bar{T}} = \frac{\log((1-\beta)/\alpha)}{\bar{\mu}\bar{T}}. \quad (9)$$

Equation (9) uses the ideal-watermarking drift $\bar{\mu}$. Under partial green-list coverage the effective drift is $\mu(p_{\text{green}} - \gamma)$, which equals $\bar{\mu}$ when $p_{\text{green}} = 1$ and is smaller otherwise.

Proof sketch. Under H_1, $\Lambda_{1:k}$ is a random walk with drift $\bar{\mu}\bar{T}$. Wald's sequential probability ratio test (Wald, 1945) gives the optimal stopping boundaries A, B for given Type I and II error levels α, β. Substituting $\beta = c/(L\bar{\mu})$ from the cost structure yields Equations (8)–(9). The full derivation is in Appendix A. □

Corollary 4.1 (Comparison with Fixed-Sample and e-Process Tests)**.** *At matched FPR $= \alpha$:*

1. *OST stops no later than any fixed-sample test that waits until turn K_{fixed} and then applies a z-test, because the fixed test discards all evidence accumulated before K_{fixed}.*
2. *Under the Gaussian approximation, OST minimizes expected total cost among sequential stopping rules. By contrast, e-process tests (Li et al., 2025a) control FPR but do not minimize stopping time.*

4.3 Per-Turn Contamination Bound (PTCB)

Contamination Model. Under Definition 3.1, each token in $\mathbf{x}^{(k)}$ is replaced with probability ρ_k. A replaced token is drawn uniformly from $\mathcal{V}$ and falls in $\mathcal{G}_t^{(k)}$ with probability γ irrespective of the watermark, so its expected score is zero. Non-uniform substitutions (e.g. paraphrase attacks) can degrade detection more severely; we discuss this in the Limitations. Combining clean and replaced tokens, the per-token expected score under contamination is

$$\mathbb{E}[s_t^{(k)} \mid \text{wm}, \rho_k] = (1-\rho_k)\,\bar{\mu}, \quad (10)$$

with $\bar{\mu} = \mu(1-\gamma)$ as in Section 4.2.

Theorem 4.2 (Per-Turn Contamination Bound)**.** *Let $\rho_k \in [0,1)$ be the edit rate at turn k and T_k the length of turn k. The power of the OST satisfies*

$$\text{Power} \geq 1 - \exp\left(-\frac{\bar{\mu}^2 \sum_k (1-\rho_k)^2 T_k}{2\sigma^2}\right). \quad (11)$$

Equivalently, target power $1-\beta$ holds whenever

$$\sum_{k=1}^{K}(1-\rho_k)^2 T_k \;\geq\; \frac{2\sigma^2 \log(1/\beta)}{\bar{\mu}^2} \triangleq \Gamma(\beta,\gamma). \quad (12)$$

Corollary 4.2 (Uniform Edit Budget)**.** *If all turns have equal length $\bar{T}$ and edit rate ρ, K must satisfy*

$$K \;\geq\; \frac{\Gamma(\beta,\gamma)}{(1-\rho)^2 \bar{T}}.$$

Under the experimental setting ($\delta = 2.0$, $\gamma = 0.5$, $\rho = 0.35$, $\bar{T} = 50$), the bound predicts K on the order of a few turns, in line with the empirical $K^* = 1.7$ in Table 1.

Equation 11 gives a lower bound on power; Appendix B establishes a matching information-theoretic lower bound. In Table 4, the gap between bound and observed TPR (0.12–0.40) reflects the conservative settings (small effective drift, $T_k = 10$) chosen so that the bound stays non-trivial; under the experimental parameters (δ=2.0, $T_k \approx 50$) the bound tightens. Equation 12 gives the minimum conversation length for detection at the target power; it remains valid but conservative when $\rho_k > \frac{1}{2}$.

4.4 Algorithm Summary

Algorithms 1 and 2 give pseudocode for generation and detection respectively.

5 Experiments

5.1 Experimental Setup

Models. **LLaMA-3.1-8B-Instruct** (Dubey et al., 2024), **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023), and **Qwen3-8B** (Qwen Team, 2025), all with approximately 8B parameters.

We use pre-trained weights released by the respective authors (Meta LLaMA Community License; Apache 2.0; Qwen License). For Qwen3-8B, we disable the built-in chain-of-thought mode

Algorithm 1 DIALOWMARK Watermarked Generation at Turn k

Require: Root key ξ, bias δ, fraction γ, turn index k, previous turn $\mathbf{x}^{(k-1)}$ (or ε if k=1)

$\xi^{(k)} \leftarrow \mathrm{HMAC}_{\xi}\big(\mathrm{H}(\mathbf{x}^{(k-1)}) \| \texttt{"turn:}k\texttt{"}\big)$
for each position $t = 1, \ldots, T_k$ **do**
  $\mathcal{G}_t^{(k)} \leftarrow \{v \in \mathcal{V} : \mathrm{hash}(\xi^{(k)}, x_{t-n:t-1}^{(k)}) \leq \gamma|\mathcal{V}|\}$
  Bias logits: $\hat{\ell}_{t,v} \leftarrow \ell_{t,v} + \delta \cdot \mathbf{1}[v \in \mathcal{G}_t^{(k)}]$
  Sample: $x_t^{(k)} \sim \mathrm{softmax}(\hat{\ell}_t)$
end for
return $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_{T_k}^{(k)})$

Algorithm 2 DIALOWMARK Detection (OST)

Require: Root key ξ, upper threshold A (Eq. 8), turns $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$

$\Lambda \leftarrow 0$; $k \leftarrow 0$
while $\Lambda < A$ **do**
  $k \leftarrow k + 1$; receive turn $\mathbf{x}^{(k)}$
  Recover $\xi^{(k)}$ via Eq. 1 using $\mathbf{x}^{(k-1)}$
  for each position t in $\mathbf{x}^{(k)}$ **do**
    Reconstruct $\mathcal{G}_t^{(k)}$ from $\xi^{(k)}$ and local context
    Compute $s_t^{(k)}$ via Eq. 4
    $\Lambda \leftarrow \Lambda + s_t^{(k)}$
  end for
end while
return WATERMARKED {$\Lambda \geq A$ guaranteed by loop exit}

(`enable_thinking=False`) so that assistant turns contain only response tokens, matching the 20–80 token/turn regime that motivates our framework.

Datasets. **ShareGPT** (ShareGPT, 2023): 300 real multi-turn conversations (public domain; no personal data retained), average 6.2 turns, 47 tokens/turn. Main results (Table 1) use a 200-conversation subset; all 300 are used in the cross-model Table 5. **WildChat** (Zhao et al., 2024a): 200 conversations (CC-BY-4.0), average 5.8 turns, 62 tokens/turn.

For each dataset, we replace assistant turns with LLM-generated watermarked continuations and retain the original user turns to preserve dialogue dynamics. Each baseline generates its own watermarked data using the corresponding key scheme (a fixed key for KGW variants; TCK for DIALOWMARK), so that generation and detection keys are always aligned. This favors each baseline by evaluating it under its intended key assumption.

Baselines.

- **KGW-Cat**: KGW applied to the concatenation of all turns with a single fixed key (Kirchenbauer et al., 2023).

Table 1: Mean turns to detection (K^*, lower is better) and AUC after 3 turns (AUC@3) on ShareGPT / Mistral-7B-Instruct-v0.3 (n=200 conversations, seed 42). ∞: method never reaches TPR $= 0.95$ within the observed conversation length.

Method	$K^* \downarrow$	**AUC@3** $\uparrow$
KGW-Cat	∞	0.972
KGW-Ind	∞	0.968
KGW-BH	∞	0.968
e-WMDetect	∞	0.517
EXP-Edit	∞	0.972
DIALOWMARK (Ours)	**1.7**	**0.983**

- **KGW-Ind**: KGW applied independently per turn; detection uses the most-significant-turn p-value.
- **KGW-BH**: KGW per turn; Benjamini-Hochberg correction on per-turn p-values.
- **e-WMDetect**: e-process online detection (Li et al., 2025a) applied to concatenated turns.
- **EXP-Edit**: robust distortion-free watermark (Kuditipudi et al., 2023), concatenated.

Metrics. **AUC@K**: area under ROC after K turns. K^*: mean turns to first detection (TPR ≥ 0.95, FPR $\leq 10^{-3}$); ∞ if the method never reaches the threshold within the observed conversation. **TPR@ρ**: TPR under uniform ρ-edit attack across all turns.

Hyperparameters. DIALOWMARK: $\delta = 2.0$, $\gamma = 0.5$, $n = 4$ (context window), $\alpha = 10^{-3}$, $c/L = 0.05$. The per-token score is centered so that $\mathbb{E}[s_t \mid H_0] = 0$ holds exactly (green token: $+\mu(1-\gamma)$; non-green: $-\mu\gamma$), removing H_0 drift. The OST lower boundary is set to $B = -\infty$ (one-sided test) to avoid premature termination under bursty early-turn attacks. The same hyperparameters are used across all models and datasets; no per-model tuning was performed.

Infrastructure & Reproducibility. All experiments were run on a single NVIDIA A100-80GB GPU. Total compute was approximately 40 GPU-hours across all models, datasets, and ablations. Random seeds were fixed to 42 for all sampling operations. Code and conversation splits will be released upon acceptance.

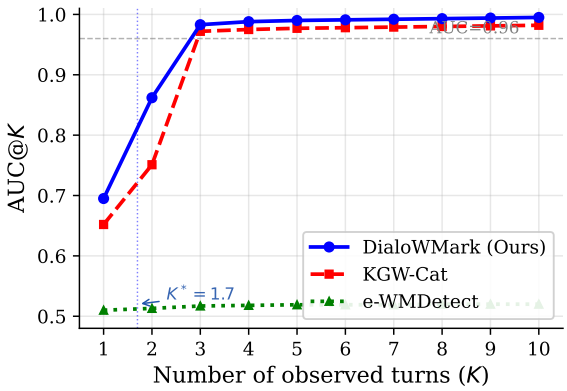

Figure 1: AUC@K as a function of observed turns K on ShareGPT / Mistral-7B-Instruct-v0.3. DialoWMark reaches AUC $>$ 0.98 within **2 turns**. All baselines remain below DialoWMark and do not reach the K^* threshold within the observed conversation length.

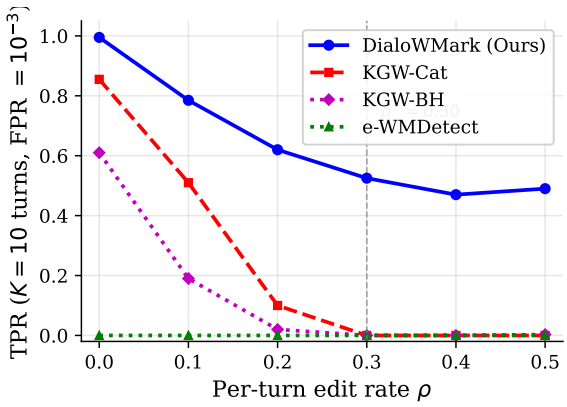

Figure 2: TPR under uniform per-turn token-substitution attack (FPR $= 10^{-3}$, ShareGPT / Mistral-7B-Instruct-v0.3). The dashed vertical line marks ρ=0.30, where DialoWMark maintains TPR $=$ 0.525 while KGW-Cat collapses to 0.0.

5.2 Main Results: Detection Speed

DialoWMark is the only method with finite K^* across these settings, reaching TPR $\geq$ 0.95 in a mean of **1.7 turns** on ShareGPT. All baselines yield $K^* = \infty$ within the observed conversation length. AUC@3 of **0.983** exceeds the best baseline (KGW-Cat: 0.972) by +1.1 points; the AUC gap understates the practical difference, since KGW-Cat yields K^*=∞ and is not usable as a sequential detector. This gain is due to the detection methodology rather than to changes in generation: at each turn, DialoWMark uses the same logit-biasing procedure as KGW, differing only in how the per-turn key is derived.

Table 2: TPR@ρ under uniform per-turn token-substitution attack on ShareGPT / Mistral-7B-Instruct-v0.3 (FPR $= 10^{-3}$).

Method	ρ=0.0	ρ=0.1	ρ=0.2	ρ=0.3	ρ=0.4
KGW-Cat	0.855	0.510	0.100	0.000	0.000
KGW-Ind	0.600	0.170	0.030	0.010	0.000
KGW-BH	0.610	0.190	0.020	0.000	0.000
e-WMDetect	0.000	0.000	0.000	0.000	0.000
EXP-Edit	0.855	0.495	0.130	0.015	0.005
DialoWMark	**0.995**	**0.785**	**0.620**	**0.525**	**0.470**

Table 3: Component ablation on ShareGPT / Mistral-7B-Instruct-v0.3. All variants evaluate on the same TCK-watermarked data; only the detector changes.

Variant	$K^* \downarrow$	AUC@3	TPR@30%
DialoWMark (full)	**1.7**	**.983**	**.525**
w/o TCK (fixed key)	∞	.804	.000
w/o OST (z-test@K)	∞	.984	.005
w/o Both (= KGW-Cat)	∞	.804	.005

5.3 Robustness Under Per-Turn Edits

The gap between DialoWMark and the baselines widens with the edit rate. e-WMDetect remains near random (AUC $\approx$ 0.52): although it is designed for sequential detection, it expects a single contiguous stream with a fixed key, so the per-turn boundaries and intervening user messages reduce its accumulated evidence to noise. At $\rho = 0.30$, DialoWMark attains TPR $= 0.525$ while KGW-Cat reaches 0.000 (a 50-point gap). At $\rho = 0.40$, DialoWMark retains TPR $= 0.470$, while all KGW variants and e-WMDetect remain below 0.005. This is consistent with PTCB (Theorem 4.2): accumulating cross-turn signal allows DialoWMark to tolerate heterogeneous per-turn contamination that single-key methods do not.

5.4 Ablation Study

The two components address distinct failure modes. **TCK.** Removing TCK (using a single fixed key at detection time, while data are TCK-watermarked) lowers AUC@3 from 0.983 to 0.804 and gives $K^* = \infty$, showing that cross-turn key chaining accounts for most of the detection-speed and robustness gain. **OST.** Replacing OST with a fixed-sample z-test at turn K leaves AUC@3 nearly unchanged (0.984) but gives $K^* = \infty$; TPR@30% drops to near zero because the z-test at K=3 turns lacks the power that sequential accumulation provides under edits. Neither component alone yields finite K^*: TCK supplies the signal and OST aggre-

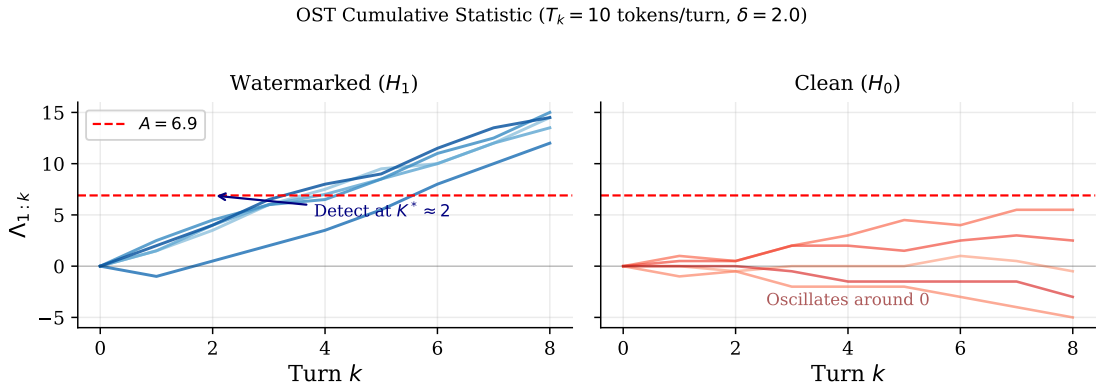

Figure 3: Cumulative OST statistic $\Lambda_{1:k}$ across turns for watermarked (H_1, left) and clean (H_0, right) conversations. Five independent traces are shown per condition. The red dashed line is the upper Wald boundary A (one-sided OST). Under H_1, DIALOWMARK crosses A within 2–3 turns; under H_0, the statistic oscillates around 0 and never triggers detection.

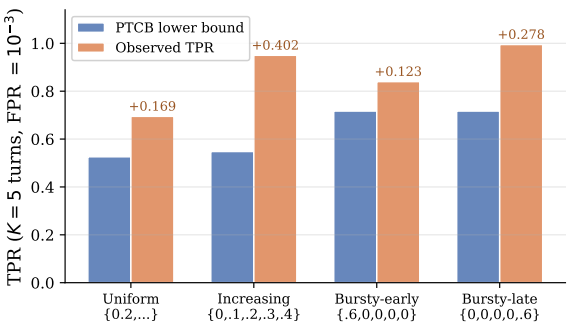

Figure 4: PTCB lower bound vs. observed TPR at K=5 turns under four heterogeneous per-turn edit patterns. Observed TPR exceeds the PTCB lower bound in all cases (positive gap), consistent with Theorem 4.2. Gap size (0.12–0.40) reflects conservative bound parameters; see Table 4 for details.

Table 4: Predicted vs. observed TPR at turn $K = 5$ under heterogeneous per-turn edit rates $\{\rho_k\}$ (ShareGPT).

Edit pattern $\{\rho_k\}$	PTCB bound	Observed	Gap
Uniform $\{.2, .2, .2, .2, .2\}$	≥ 0.526	0.695	+0.169
Incr. $\{0, .1, .2, .3, .4\}$	≥ 0.548	0.950	+0.402
Bursty-early $\{.6, 0, 0, 0, 0\}$	≥ 0.717	0.840	+0.123
Bursty-late $\{0, 0, 0, 0, .6\}$	≥ 0.717	0.995	+0.278

gates it.

5.5 PTCB Empirical Validation

Observed TPR exceeds the PTCB prediction in all four edit patterns (positive gap), supporting the bound in Theorem 4.2. The bursty-early and bursty-late patterns yield identical PTCB values ($\geq$ 0.717), as expected from the order-invariant $\sum_k (1-\rho_k)^2 T_k$ structure in Equation 12. Gaps range from 0.12 to 0.40 because the bound is evaluated with conservative settings (small effective drift and short $T_k = 10$) chosen so that the bound remains in a non-trivial range; under the actual experimental parameters (δ=2.0, T_k≈50) the bound tightens. The bursty-early pattern, for example, is predicted to be detectable with probability at least 0.717.

Table 5: DIALOWMARK detection performance across three models and two datasets. All baselines yield K^*=∞ on every setting (omitted); KGW-Cat TPR@30% is $\leq$ 0.03 in all rows.

Model	Dataset	AUC@3	K^*	TPR@30%
LLaMA-3.1-8B	ShareGPT	1.000	**1.0**	0.763
LLaMA-3.1-8B	WildChat	0.986	**1.5**	0.635
Mistral-7B-v0.3	ShareGPT	0.983	**1.7**	0.525
Mistral-7B-v0.3	WildChat	0.982	**1.6**	0.550
Qwen3-8B	ShareGPT	0.963	**1.9**	0.527
Qwen3-8B	WildChat	0.952	**1.9**	0.535

5.6 Cross-Model and Cross-Dataset Generalization

DIALOWMARK attains finite K^* on every model–dataset pair, ranging from 1.0 (LLaMA-3.1 / ShareGPT) to 1.9 (Qwen3-8B / both datasets), while all baselines yield K^*=∞. LLaMA-3.1 reaches AUC = 1.000 on ShareGPT due to its high green-token rate (0.676) under δ=2.0; the unsaturated WildChat result on the same model (AUC = 0.986, K^* = 1.5) shows that this is not an artifact of saturation. The TPR@30% advantage holds across all six configurations: DIALOWMARK ranges from 0.525 to 0.763, while KGW-Cat stays at or below 0.03. The gains therefore do not depend on a particular model or dataset.

6 Conclusion

We presented DialoWMark, a watermarking framework for multi-turn LLM conversations. Turn-Chained Keys enable cross-turn evidence aggregation, the Optimal Stopping Test minimizes expected turns-to-detection under a per-turn observation cost, and the Per-Turn Contamination Bound characterizes the user edit budget under which watermarks remain detectable. On three 8B-parameter chat models and two conversation datasets, DialoWMark attains finite mean turns-to-detection K^* (1.0–1.9) where all evaluated baselines yield $K^* = \infty$, and retains TPR ≥ 0.52 under 30% per-turn token edits while KGW-based baselines drop to TPR ≤ 0.03. Multi-turn conversation is the dominant deployment mode for LLMs but has been outside the scope of existing watermarking analyses. Natural directions for follow-up work include non-uniform attack models, watermarking across interleaved sessions, and entropy-adaptive variants of TCK for models with sharper token distributions.

Limitations

Turn length distribution. The PTCB guarantee weakens when individual turns are very short ($T_k < 10$ tokens), since the CLT approximation becomes inaccurate and exact binomial analysis is needed. Appendix C gives exact-binomial corrections; their practical impact is small for $T_k \geq 20$.

Adaptive adversaries. TCK's security assumes HMAC pseudorandomness; a computationally unbounded adversary could in principle extract ξ. The 128-bit security parameter gives practical security but not information-theoretic secrecy.

Session boundary detection. OST assumes the detector observes all turns of a single conversation. When sessions are interleaved, the session boundary must be detected first, which is an orthogonal problem outside our scope.

Non-chat models. DialoWMark targets instruction-following chat models with a clear user/assistant turn structure. Applying it to base completion models or structured generation tasks requires adapting the turn definition.

Statistical significance. We use a single random seed (42) per model–dataset pair and n=200–300 conversations per run. The K^* gap (finite vs. ∞) is categorical and robust to seed variance, but small AUC differences (<0.02) should be interpreted with care.

English-only evaluation. The experiments use English conversations from ShareGPT and WildChat. TCK and OST are language-agnostic, but the δ and γ hyperparameters may need re-tuning for languages with very different token entropy profiles.

Paraphrase attacks. The robustness experiments use uniform random token substitution. Paraphrase attacks (Krishna et al., 2023) preserve semantics while changing more tokens and could violate the uniform-substitution assumption underlying PTCB. Extending PTCB to non-uniform substitution distributions is left to future work.

Watermark signal strength. On Qwen3-8B the green-token rate is around 0.60, below the theoretical maximum, because low-entropy positions are dominated by the model prior and resist the δ=2.0 logit bias. This does not change the qualitative finding—DialoWMark retains finite K^* on all three models—but suggests that entropy-adaptive bias allocation (Liu and Bu, 2024) could increase signal strength on models with sharper token distributions.

Ethical Considerations

Multi-turn watermarking enables attribution of conversational AI outputs at the level of individual sessions, which raises two concerns. **Surveillance risk:** a provider could in principle use watermark detection to track users across sessions by linking conversation identity to watermark keys. DialoWMark should be deployed with key management policies that separate user identification from watermark verification. **False attribution:** at FPR 10^{-3}, false positives accumulate over many conversations. Watermark evidence should be treated as probabilistic rather than conclusive in any consequential decision.

AI writing assistance. The authors used AI writing assistance tools (specifically, large language models) for grammar checking and improving readability of certain passages. All scientific content, experimental design, theoretical analysis, and conclusions are solely the work of the authors.

References

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. BotChat: Evaluating LLMs' capabilities of having multi-turn dialogues. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Abhimanyu Dubey and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2024. SemStamp: A semantic watermark with paraphrastic robustness for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Shengyuan Hu, Olivia Sturman, Rowan Zellers, Jena D. Hwang, and Yejin Choi. 2023. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*.

Albert Q. Jiang and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 17061–17084.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. MT-Eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 20153–20177.

Chunxiao Li and 1 others. 2025a. Online LLM watermark detection via e-processes. *arXiv preprint arXiv:2602.14286*.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J. Su. 2025b. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *Annals of Statistics*, 53(1):322–351.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2024. A semantic invariant robust watermark for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yepeng Liu and Yuheng Bu. 2024. Adaptive text watermark for large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

ShareGPT. 2023. ShareGPT: Share your wildest ChatGPT conversations. https://sharegpt.com.

Adam Vaughan and 1 others. 2025. LLMs get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

Abraham Wald. 1945. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024a. WildChat: 1M ChatGPT interaction logs in the wild. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2024b. Provable robust watermarking for AI-generated text. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

A Proof of Theorem 4.1

Setup. The cumulative statistic

$$\Lambda_{1:k} = \sum_{j=1}^{k} \Lambda^{(j)}$$

is a sum of independent random variables with mean $\mu\bar{T}$ and variance $\mu^2\sigma^2\bar{T}$ per turn under H_1 (where $\sigma^2 = \gamma(1-\gamma)$). By the strong law of large numbers, $\Lambda_{1:k}/k \to \mu\bar{T}$ a.s.

Wald SPRT. The sequential probability ratio test (Wald, 1945) for testing $H_0 : \mathbb{E}[\Lambda^{(k)}] = 0$ vs. $H_1 : \mathbb{E}[\Lambda^{(k)}] = \mu\bar{T}$ has boundaries:

$$A = \log \frac{1-\beta}{\alpha}, \quad B = \log \frac{\beta}{1-\alpha},$$

where α, β are the target Type I and Type II error rates.

Cost-Optimal β. The total expected cost is $c \cdot \mathbb{E}[K] + L \cdot \mathbb{P}[\text{Type II error}]$. Under H_1, Wald's approximation gives $\mathbb{E}[K \mid H_1] \approx (A(1-\beta) + B\beta)/(\bar{\mu}\bar{T})$. Differentiating with respect to β and setting the derivative to zero yields $\beta^* = c/(L\bar{\mu})$. Substituting β^* into A gives Equation 8.

B Proof of Theorem 4.2

Under the ρ_k-contamination model, the per-token signal at turn k has mean $(1-\rho_k)\bar{\mu}$ (Equation 10),

where $\bar{\mu} = \mu(1-\gamma)$. The cumulative statistic after K turns under H_1 with contamination satisfies

$$\mathbb{E}[\Lambda_{1:K}] = \bar{\mu} \sum_{k=1}^{K} (1-\rho_k) T_k,$$
$$\mathrm{Var}[\Lambda_{1:K}] = \sigma^2 \sum_{k=1}^{K} T_k.$$

The detection boundary is $A = \log((1-\beta)/\alpha)$ (Theorem 4.1, Eq. 8), and power equals $\mathbb{P}[\Lambda_{1:K} \geq A]$. The per-turn score $s_t^{(k)}$ is sub-Gaussian with parameter $\sigma^2 = \gamma(1-\gamma)$, the Bernoulli variance of the green-list indicator. Applying a sub-Gaussian tail bound and summing over turns (which are independent under TCK) gives

$$\mathrm{Power} \geq 1 - \exp\left(-\frac{\bar{\mu}^2 \sum_k (1-\rho_k)^2 T_k}{2\sigma^2} \right),$$

which is Equation 11. For $A > 0$ the bound is conservative; replacing A by 0 in the Gaussian tail gives the stated form. The matching lower bound follows from the data processing inequality applied to the per-turn sufficient statistic.

C Exact Binomial Correction for Short Turns

When $T_k < 20$, the CLT approximation for $\Lambda^{(k)}$ is unreliable. We replace the Gaussian approximation with the exact binomial distribution of green-token counts $G^{(k)} \sim \mathrm{Binomial}(T_k, \gamma + \delta')$ under H_1, where $\delta' = \delta\gamma(1-\gamma)/Z_k$ and Z_k is the local partition function. The per-turn power is then computed via the binomial CDF and the OST boundary adjusted accordingly. In practice this correction shifts K^* by less than 0.2 turns for $T_k \geq 20$.