

# 000 001 DISTILLING THE THOUGHT, WATERMARKING THE AN- 002 SWER: A PRINCIPLE SEMANTIC GUIDED WATERMARK 003 FOR LARGE REASONING MODELS 004 005

006 **Anonymous authors**  
007 Paper under double-blind review  
008  
009  
010

## 011 ABSTRACT 012

013 Reasoning Large Language Models (RLLMs) excelling in complex tasks present  
014 unique challenges for digital watermarking, as existing methods often disrupt  
015 logical coherence or incur high computational costs. Token-based watermarking  
016 techniques can corrupt the reasoning flow by applying pseudo-random biases, while  
017 semantic-aware approaches improve quality but introduce significant latency or  
018 require auxiliary models. This paper introduces **ReasonMark**, a novel watermarking  
019 framework specifically designed for reasoning-intensive LLMs. Our approach  
020 decouples generation into an undisturbed Thinking Phase and a watermarked An-  
021 swering Phase. We propose a Criticality Score to identify semantically pivotal  
022 tokens from the reasoning trace, which are distilled into a Principal Semantic  
023 Vector (PSV). The PSV then guides a semantically-adaptive mechanism that mod-  
024 ulates watermark strength based on token-PSV alignment, ensuring robustness  
025 without compromising logical integrity. Extensive experiments show ReasonMark  
026 surpasses state-of-the-art methods by reducing text Perplexity by 0.35, increasing  
027 translation BLEU score by 0.164, and raising mathematical accuracy by 0.67 points.  
028 These advancements are achieved alongside a 0.34% higher watermark detection  
029 AUC and stronger robustness to attacks, all with a negligible increase in latency.  
030 This work enables the traceable and trustworthy deployment of reasoning LLMs in  
031 real-world applications.

## 032 1 INTRODUCTION 033

034 Large Language Models (LLMs) have demonstrated remarkable advancements in recent years,  
035 achieving state-of-the-art performance across a multitude of domains including information retrieval  
036 (Labruna et al., 2024; Zhu et al., 2023; Jin et al., 2025), medical diagnosis (Zhou et al., 2024),  
037 financial analysis (Li et al., 2023b; Lopez-Lira et al., 2025), legal assistance (Kuk & Harasta, 2025;  
038 Fei et al., 2023), and academic research (Liao et al., 2024; Naveed et al., 2023). More recently, a new  
039 wave of models, exemplified by systems like ChatGPT-4o (Jaech et al., 2024) and DeepSeek-V2 (Guo  
040 et al., 2025), have showcased superior capabilities in complex reasoning tasks such as mathematical  
041 problem-solving, strategic planning, code generation, and scientific discovery (Guo et al., 2025; Min  
042 et al., 2024; Wei et al., 2025; Wen et al., 2025). These reasoning-intensive LLMs often employ distinct  
043 training paradigms and inference mechanisms, such as internal monologues or chain-of-thought  
044 (CoT) prompting (Wei et al., 2022), which differentiate them significantly from their predecessors.

045 The burgeoning capabilities and widespread adoption of LLMs, particularly those adept at reasoning,  
046 necessitate robust mechanisms for ensuring content authenticity, traceability, and intellectual property  
047 protection. Digital watermarking (Kirchenbauer et al., 2023b) has emerged as a promising technique  
048 to invisibly embed identifiable signals within model-generated text, thereby enabling source tracking  
049 and mitigating misuse (Abdelnabi & Fritz, 2021; Chang et al., 2024; Hou et al., 2023). However,  
050 existing watermarking algorithms, largely developed for general-purpose LLMs, face significant  
051 challenges when applied to reasoning-based models. For instance, methods like KGW (Kirchenbauer  
052 et al., 2023b), which rely on pseudo-random vocabulary partitioning, can inadvertently disrupt the  
053 logical consistency of the model’s internal reasoning—the *thinking phase*—thereby compromising  
the coherence and accuracy of the final answer. Other approaches that focus on preserving text quality,  
such as unbiased sampling techniques (Hu et al., 2023), often do so at the cost of detection efficiency.

054 Conversely, methods like EWD (Lu et al., 2024) and SWEET (Lee et al., 2023), while achieving higher  
 055 detection rates, may introduce perceptible artifacts that degrade text quality. More sophisticated  
 056 strategies like WaterMax (Giboulot & Furion, 2024), which perform multiple generation runs to  
 057 find optimally watermarked outputs, achieve a better balance but incur substantial computational  
 058 overhead and increased inference latency. This persistent trade-off among text quality, watermark  
 059 detectability, and computational efficiency has hindered the practical deployment of watermarking in  
 060 many real-world applications (Liu et al., 2024b).

061 To address these challenges, we introduce **ReasonMark**, a novel watermarking framework specifically  
 062 designed for large reasoning models, centered on the principle of *Distilling the Thought, Watermarking*  
 063 *the Answer*, as illustrated in Fig. 1. Our approach decouples the generation process into two distinct  
 064 stages: an undisturbed internal Thinking Phase, where the model performs its reasoning, and a  
 065 subsequent Answering Phase, where the final response is generated. The core innovation lies in  
 066 preserving the integrity of the thinking phase entirely. We analyze it to identify a set of **Criticality**  
 067 **Tokens (CTs)** that encapsulate the most salient semantic anchors of the reasoning process. These  
 068 tokens are then distilled into a continuous vector representation, the **Principal Semantic Vector**  
 069 (**PSV**), which serves as a dynamic semantic compass for the answering phase. The PSV guides a  
 070 semantically-adaptive watermarking mechanism, where the watermark strength applied to candidate  
 071 tokens is modulated by their alignment with the model’s established reasoning trajectory. By aligning  
 072 the watermark with the model’s own logical flow, ReasonMark can embed a robust and detectable  
 073 signal without disrupting coherence or accuracy. This effectively resolves the debilitating trade-off  
 074 between watermark strength and semantic integrity, all while avoiding the additional inference latency  
 075 common in other semantic-aware techniques. Our main contributions are threefold:  
 076

- 076 • We propose a novel two-phase watermarking framework that decouples a model’s internal  
 077 reasoning from its final answer generation. This is the first approach specifically designed to  
 078 protect the outputs of RLLMs without corrupting their logical integrity.
- 079 • We design a principled method to distill the semantic essence of the model’s reasoning pro-  
 080 cess, involving a Criticality Score to identify key tokens and their subsequent transformation  
 081 into a PSV that provides a continuous, directional guide for watermarking.
- 082 • Extensive experiments show ReasonMark surpasses baselines by reducing text PPL by 0.35,  
 083 increasing translation BLEU score by 0.164, and raising mathematical accuracy by 0.67  
 084 points, while also improving detection AUC by 0.34% with negligible latency.

## 086 2 PRELIMINARY

089 In this section, we introduce token-based and semantic-based watermarking methods, outlining their  
 090 respective strengths and limitations. We then formalize the framework of our proposed algorithm and  
 091 define its key concepts.

093 **Related Work.** Existing LLM watermarking research is primarily divided into token-based and  
 094 semantic-based approaches. Token-based methods, such as the seminal work by Kirchenbauer et al.  
 095 (2023b), partition the vocabulary and apply a statistical bias during generation (Hu et al., 2023).  
 096 While effective for detection, their pseudo-random nature can disrupt the logical flow and semantic  
 097 consistency crucial for reasoning tasks (Yoo et al., 2024; Chang et al., 2024). Adaptive watermark  
 098 strength methods like (Wang et al., 2025b) balance the effectiveness-quality trade-off at the list level  
 099 but ignore the intra-list importance of individual tokens. Conversely, semantic-based methods operate  
 100 in the embedding space to improve text quality and robustness against paraphrasing (Ren et al., 2023;  
 101 Hou et al., 2023). However, they often introduce significant computational overhead by requiring  
 102 auxiliary models or architectural modifications (Baldassini et al., 2024), and are not specifically  
 103 tailored to preserve the step-by-step integrity of complex reasoning (Dasgupta et al., 2024). Our PSV-  
 104 guided watermarking framework bridges this gap by distilling the reasoning phase into a continuous  
 105 Principal Semantic Vector, which dynamically modulates watermark strength based on token-PSV  
 106 alignment. This approach ensures the watermark is semantically coherent, thus preserving logical  
 107 integrity—a key weakness of token-based methods—while maintaining high detection efficiency  
 108 and avoiding the extra inference costs typical of many semantic-aware techniques (Baldassini et al.,  
 109 2024). A more comprehensive review of related work is available in Appendix B.

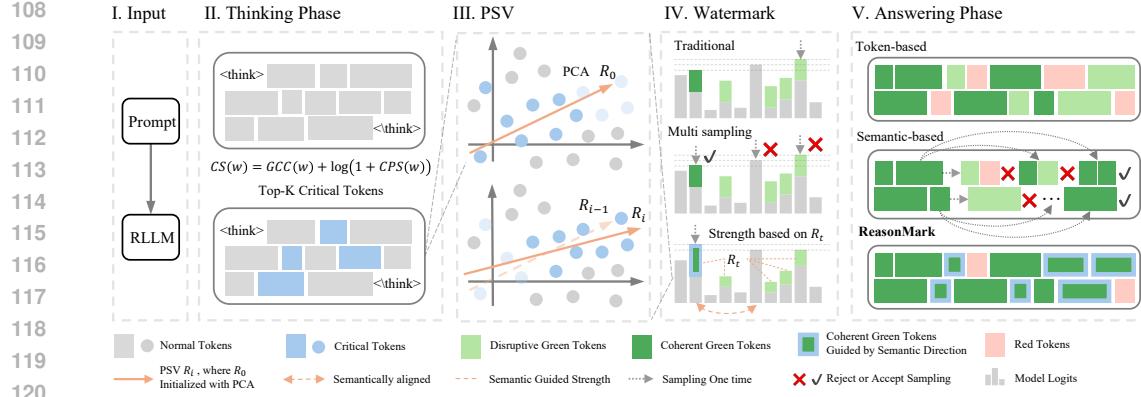


Figure 1: ReasonMark identifies top-K critical tokens during the thinking phase (II.) and uses PCA (III.) to establish an initial Principal Semantic Vector (PSV). This semantic vector then guides the watermarking process (IV.) by dynamically adjusting the logits to favor semantically coherent green tokens and penalize disruptive ones. This enables the efficient generation of a semantically coherent watermarked sequence with a high proportion of green tokens (V.) by sampling only one time.

**Core Framework illustrated.** Our proposed watermarking framework introduces a semantics-aware approach that preserves logical coherence in reasoning-based language models by dynamically aligning watermark strength with semantic relevance. Unlike conventional pseudo-random partitioning methods, our approach first identifies Critical Tokens from the reasoning phase and distills them into a continuous Principal Semantic Vector. The PSV then guides watermark embedding by selectively boosting green-list tokens that are semantically aligned with the reasoning trajectory, even if initially lower in probability. This resolves the fundamental trade-off between watermark strength and semantic integrity—preventing distortion while maintaining high detectability—through adaptive, semantically-grounded modulation.

**Phase Segmentation.** Let  $T = \{t_1, t_2, \dots, t_S\}$  denote the full sequence of  $S$  tokens generated by a language model. We partition  $T$  into two distinct phases. The **Thinking Phase**:  $T_{\text{think}} = \{t_i\}_{i=1}^N = \{t_1, t_2, \dots, t_N\}$ , comprising the model’s internal chain-of-thought or reasoning steps. The **Answering Phase**:  $T_{\text{answer}} = \{t_i\}_{N+1}^S = \{t_{N+1}, \dots, t_S\}$ , representing the final response intended for the user. The delineation point  $k$  is identified via a *Marker-Based Separation Algorithm* (Guo et al., 2025) that detects structural delimiters (e.g., `<think>`, `<\think>`), as illustrated in Fig 1.II.

**Definition 2.1** (Semantic Guidance via Principal Semantic Vector). *Given the thinking phase  $T_{\text{think}}$ , the sequence of probability distributions over the vocabulary  $\mathcal{V}$  during thinking phase  $\{P_i\}_{i=1}^N$ , and the previously generated answer tokens  $\{t_{N+1}, \dots, t_i\}$ , we define:*

$$\mathcal{R}_0 = f_\eta(\{t_i\}_{i=1}^N, \{P_i\}_{i=1}^N), \quad \delta_{i,t_i} = g_\sigma(\mathcal{R}_{i-1}, t_i), \quad \mathcal{R}_i = f_\mu(\mathcal{R}_{i-1}, t_i) \quad (1)$$

Here, the initial PSV  $\mathcal{R}_0$  captures the reasoning trajectory, while  $\delta_{i,t_i}$  provides watermark guidance to token  $t_i$  at step  $i$ . The PSV  $\mathcal{R}_i$  updates dynamically to reflect the evolving semantic context, ensuring watermark strength aligns with the model’s logical flow throughout answer generation. The design of  $f_\eta$  (Fig. 1.II. to III.) poses the primary challenge; thus, the rest of this section details its design, while the implementations of  $g_\sigma$  (Fig. 1.IV.) and  $f_\mu$  (Fig. 1.III.) are deferred to Section 3.2.

**Principle Semantic Vector Construction.** The foundation for constructing PSV rests upon identifying a curated set of **Critical Tokens** (CTs) within the thinking phase  $T_{\text{think}}$ , inspired by (Liu et al., 2025a). These CTs are hypothesized to encapsulate the most salient semantic anchors of the reasoning process and thus provide the essential raw material for deriving the initial PSV. We then formalize the notion that a token’s criticality is a function of both its influence on the generative trajectory and its ability to resolve uncertainty. This principle is encapsulated in the following theorem, which defines an optimal set of CTs.

**Theorem 2.2** (Optimal Representation of Critical Tokens). *The optimal set of Critical Tokens, denoted  $\mathcal{C}^* \subseteq \mathcal{V}$ , is the set that maximizes a joint measure of causal influence and competitive significance, subject to a constraint  $|\mathcal{C}| \leq K$  on its size:*

$$162 \quad \mathcal{C}^* = \arg \max_{\mathcal{C} \subseteq \mathcal{V}, |\mathcal{C}| \leq K} \sum_{w \in \mathcal{C}} [D_{causal}(w\|\theta) + \omega \cdot \mathbb{E}_{j>i} [\Delta S_{i \rightarrow j}(w)]] \quad (2)$$

$$163$$

$$164$$

165 where  $K$  is the maximum desired number of Critical Tokens,  $\theta$  represents model parameters, and  $\omega$   
166 balances the two measures. The Causal Divergence  $D_{causal}(w\|\theta)$  is:

$$167 \quad D_{causal}(w\|\theta) = \sum_{i=1}^N \lambda_i \cdot \|\nabla_{\theta} \mathbb{E}_{w \sim P_i} [Sim(w, w)]\|^2 \quad (3)$$

$$168$$

$$169$$

170 and the Competitive Entropy Reduction  $\Delta S_{i \rightarrow j}(w)$  is:

$$172 \quad \Delta S_{i \rightarrow j}(w) = S(P_j) - S(P_j | w \in Top_k(P_i)) \quad (4)$$

$$173$$

174 Further explanation can be seen at Appendix C. While Theorem 2.2 provides a principled foundation,  
175 the direct computation of the causal divergence term, which requires evaluating gradients with respect  
176 to all model parameters  $\theta$ , is computationally prohibitive for large models.

### 178 3 METHODOLOGY

#### 180 3.1 ALGORITHMIC REALIZATION OF PSV CONSTRUCTION $f_{\eta}$

182 This construction is primarily realized through Critical Tokens (CTs). Consequently, this section  
183 is organized into two main parts: the first details the method for identifying CTs, and the second  
184 explains how these tokens are utilized to construct PSV, corresponding to the function  $f_{\eta}$  in Eq. 1.

##### 186 3.1.1 CRITICALITY SCORE

188 Translating Theorem 2.2 into a practical algorithm, we devise a Criticality Score for each word  
189  $w \in \mathcal{V}$ . The proof of this translation's validity is discussed in detail in Appendix D. This score is a  
190 composite measure reflecting both the global causal influence and the local competitive persistence  
191 of  $w$ .

192 **Global Causal Contribution (GCC).** This component aims to approximate  $D_{causal}(w\|\theta)$  (Eq. equation  
193 3) by quantifying a word  $w$ 's capacity to indirectly shape the reasoning trajectory through  
194 sustained high probability in causally interconnected steps. The GCC is formulated as:

$$196 \quad GCC(w) = \sum_{i=1}^N \left[ P_i(w) \cdot \lambda_i \cdot \sum_{j=i+1}^M \alpha_{i \rightarrow j} \cdot P_j(w) \right] \quad (5)$$

$$197$$

$$198$$

199 The weight  $\lambda_i = JS(P_i \| P_{i-1})$  captures the magnitude of change in the models predictive distribution  
200 from step  $i-1$  to  $i$ . A large JS divergence signals a critical juncture in the reasoning process,  
201 amplifying the contribution of words prominent at such points. The term  $\alpha_{i \rightarrow j} = \frac{Sim(P_i, P_j)}{\sum_{k'=1}^N Sim(P_i, P_{k'})}$   
202 represents the normalized semantic influence of the distributional state at step  $i$  on that of step  $j$ . Here,  
203  $Sim(P_i, P_j) = \frac{P_i \cdot P_j}{\|P_i\| \|P_j\|}$  is the cosine similarity between the vector representations of probability  
204 distributions  $P_i$  and  $P_j$ . This factor ensures that the influence of early critical steps is appropriately  
205 propagated and weighted in assessing a words contribution to later stages of reasoning.

207 **Competitive Persistence Scoring (CPS).** This component approximates  $\mathbb{E}_{j>i} [\Delta S_{i \rightarrow j}(w)]$  (from  
208 Eq. equation 4) by rewarding words that not only feature prominently in competitive generation  
209 contexts but also maintain this prominence over subsequent steps. The CPS for a word  $w$  is calculated  
210 as:

$$212 \quad CPS(w) = \sum_{i=1}^N \left[ S(t_i)^{-1} \cdot (1 - \Delta_i(w)) \cdot \sum_{j=i+1}^M \mathbb{I}(w \in top_k(P_j)) \right] \quad (6)$$

$$213$$

$$214$$

215 The term  $S(t_i)^{-1} = (-\log P_i(t_i))^{-1}$  inversely weights the contribution by the surprisal of the token  
 $t_i$  actually generated at step  $i$ . This rewards contexts where the model makes a high-confidence

choice, suggesting that such choices are more deliberate and impactful. The core of this reward lies in  $\Delta_i(w)$ , which measures the competitive pressure surrounding  $w$  at step  $i$ :

$$\Delta_i(w) = \begin{cases} |L_i(w) - \max_{v \neq w} L_i(v)|, & \text{if } w = t_i \text{ (i.e., } w \text{ was selected)} \\ |L_i(w) - L_i(t_i)|, & \text{if } w \in \text{top}_k(L_i) \text{ and } w \neq t_i \text{ (i.e., } w \text{ was a close competitor)} \\ 1, & \text{otherwise (not competitive)} \end{cases} \quad (7)$$

When  $w$  is the selected token  $t_i$ ,  $\Delta_i(w)$  is its logit margin over the strongest competitor. If  $w$  was a top- $k$  candidate but not selected,  $\Delta_i(w)$  is its logit difference from the winner  $t_i$ . A smaller  $\Delta_i(w)$  indicates more intense competition. The reward thus assigns higher rewards to tokens that emerge from, or are central to, highly contested selection points.

$\sum_{j=i+1}^M \mathbb{I}(w \in \text{top}_k(P_j))$  counts the number of times  $w$  appears among the top- $k$  probability candidates in the  $M - i$  steps immediately following step  $i$ . This serves as empirical validation of  $w$ 's enduring relevance and high-frequency consideration throughout the local reasoning window, reinforcing its status as a critical element.

**Consolidated Criticality Score (CS).** The final score synergistically combines these two aspects to provide a holistic measure of a token's importance.

$$\text{CS}(w) = \text{GCC}(w) \cdot \log(1 + \text{CPS}(w)) \quad (8)$$

The set of Critical Tokens  $\mathcal{C}'$ , is then formed by selecting the  $K$  tokens with the highest CS values, as Fig 1.II illustrated, providing the semantic anchors for the next stage of our methodology. The case study in Appendix I examines the distribution of normalized CS for CTs on different datasets, revealing their correspondence with the input and output of the model.

### 3.1.2 FROM CRITICAL TOKENS TO PRINCIPAL SEMANTIC VECTOR

While the discrete set  $\mathcal{C}'$  identifies key semantic anchors, it falls short of capturing the holistic, relational logic inherent in complex reasoning. To overcome this limitation, we transform this discrete set of tokens into a continuous vector representation, the **PSV**, that encapsulates the dominant semantic direction of the entire thinking phase.

Let  $E(\cdot)$  be the model's token embedding function. We first construct an embedding matrix  $H \in \mathbb{R}^{K \times d}$  by stacking the embeddings of the  $K$  identified Critical Tokens from  $\mathcal{C}'$ , where  $d$  is the embedding dimension.

$$H = [E(w_1), E(w_2), \dots, E(w_K)]^T, \quad \forall w_i \in \mathcal{C}' \quad (9)$$

We then apply Principal Component Analysis (PCA) to  $H$ . The first principal component  $\mathbf{v}_1$ , represents the direction of maximum variance within the embeddings of the most critical tokens. This direction captures their most significant shared semantic properties and reflects the primary axis of the model's reasoning. We define the initial PSV  $\mathcal{R}_0$  as this first principal component:

$$\mathcal{R}_0 = \mathbf{v}_1 = \text{PCA}_1(H) \quad (10)$$

This initial PSV  $\mathcal{R}_0$ , described in Definition 2.1 and Fig. 1.III, acts as a global semantic compass, providing a stable, overarching directional guide for the subsequent answering phase.

### 3.2 SEMANTICALLY-ADAPTIVE WATERMARK EMBEDDING $g_\sigma$ AND $f_\mu$

Our framework departs from conventional methods that employ a fixed watermark strength (Kirchenbauer et al., 2023b), inspired by Wang et al. (2025b). Instead, we introduce a semantically-adaptive mechanism where the watermark's intensity is dynamically modulated based on the alignment of candidate tokens with the current PSV. This allows for a strong watermark signature on semantically coherent tokens while minimizing interference with the model's natural generation process.

**Dynamic Watermark Strength.** At each generation step  $i$  in the answer phase, we partition the vocabulary  $\mathcal{V}$  into a green list  $\mathcal{V}_g$  and a red list  $\mathcal{V}_r$ , based on a standard cryptographic hash of the previous token, following Kirchenbauer et al. (2023b). However, instead of applying a fixed bonus  $\delta$

270 Table 1: Experimental results on C4, WMT16-DE-EN, AIME, and GSM8K datasets for Qwen3-32B  
 271 and Deepseek-R1 32B models. We report Perplexity (PPL)( $\downarrow$ ) for text quality, BLEU and mACC( $\uparrow$ ),  
 272 short for math ACC, for task performance, and AUC ( $\uparrow$ ) for watermark detection. The best result  
 273 among watermarking methods for each metric is in **bold**.

Method	C4				WMT				AIME				GSM8K			
	Qwen3		Deepseek		Qwen3		Deepseek		Qwen3		Deepseek		Qwen3		Deepseek	
	PPL	AUC	PPL	AUC	BLEU	AUC	BLEU	AUC	mACC	AUC	mACC	AUC	mACC	AUC	mACC	AUC
No Watermark	10.55	-	10.82	-	7.851	-	7.622	-	70.03	-	71.52	-	94.01	-	95.21	-
KGW	12.15	98.78	12.52	98.55	7.351	82.36	7.185	81.95	69.23	98.16	70.67	98.43	92.98	94.11	94.16	94.57
UPV	11.41	97.01	11.62	97.15	7.493	82.75	7.288	82.50	63.04	86.94	64.23	87.46	92.51	81.92	93.67	82.45
Unigram	10.66	97.10	10.97	96.90	8.441	83.26	8.157	83.01	56.02	93.27	57.36	93.62	93.88	82.53	94.94	83.07
Unbiased	11.52	93.06	11.92	92.58	7.516	84.50	7.305	84.21	65.51	85.73	66.83	86.26	92.17	80.52	93.21	81.04
SynthID	12.69	87.61	13.22	87.11	6.953	78.15	6.781	77.86	52.34	83.12	53.65	83.67	90.53	74.24	91.46	74.88
SWEET	12.46	97.27	12.62	97.20	7.209	85.10	7.016	84.88	66.01	99.86	67.28	99.89	93.74	92.51	94.82	93.13
EWD	11.89	99.22	12.12	99.18	7.413	86.80	7.228	86.45	69.52	99.91	71.04	99.94	93.67	95.82	94.76	96.41
WatMe	11.27	98.53	11.67	98.60	8.038	86.93	7.893	<b>86.55</b>	67.03	88.11	68.46	88.53	93.82	84.25	94.87	84.74
MorphMark	11.01	94.16	11.22	94.55	9.752	76.08	9.463	75.82	68.74	88.31	70.17	88.79	93.52	76.76	94.63	77.15
SemStamp	11.42	97.85	11.73	97.65	7.912	85.20	7.682	84.80	68.90	98.95	70.31	99.15	93.05	94.80	94.28	95.38
k-SemStamp	11.22	98.10	11.51	97.90	8.123	85.50	7.886	85.15	69.15	99.25	70.55	99.35	93.25	95.10	94.45	95.65
SimMark	11.18	97.95	11.46	97.75	8.191	85.40	7.954	85.00	69.05	99.10	70.48	99.23	93.18	94.95	94.39	95.52
<b>ReasonMark</b>	<b>10.31</b>	<b>99.31</b>	<b>10.54</b>	<b>99.52</b>	<b>9.916</b>	<b>87.25</b>	<b>9.653</b>	85.10	<b>69.86</b>	<b>99.95</b>	<b>71.34</b>	<b>99.98</b>	<b>93.96</b>	<b>95.94</b>	<b>95.14</b>	<b>96.56</b>

289 to the logits of all green-list tokens, we compute a token-specific bonus  $\delta_{i,w}$  for each candidate token  
 290  $w \in \mathcal{V}_g$ . This bonus is proportional to the token’s semantic relevance to the current PSV  $\mathcal{R}_{i-1}$ :

$$s_{w,i} = \frac{E(w) \cdot \mathcal{R}_i}{\|E(w)\| \|\mathcal{R}_i\|}, \quad \delta_{i,w} = \delta_0 + \delta_\lambda \cdot s_{w,i-1} \quad (11)$$

294 where  $s_{w,i-1}$  is the cosine similarity between the embedding of token  $v$  and the PSV  $\mathcal{R}_{i-1}$ .  $\delta_0$  is a  
 295 base watermark strength, and  $\delta_\lambda$  is a scaling factor that controls the sensitivity to semantic alignment.  
 296 The logit for a green-list token  $w$  is then modified as  $L_i(w) \leftarrow L_i(w) + \delta_{i,w}$ . This ensures that  
 297 green-list tokens that are highly aligned with the intended reasoning trajectory receive a stronger  
 298 watermark, reinforcing logical consistency. If a highly probable, contextually appropriate token falls  
 299 into the red list, the relatively lower bonuses on green-list alternatives prevent significant quality  
 300 degradation. The effect of these two parameters on model performance is analyzed in Section 4.5.

301 **Dynamic PSV Update.** The PSV is not static; it evolves with the generation of the answer to act as  
 302 a semantic compass, tracking the local semantic context. After a token  $t_i$  is generated at step  $i$ , we  
 303 update the PSV using an exponential moving average:

$$\mathcal{R}_i = (1 - \beta_i)\mathcal{R}_{i-1} + \beta_i E(t_i), \quad \text{where } \beta_i = \beta_{\text{base}} \cdot s_{t_i, i-1} \quad (12)$$

306 The update rate  $\beta_i$  is itself adaptive, depending on the semantic contribution of the newly generated  
 307 token, where  $\beta_{\text{base}} \in [0, 1]$  is a small base learning rate, which is also analyzed in Section 4.5. This  
 308 mechanism ensures that the PSV gradually incorporates the semantic content of the unfolding answer,  
 309 allowing for smooth topical transitions while remaining anchored to the initial reasoning established  
 310 in  $T_{\text{think}}$ .

311 **Watermark Detection.** A significant advantage of our approach is that the detection process  
 312 requires no modification to standard procedures in Kirchenbauer et al. (2023b). Despite the dynamic  
 313 nature of the watermark embedding, detection remains stateless and does not require access to the  
 314 PSV or the original prompt. It is performed using the same statistical z-test as in KGW by checking  
 315 for a statistically significant bias towards green-list tokens in the generated text. The performance  
 316 gain of the algorithm is attributed to its ability to identify a larger set of valid green tokens from the  
 317 candidate list at each step  $i$ , thereby reducing the number of red tokens.

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETTINGS

323 **Datasets and Prompts.** We evaluate our method on datasets spanning both text generation and  
 324 reasoning tasks. (1) Our evaluation of text generation encompasses two primary tasks. The first

324 is text completion, for which we adopt the C4 dataset (Raffel et al., 2023), which is widely used  
 325 in prior watermarking studies. The first 30 tokens of each sample are taken as prompts, and the  
 326 model generates the continuation. The second is machine translation, and we use the WMT16  
 327 German–English dataset (Bojar et al., 2016), where the task is to translate German sentences into  
 328 English. (2) For reasoning, we employ the AIME (Veeraboina, 2023) and GSM8K (Cobbe et al., 2021)  
 329 mathematical benchmark, which provides standardized solutions, enabling a rigorous evaluation  
 330 of watermarking in tasks where correctness can be objectively assessed. The prompts used in our  
 331 experiments are detailed in Appendix F.1.

332 **Models and Baselines.** We conduct experiments with Qwen3-32B (Yang et al., 2025) and DeepSeek-  
 333 R1-Distill-Qwen-32B (abbreviated as DeepSeek in subsequent section) (Guo et al., 2025) models. To  
 334 ensure a comprehensive comparison, we benchmark our method against a range of representative wa-  
 335 termarking algorithms, including token-based methods KGW (Kirchenbauer et al., 2023b), UPV (Liu  
 336 et al., 2023), Unigram (Zhao et al., 2023), Unbiased (Hu et al., 2023), SWEET (Lee et al., 2023),  
 337 EWD (Lu et al., 2024), WatMe (Chen et al., 2024), and MorphMark (Wang et al., 2025b), as well  
 338 as semantic-based approaches SemStamp (Hou et al., 2023), k-SemStamp (Hou et al., 2024), and  
 339 SimMark (Dabiriaghdam & Wang, 2025). Implementations are facilitated by the MarkLLM (Pan  
 340 et al., 2024) repository. Evaluation Metrics is detailed discussed in Appendix F.2.

341 **Hyperparameters.** For methods requiring a  $\delta$  parameter (e.g., KGW, Unigram), we set  $\delta = 2$  by de-  
 342 fault. **And we set  $\delta_0 = 1.5, \delta_\lambda = 3.0$  in Eq. 11, which is analysed in Sec 4.5.** For text generation tasks,  
 343 we apply repetition penalties to reduce duplicate outputs, including the `no_repeat_ngram_size`  
 344 constraint. For mathematical reasoning tasks, however, we refrain from imposing such penalties, as  
 345 preliminary experiments showed that these constraints significantly reduce problem-solving accuracy,  
 346 regardless of watermarking.

## 347 4.2 MAIN RESULTS

349 As presented in Table 1, ReasonMark demonstrates a superior balance between output quality, task  
 350 performance, and watermark detectability across all evaluated datasets and models, consistently out-  
 351 performing existing state-of-the-art methods. A comparative analysis of inference latency, averaged  
 352 over multiple executions, is detailed in Appendix G.2. The results demonstrate that our method’s  
 353 computational overhead is highly competitive. Furthermore, we provide case studies in Appendix I  
 354 and visualization of PSV and CTs in Appendix G.4, examining the identified critical tokens and the  
 355 embedded watermark to offer qualitative insights into the efficacy of our algorithm.

356 On text generation tasks, our method achieves the lowest perplexity (PPL) on the C4 dataset (10.31  
 357 for Qwen3-32B and 10.54 for Deepseek-R1 32B), indicating the highest text quality that is nearly on  
 358 par with non-watermarked text. For machine translation on WMT16-DE-EN, ReasonMark obtains  
 359 the highest BLEU scores among all watermarking techniques (9.916 and 9.653), showcasing its  
 360 ability to preserve translation fidelity. A more detailed breakdown of the results for various translation  
 361 metrics can be found in Appendix G.1.

362 Crucially, in reasoning-intensive benchmarks, our approach excels at maintaining logical integrity. On  
 363 both AIME and GSM8K, ReasonMark achieves the highest mathematical accuracy (mACC), closely  
 364 matching or even slightly exceeding the baseline performance without a watermark, while other  
 365 methods often lead to a noticeable degradation in performance. For instance, on the AIME dataset  
 366 with the Deepseek model, our method scores 71.34 in mACC, surpassing all other watermarking  
 367 techniques and nearing the 71.52 of the non-watermarked baseline.

368 Across all these tasks, ReasonMark consistently delivers the highest or near-highest detection rates,  
 369 with an AUC of 99.31 and 99.52 on C4, and over 99.9 on the AIME reasoning task. This empiri-  
 370 cally validates that our framework effectively resolves the trade-off between semantic integrity and  
 371 watermark robustness, preserving the performance of reasoning LLMs while embedding a strong,  
 372 detectable signal.

## 373 4.3 ATTACK ROBUSTNESS ANALYSIS

375 In our robustness experiments, we evaluated multiple watermarking algorithms under two attack  
 376 settings, A1 and A2 (Lau et al., 2024) on C4 dataset using Qwen3 model. Attack type A1 applies  
 377 random word-level perturbations, including insertions, deletions, and synonym substitutions, with  
 30 of the tokens modified. Attack type A2 consists of semantic-level transformations via translation

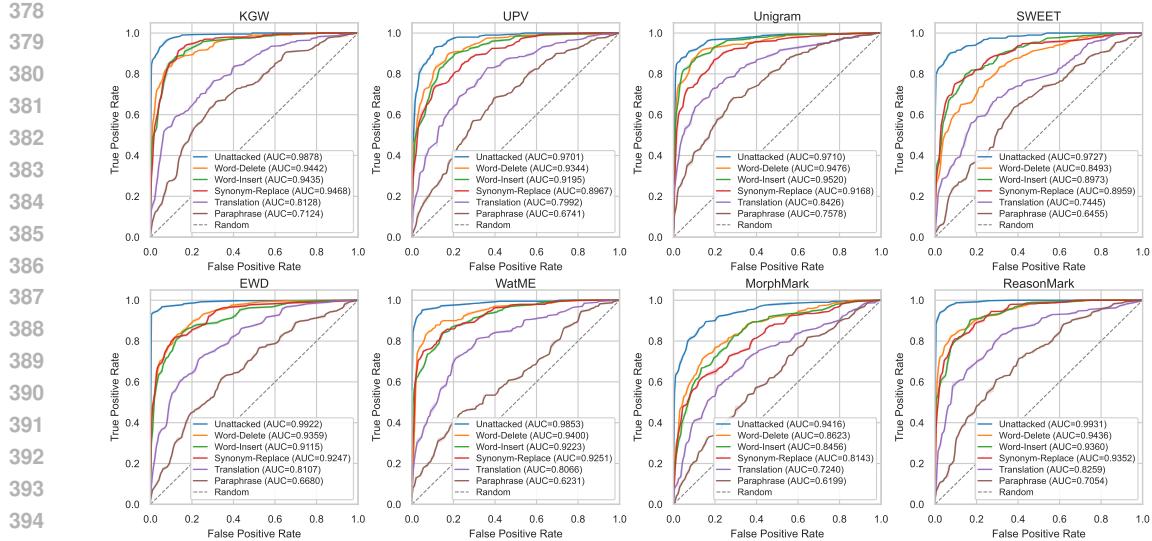


Figure 2: ROC curves under different attack methods for various watermarking approaches.

and paraphrasing, which are implemented through calls to the DeepSeek-V3 API. For the translation attack, each text is translated into Chinese and subsequently back into English.

As shown in Fig 2, ReasonMark demonstrates superior robustness against both word-level and semantic-level attacks. Achieving a near-perfect unattacked AUC of 99.31, it maintains high detectability above 93.5 under word deletion, insertion, and synonym replacement. Crucially, it shows strong resilience to semantic attacks, retaining a high AUC of 82.58 against translation and 70.54 against paraphrasing. This resilience stems from our core principle of embedding the watermark in alignment with the model’s reasoning, captured by the PSV. By tying the watermark to the core semantic meaning rather than the syntactic structure, ReasonMark ensures persistence against such modifications, validating the efficacy of our semantically-grounded approach in adversarial settings.

#### 4.4 ABLATION STUDY

To validate our core components, we conducted an ablation study in Table 2. We tested variants by replacing our Critical Token (CT) selection with random sampling (**w/o CTs**), and by individually removing the Global Causal Contribution (**w/o GCC**) and Competitive Persistence Scoring (**w/o CPS**) modules.

The results confirm that all components are essential. The full ReasonMark model achieves the best performance, with the lowest perplexity (10.3080) and a high AUC (0.9931). The most significant performance drop occurred in the **w/o CTs** variant, where PPL increased to 12.8801. This demonstrates that our principled, semantic-based token selection is critical for maintaining text quality, as random tokens fail to provide coherent guidance for the watermark. Furthermore, removing the GCC or CPS modules also degrades performance. The absence of GCC (**w/o GCC**) primarily impacts text quality (PPL increases to 11.1510), while removing CPS (**w/o CPS**) leads to a more noticeable drop in watermark detectability (AUC falls to 98.69). This shows that GCC is vital for semantic coherence, and CPS is crucial for embedding a robust watermark. In conclusion, the components are synergistic and indispensable for achieving the optimal balance between text quality and detection robustness.

Table 2: Ablation study on the C4 dataset.

Method / Variant	PPL	AUC
No Watermark	10.5488	-
<b>ReasonMark</b>	<b>10.3080</b>	<b>99.31</b>
w/o CTs	12.8801	99.21
w/o GCC	11.1510	99.11
w/o CPS	11.0597	98.69

432 4.5 HYPERPARAMETER SENSITIVITY ANALYSIS  
433

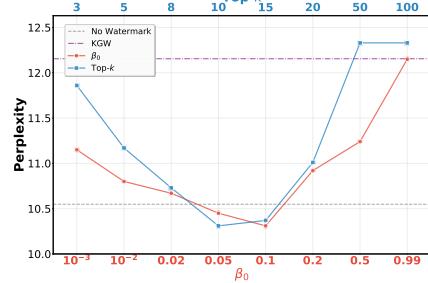
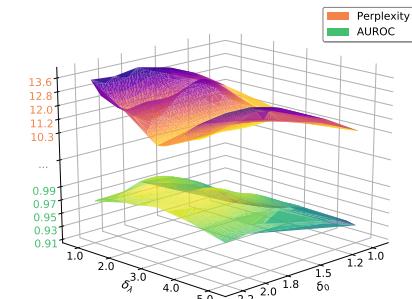
434  **$\beta_0$  and Top-k Parameter Analysis.** To understand the  
435 impact of key hyperparameters on our method’s performance,  
436 we conduct a sensitivity analysis for the PSV update rate  $\beta_0$  and the top-k sampling value. Figure 3 illus-  
437 trates how text quality, measured by perplexity, varies with  
438 these parameters. The analysis reveals that ReasonMark  
439 is robust, showing stable performance across a wide range  
440 of values for both hyperparameters.  
441

442 For the PSV update rate  $\beta_0$ , perplexity follows a U-shaped  
443 curve, starting at 11.1 for a value of  $10^{-3}$ , reaching its  
444 minimum of approximately 10.3 around 0.1, and then  
445 increasing again. Similarly, the top-k parameter shows  
446 that perplexity is highest at a small k of 3, drops to its lowest point around k=10, and then gradually  
447 rises as k increases to 100. Notably, the perplexity of ReasonMark consistently remains well below  
448 the KGW baseline across all tested settings, underscoring a persistent advantage in text quality.  
449 Critically, the results highlight that with careful tuning, our method’s performance can even surpass  
450 the non-watermarked baseline. The optimal configuration, with a  $\beta_0$  value in the range of 0.01 to 0.1  
451 and a top-k value between 10 and 50, yields a perplexity score that is superior to that of the original,  
452 non-watermarked text. This demonstrates that ReasonMark not only embeds a robust watermark but  
453 can also enhance text fluency.  
454

455  **$\delta_0$  and  $\delta_\lambda$  Interaction Analysis.** To maintain the overall  
456 watermark strength approximately consistent with other  
457 methods in our hyperparameter settings, specifically  $\delta = 2$ ,  
458 as outlined in Section 2, and considering the formulation  
459 in Eq. 7, we set  $\delta_0$  within a range of 1 to 2. This ensures  
460 that the overall watermark strength does not deviate signif-  
461 icantly from the baseline methods. Thus, we primarily  
462 focus on adjusting the parameter  $\delta_\lambda$ , varying it from 1 to  
463 5. Figure 4 provides a segmented surface visualization  
464 that reveals the complex interaction patterns between these  
465 critical hyperparameters. The results from the figure indi-  
466 cate that variations in  $\delta_0$  have a more substantial impact on  
467 the AUC, while variations in  $\delta_\lambda$  exert a greater influence  
468 on the PPL. This observation aligns with our algorithmic  
469 design:  $\delta_0$  ensures a fundamental watermark strength, whereas  
470  $\delta_\lambda$  dynamically adjusts the intensity to assign higher watermark  
471 strength to semantically critical tokens, thereby achieving the dual objectives  
472 of effective watermark detection and high text quality. Furthermore, although parameter adjustments  
473 lead to performance variations, the overall efficacy remains within a stable range, demonstrating the  
474 robustness of the ReasonMark algorithm.  
475

476 5 CONCLUSION  
477

478 In this work, we introduced the **ReasonMark**, the first watermarking framework specifically designed  
479 to protect the outputs of reasoning-based Large Language Models. By distinguishing between the  
480 model’s internal thinking process and the final answer generation, our method effectively preserves the  
481 integrity of the model’s reasoning capabilities a critical vulnerability of conventional watermarking  
482 techniques. The core innovations of our approach, including the identification of Critical Tokens  
483 through a principled Criticality Score and their distillation into a continuous Principal Semantic  
484 Vector (PSV), allow for a semantically-aware embedding process. This ensures that the watermark  
485 aligns with the model’s own logical trajectory, resolving the persistent trade-off between watermark  
486 detectability, text quality, and inference cost. Our experiments confirm that the **ReasonMark**  
487 maintains high-quality, logically coherent outputs and robust watermark detection with minimal  
488 inference latency. This work represents a significant step towards enabling safe, traceable, and  
489 accountable deployment of advanced reasoning LLMs in real-world applications. Usage of LLMs  
490 when drafting the manuscript is detailed in Appendix A.  
491

Figure 3: Visualization of  $\beta_0$  and top-k.Figure 4: Visualization of  $\delta_0$  and  $\delta_\lambda$ .

486 REFERENCES  
487

- 488 Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text  
489 provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 121–140.  
490 IEEE, 2021.
- 491 Folco Bertini Baldassini, Huy H Nguyen, Ching-Chung Chang, and Isao Echizen. Cross-attention  
492 watermarking of large language models. In *ICASSP 2024-2024 IEEE International Conference on  
493 Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4625–4629. IEEE, 2024.
- 494
- 495 Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias  
496 Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri,  
497 Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia  
498 Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference  
499 on machine translation. In *Proceedings of the First Conference on Machine Translation*, pp.  
500 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics. URL  
501 <http://www.aclweb.org/anthology/W/W16/W16-2301>.
- 502 Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Wieting, and Mohit Iyyer. Postmark: A  
503 robust blackbox watermark for large language models. *arXiv preprint arXiv:2406.14517*, 2024.
- 504
- 505 Liang Chen, Yatao Bian, Yang Deng, Deng Cai, Shuaiyi Li, Peilin Zhao, and Kam-Fai Wong. Watme:  
506 Towards lossless watermarking through lexical redundancy. In *Proceedings of the 62nd Annual  
507 Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9166–9180,  
508 2024.
- 509 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,  
510 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve  
511 math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- 512 Amirhossein Dabiriaghdam and Lele Wang. Simmark: A robust sentence-level similarity-based  
513 watermarking algorithm for large language models. *arXiv preprint arXiv:2502.02787*, 2025.
- 514
- 515 Agnihib Dasgupta, Abdullah Tanvir, and Xin Zhong. Watermarking language models through language  
516 models. *arXiv preprint arXiv:2411.05091*, 2024.
- 517
- 518 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha  
519 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.  
520 *arXiv e-prints*, pp. arXiv–2407, 2024.
- 521 Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen,  
522 Zongwen Shen, and Jidong Ge. Lawbench: Benchmarking legal knowledge of large language  
523 models. *arXiv preprint arXiv:2309.16289*, 2023.
- 524
- 525 Eva Giboulot and Teddy Furon. Watermax: breaking the llm watermark detectability-robustness-  
526 quality trade-off. *arXiv preprint arXiv:2403.04808*, 2024.
- 527
- 528 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,  
529 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms  
530 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- 531
- 532 Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang,  
533 Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. Semstamp: A semantic  
534 watermark with paraphrastic robustness for text generation. *arXiv preprint arXiv:2310.03991*,  
2023.
- 535
- 536 Abe Bohan Hou, Jingyu Zhang, Yichen Wang, Daniel Khashabi, and Tianxing He. k-semstamp:  
537 A clustering-based semantic watermark for detection of machine-generated text. *arXiv preprint  
538 arXiv:2402.11399*, 2024.
- 539 Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased  
watermark for large language models. *arXiv preprint arXiv:2310.10669*, 2023.

- 540 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec  
 541 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint*  
 542 *arXiv:2412.16720*, 2024.
- 543 Jarosław Janas, Paweł Morawiecki, and Josef Pieprzyk. Llm-text watermarking based on lagrange  
 544 interpolation. *arXiv preprint arXiv:2505.05712*, 2025.
- 545 Bowen Jin, Jinsung Yoon, Zhen Qin, Ziqi Wang, Wei Xiong, Yu Meng, Jiawei Han, and Sercan O  
 546 Arik. Llm alignment as retriever optimization: An information retrieval perspective. *arXiv preprint*  
 547 *arXiv:2502.03699*, 2025.
- 548 John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A  
 549 watermark for large language models. In *International Conference on Machine Learning*, pp.  
 550 17061–17084. PMLR, 2023a.
- 551 John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A  
 552 watermark for large language models. In *International Conference on Machine Learning*, pp.  
 553 17061–17084, 2023b.
- 554 Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for  
 555 uncertainty estimation in natural language generation. In *International Conference on Learning*  
 556 *Representations*, 2023.
- 557 Michal Kuk and Jakub Harasta. Llms & legal aid: Understanding legal needs exhibited through user  
 558 queries. *arXiv preprint arXiv:2501.01711*, 2025.
- 559 Tiziano Labruna, Jon Ander Campos, and Gorka Azkune. When to retrieve: Teaching llms to utilize  
 560 information retrieval effectively. *arXiv preprint arXiv:2404.19705*, 2024.
- 561 Gregory Kang Ruey Lau, Xinyuan Niu, Hieu Dao, Jiangwei Chen, Chuan-Sheng Foo, and Bryan  
 562 Kian Hsiang Low. Waterfall: Framework for robust and scalable text watermarking and provenance  
 563 for llms. *arXiv preprint arXiv:2407.04411*, 2024.
- 564 Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin,  
 565 and Gunhee Kim. Who wrote this code? watermarking for code generation. *arXiv preprint*  
 566 *arXiv:2305.15060*, 2023.
- 567 Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time  
 568 intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information*  
 569 *Processing Systems*, volume 36, pp. 41451–41483, 2023a.
- 570 Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey.  
 571 In *Proceedings of the fourth ACM international conference on AI in finance*, pp. 374–382, 2023b.
- 572 Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S Yu. Watermarking techniques for large  
 573 language models: A survey. *arXiv preprint arXiv:2409.00089*, 2024.
- 574 Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo,  
 575 Joseph Chee Chang, and Amy X Zhang. Llms as research tools: A large scale survey of re-  
 576 searchers' usage and perceptions. *arXiv preprint arXiv:2411.05025*, 2024.
- 577 Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization*  
 578 *Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.  
 579 URL <https://aclanthology.org/W04-1013/>.
- 580 Aiwei Liu, Leyi Pan, Xuming Hu, Shuang Li, Lijie Wen, Irwin King, and S Yu Philip. An unforgeable  
 581 publicly verifiable watermark for large language models. In *The Twelfth International Conference*  
 582 *on Learning Representations*, 2023.
- 583 Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark  
 584 for large language models, May 2024a.
- 585 Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui  
 586 Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM*  
 587 *Computing Surveys*, 57(2):1–36, 2024b.

- 594 Shuliang Liu, Qi Zheng, Jesse Jiaxi Xu, Yibo Yan, He Geng, Aiwei Liu, Peijie Jiang, Jia Liu, Yik-  
 595 Cheung Tam, and Xuming Hu. Vla-mark: A cross modal watermark for large vision-language  
 596 alignment model. *arXiv preprint arXiv:2507.14067*, 2025a.
- 597
- 598 Yepeng Liu and Yuheng Bu. Adaptive text watermark for large language models. *arXiv preprint*  
 599 *arXiv:2401.13927*, 2024.
- 600 Yepeng Liu, Xuandong Zhao, Christopher Kruegel, Dawn Song, and Yuheng Bu. In-context water-  
 601 marks for large language models. *arXiv preprint arXiv:2505.16934*, 2025b.
- 602
- 603 Alejandro Lopez-Lira, Jihoon Kwon, Sangwoon Yoon, Jy-yong Sohn, and Chanyeol Choi. Bridging  
 604 language models and financial analysis. *arXiv preprint arXiv:2503.22693*, 2025.
- 605 Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. An entropy-based text watermarking  
 606 detection method. *arXiv preprint arXiv:2403.13485*, 2024.
- 607
- 608 Minjia Mao, Dongjun Wei, Zeyu Chen, Xiao Fang, and Michael Chau. A watermark for low-entropy  
 609 and unbiased generation in large language models. *arXiv preprint arXiv:2405.14604*, 2024.
- 610
- 611 Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual  
 612 associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pp.  
 613 17359–17372, 2022.
- 614 Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang,  
 615 Xiaoxue Cheng, Huatong Song, et al. Imitate, explore, and self-improve: A reproduction report on  
 616 slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*, 2024.
- 617
- 618 Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman,  
 619 Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language  
 620 models. *arXiv preprint arXiv:2307.06435*, 2023.
- 621
- 622 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong  
 623 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow  
 624 instructions with human feedback. In *Advances in Neural Information Processing Systems*,  
 volume 35, pp. 27730–27744, 2022.
- 625
- 626 Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang  
 627 Liu, Xuming Hu, Lijie Wen, et al. Markllm: An open-source toolkit for llm watermarking. *arXiv*  
 628 *preprint arXiv:2405.10051*, 2024.
- 629
- 630 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic  
 631 evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.),  
 632 *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.  
 633 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.  
 634 doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040/>.
- 635
- 636 Wenjie Qu, Wengrui Zheng, Tianyang Tao, Dong Yin, Yanze Jiang, Zhihua Tian, Wei Zou, Jinyuan  
 637 Jia, and Jiaheng Zhang. Provably robust multi-bit watermarking for {AI-generated} text. In *34th*  
 638 *USENIX Security Symposium (USENIX Security 25)*, pp. 201–220, 2025.
- 639
- 640 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi  
 641 Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text  
 642 transformer, 2023. URL <https://arxiv.org/abs/1910.10683>.
- 643
- 644 Saksham Rastogi and Danish Pruthi. Revisiting the robustness of watermarking to paraphrasing  
 645 attacks. *arXiv preprint arXiv:2411.05277*, 2024.
- 646
- 647 Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. A  
 648 robust semantics-based watermark for large language model against paraphrasing. *arXiv preprint*  
 649 *arXiv:2311.08721*, 2023.
- 650
- 651 Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27  
 652 (3):379–423, 1948.

- 648 Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In  
 649 *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*,  
 650 pp. 368–377, 1999.
- 651
- 652 Hemish Veeraboina. Aime problem set 1983-2024, 2023. URL <https://www.kaggle.com/datasets/hemishveeraboina/aime-problem-set-1983-2024>.
- 653
- 654 Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and  
 655 Xu Sun. Towards codable watermarking for injecting multi-bits information to llms. In *The Twelfth  
 656 International Conference on Learning Representations*, NA.
- 657
- 658 Yidan Wang, Yubing Ren, Yanan Cao, and Binxing Fang. From trade-off to synergy: A versatile  
 659 symbiotic watermarking framework for large language models. *arXiv preprint arXiv:2505.09924*,  
 660 2025a.
- 661 Zongqi Wang, Tianle Gu, Baoyuan Wu, and Yujiu Yang. Morphmark: Flexible adaptive watermarking  
 662 for large language models. *arXiv preprint arXiv:2505.11541*, 2025b.
- 663
- 664 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny  
 665 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in  
 666 neural information processing systems*, 35:24824–24837, 2022.
- 667
- 668 Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao  
 669 Zhang, Yunzhuo Hao, Xuchen Song, et al. Skywork r1v2: Multimodal hybrid reinforcement  
 learning for reasoning. *arXiv preprint arXiv:2504.16656*, 2025.
- 670
- 671 Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu  
 672 Tang, Xiaowei Lv, et al. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond.  
 673 *arXiv preprint arXiv:2503.10460*, 2025.
- 674
- 675 Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement  
 676 learning. *Machine learning*, 8(3):229–256, 1992.
- 677
- 678 Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and  
 679 accessible distribution-preserving watermark for large language models, 2023.
- 680
- 681 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang  
 682 Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*,  
 683 2025.
- 684
- 685 KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. Advancing beyond identification: Multi-bit watermark  
 686 for large language models. In *Proceedings of the 2024 Conference of the North American Chapter  
 687 of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long  
 688 Papers)*, pp. 4031–4055, 2024.
- 689
- 690 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating  
 691 text generation with BERT. *CoRR*, abs/1904.09675, 2019. URL <http://arxiv.org/abs/1904.09675>.
- 692
- 693 Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for  
 694 ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023.
- 695
- 696 Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang,  
 697 Kaishuai Xu, Yi Fang, et al. Large language models for disease diagnosis: A scoping review. *arXiv  
 698 preprint arXiv:2409.00097*, 2024.
- 699
- 700 Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen,  
 701 Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A  
 survey. *arXiv preprint arXiv:2308.07107*, 2023.

702 **A USE OF LLMs**  
703704 The Large Language Model (LLM) was utilized to assist in the language editing and polishing of  
705 this manuscript. Specifically, its application was confined to correcting grammatical errors, refining  
706 sentence structure, and enhancing the overall readability of the text. The LLM was not used for  
707 any part of the scientific process, including the generation of hypotheses, data analysis, or the  
708 interpretation of results. The intellectual content of this paper is entirely the product of the authors,  
709 who have reviewed all revisions and take complete responsibility for the work presented.  
710711 **B DETAILED RELATED WORK**  
712713 The field of digital watermarking for Large Language Models (LLMs) has rapidly expanded to  
714 address the growing need for content authenticity and intellectual property protection (Liu et al.,  
715 2024b; Liang et al., 2024). Existing methods can be broadly categorized into three main paradigms:  
716 vocabulary partitioning-based, semantic-aware, and those specifically targeting the unique challenges  
717 of reasoning-based models.  
718719 **B.1 VOCABULARY PARTITIONING-BASED WATERMARKING**  
720721 The foundational approach in LLM watermarking involves partitioning the model’s vocabulary into a  
722 green list and a red list based on a cryptographic hash of the preceding tokens. During generation,  
723 a positive bias is added to the logits of green-list tokens, embedding a detectable statistical signal  
724 into the output text. The seminal work by Kirchenbauer et al. (2023b) established this paradigm,  
725 demonstrating its effectiveness for generating detectable signals even in short text spans.  
726727 Numerous variants have since been proposed to improve upon this core idea. Some methods focus  
728 on preserving the original output distribution to enhance text quality, employing techniques like  
729 reweighting (Hu et al., 2023), permutation-based schemes (Wu et al., 2023), or sampling-acceptance  
730 protocols (Mao et al., 2024). While adaptive watermarking methods like MorphMark (Wang et al.,  
731 2025b) balance the trade-off between watermark effectiveness and text quality across the red-green  
732 list partition, they do not modulate the watermark strength in proportion to the varying importance of  
733 individual tokens within the green list itself. Others selectively apply watermarking to high-entropy  
734 tokens, particularly in specialized domains like code generation (Lee et al., 2023). Researchers have  
735 also extended this framework to encode multi-bit messages using probability-balanced partitioning  
(Wang et al., NA) or error-correction codes (Qu et al., 2025).  
736737 Despite their widespread adoption, vocabulary partitioning methods share a fundamental limitation:  
738 their reliance on pseudo-random token selection can inadvertently disrupt the logical flow and  
739 semantic consistency of the generated text (Yoo et al., 2024; Chang et al., 2024). This drawback  
740 is particularly pronounced in reasoning-intensive models, where even minor perturbations to the  
741 chain-of-thought can corrupt the entire reasoning process and lead to an incorrect final answer  
(Kirchenbauer et al., 2023b).  
742743 **B.2 SEMANTIC-AWARE AND CONTEXT-GUIDED WATERMARKING**  
744745 To mitigate the quality degradation issues of vocabulary partitioning, a second wave of research has  
746 focused on developing semantic-aware and context-guided watermarking techniques. These methods  
747 move beyond statistical manipulation of token frequencies and instead operate in the semantic space  
748 to embed watermarks in a more natural and robust manner.  
749750 One line of work leverages the semantic embedding space directly. For instance, SemaMark (Ren  
751 et al., 2023) discretizes the embedding space and uses contrastive learning, while SEMSTAMP (Hou  
752 et al., 2023) employs locality-sensitive hashing for semantic partitioning. Other approaches utilize  
753 an auxiliary model to guide the watermarking process, such as generating watermark logits from  
754 semantic embeddings (Liu et al., 2024a), measuring token distribution entropy to decide when to  
755 apply the watermark (Liu & Bu, 2024), or generating dynamic, context-aware instructions for the  
LLM to follow (Dasgupta et al., 2024). Post-hoc methods that do not require access to model internals  
have also been explored; POSTMARK (Chang et al., 2024) instructs an LLM to insert specific words  
756

756 to embed a signal, while In-Context Watermarking (ICW) (Liu et al., 2025b) achieves embedding  
 757 solely through prompt engineering.  
 758

759 While these methods often yield higher text quality and improved robustness against paraphrasing  
 760 attacks (Rastogi & Pruthi, 2024), they typically introduce significant trade-offs. Many require training  
 761 auxiliary models or modifying the model architecture (Baldassini et al., 2024), incurring substantial  
 762 computational overhead and complexity. Furthermore, most of these approaches are designed for  
 763 general text generation and are not specifically tailored to preserve the delicate, step-by-step logical  
 764 integrity required by reasoning-intensive tasks.  
 765

### 766 B.3 WATERMARKING FOR REASONING LLMs

767 The emergence of powerful reasoning LLMs, such as DeepSeek-R1 (Guo et al., 2025) and Skywork  
 768 R1V2 (Wei et al., 2025), has introduced a new frontier for watermarking. These models often employ  
 769 a distinct two-phase generation process, producing an internal thinking phase (e.g., chain-of-thought)  
 770 before delivering the final answer. As noted, applying conventional watermarks uniformly across  
 771 both phases can severely degrade reasoning performance.  
 772

773 A nascent body of work has begun to address this specific challenge. These pioneering approaches  
 774 recognize the importance of the thinking phase and attempt to leverage it for more intelligent  
 775 watermark embedding. For example, some methods extract semantic patterns or critical tokens from  
 776 the reasoning steps to guide the watermarking process in the subsequent answer phase (Yoo et al.,  
 777 2024; Janas et al., 2025; Wang et al., 2025a). This strategy aims to align the watermark with the  
 778 model’s established logical trajectory, thereby preserving coherence.  
 779

780 Our work builds upon this insight but introduces a key innovation. Whereas prior methods rely on  
 781 discrete semantic anchors, we propose distilling the collective essence of the reasoning phase into  
 782 a continuous Principal Semantic Vector (PSV). This vector serves as a dynamic semantic compass  
 783 that guides a semantically-adaptive embedding mechanism in the answer phase. By modulating  
 784 the watermark strength based on each token’s alignment with the overall reasoning trajectory, our  
 785 framework achieves a superior balance of text quality, logical consistency, and watermark detectability  
 786 without incurring additional inference latency.  
 787

## 788 C THEOREM FURTHER EXPLAINED

789 Theorem 2.2 provides a principled basis for identifying tokens that are not merely frequent but are  
 790 integral to the structure and direction of the models thought process. It posits that the optimal set of  
 791 Critical Tokens,  $\mathcal{C}^*$ , is found by maximizing the objective function in Eq. 2. This function combines  
 792 two key aspects: the causal influence of a token and its role in reducing predictive uncertainty,  
 793 balanced by a coefficient  $\omega$ . The maximization is constrained by  $K$ , the maximum allowable number  
 794 of critical tokens, and depends on model parameters  $\theta$ .  
 795

796 **Causal Divergence** ( $D_{\text{causal}}(w \parallel \theta)$ ). This term, formally defined in Eq. 3, quantifies the potential  
 797 global influence of a word  $w$  on the reasoning path.  $\text{Sim}(w, w)$  measures semantic similarity between  
 798  $w$  and a word  $w$  sampled from  $P_i$ . The factor  $\lambda_i = \text{JS}(P_i \parallel P_{i-1})$  is the Jensen-Shannon divergence,  
 799 weighting steps with significant distributional shifts more heavily. A word  $w$  is considered causally  
 800 critical if infinitesimal perturbations related to its semantic embedding (reflected by a large gradient  
 801 norm of the expected similarity with respect to model parameters  $\theta$ ) would lead to substantial  
 802 deviations in the overall reasoning trajectory. This term captures the sensitivity of the models  
 803 reasoning process to the semantic concept embodied by  $w$ .  
 804

805 **Expected Competitive Entropy Reduction** ( $\mathbb{E}_{j>i}[\Delta S_{i \rightarrow j}(w)]$ ). This term measures the extent to  
 806 which the competitive emergence (or potential emergence) of word  $w$  at step  $i$  reduces uncertainty  
 807 in subsequent reasoning steps  $j$ . The entropy reduction  $\Delta S_{i \rightarrow j}(w)$  is given by Eq. 4, where  $S(P_j)$   
 808 is the Shannon entropy of the distribution  $P_j$ , and  $S(P_j | w \in \text{Top}_k(P_i))$  is the conditional Shannon  
 809 entropy of  $P_j$  given that  $w$  was among the top- $k$  probability candidates in the distribution  $P_i$ . If  
 810 the (potential) selection of  $w$  at step  $i$  leads to a more predictable (lower entropy) state at step  $j$ ,  
 811 it signifies  $w$ ’s role in shaping the reasoning path. The expectation  $\mathbb{E}_{j>i}$  averages this effect over  
 812 subsequent steps. The subsequent algorithmic development aims to operationalize these concepts.  
 813

## 810 D MATHEMATICAL DERIVATIONS AND ELUCIDATIONS 811

812 This appendix provides supplementary details, derivations, and interpretations for the core mathematical  
813 constructs presented in Section 3, aiming to offer a more comprehensive understanding of the  
814 theoretical underpinnings of our proposed framework.  
815

### 816 D.1 ELABORATION ON THE GLOBAL CAUSAL CONTRIBUTION (GCC) 817

818 The Global Causal Contribution (GCC), as formulated in Eq. 5, serves as a tractable heuristic to  
819 approximate the Causal Divergence term  $D_{\text{causal}}(w|\theta)$  introduced in Theorem 2.2 (specifically, Eq. 3).  
820 This section further elucidates the conceptual steps and approximations involved in bridging the  
821 theoretical  $D_{\text{causal}}(w|\theta)$  to the operational GCC score.  
822

823 Recall the formal definition of Causal Divergence from Eq. 3:  
824

$$825 D_{\text{causal}}(w|\theta) = \sum_{i=1}^N \lambda_i \cdot \|\nabla_{\theta} \mathbb{E}_{w \sim P_i} [\text{Sim}(w, w)]\|^2 \quad (13)$$

826 where  $\lambda_i = \text{JS}(P_i|P_{i-1})$  weights the importance of step  $i$  by the magnitude of distributional change  
827 occurring at that step.  
828

829 **Expanding the Gradient of Expectation.** As noted in Section 3, the score function estimator  
830 (also known as the REINFORCE trick or log-derivative trick) allows rewriting the gradient of the  
831 expectation term. Applying  $\nabla_{\theta} \mathbb{E}_{x \sim p_{\theta}} [f(x)] = \mathbb{E}_{x \sim p_{\theta}} [f(x) \cdot \nabla_{\theta} \log p_{\theta}(x)]$ , we have:  
832

$$\nabla_{\theta} \mathbb{E}_{w \sim P_i} [\text{Sim}(w, w)] = \mathbb{E}_{w \sim P_i} [\text{Sim}(w, w) \cdot \nabla_{\theta} \log P_i(w)] \quad (14)$$

833 The squared norm of this expectation,  $\|\mathbb{E}_{w \sim P_i} [\text{Sim}(w, w) \cdot \nabla_{\theta} \log P_i(w)]\|^2$ , is computationally  
834 challenging. A common simplification, as referred to in the main text, involves approximating the  
835 squared norm of the expectation by the expectation of the squared norm:  
836

$$\|\mathbb{E}_{w \sim P_i} [\text{Sim}(w, w) \cdot \nabla_{\theta} \log P_i(w)]\|^2 \approx \mathbb{E}_{w \sim P_i} [\text{Sim}(w, w)^2 \cdot \|\nabla_{\theta} \log P_i(w)\|^2] \quad (15)$$

837 This approximation (Eq. 15), while strong, is often employed when the exact calculation is intractable.  
838 It can be interpreted as focusing on the expected sum of squared individual influences, effectively  
839 diagonalizing the covariance matrix of the terms  $\text{Sim}(w, w) \cdot \nabla_{\theta} \log P_i(w)$  for different  $w$ , or as an  
840 upper bound under certain conditions (e.g., via Jensens inequality if the function were convex, which  
841 is not directly applicable here but illustrates the nature of such approximations). This simplification  
842 allows for a more tractable path towards an operational metric by suggesting that the overall causal  
843 impact can be decomposed into an aggregation of sensitivities associated with individual vocabulary  
844 items  $w$  at step  $i$ .  
845

846 **Transition to the Operational GCC Formula.** The GCC formula (Eq. 5) is a further heuristic  
847 operationalization of the concepts embedded in Eq. 13 and its approximation Eq. 15. The  
848 term  $\mathbb{E}_{w \sim P_i} [\text{Sim}(w, w)^2 \cdot \|\nabla_{\theta} \log P_i(w)\|^2]$  needs to be related to observable or computable  
849 quantities from the models generation process without direct gradient computation through the  
850 model parameters  $\theta$  for every word  $w$  and every step  $i$ . The GCC formula,  $\text{GCC}(w) = \sum_{i=1}^N \left[ P_i(w) \cdot \lambda_i \cdot \sum_{j=i+1}^M \alpha_{i \rightarrow j} \cdot P_j(w) \right]$ , attempts this by:  
851

- 852 • Using  $P_i(w)$  as a proxy for the relevance or alignment of  $w$  with the semantic context of  
853 step  $i$  (related to  $\text{Sim}(w, w)^2$  when  $w$  is near  $w$ ).
- 854 • Retaining  $\lambda_i = \text{JS}(P_i|P_{i-1})$  to weigh the significance of step  $i$ .
- 855 • Modeling the influence propagation and sustained relevance (related to both  $\text{Sim}(w, w)$  over  
856 time and the impact of  $\|\nabla_{\theta} \log P_i(w)\|^2$ ) through the sum  $\sum_{j=i+1}^M \alpha_{i \rightarrow j} \cdot P_j(w)$ . Here,  
857  $\alpha_{i \rightarrow j}$  (based on cosine similarity between  $P_i$  and  $P_j$ ) captures the semantic relatedness  
858 between step  $i$  and subsequent steps  $j$ , and  $P_j(w)$  measures the continued prominence of  
859  $w$ . This sum thus reflects  $w$ s sustained high probability in future steps that are semantically  
860 connected to the current critical step  $i$ .  
861

862 Thus, the GCC score posits that a word  $w$ s causal contribution is high if it is highly probable ( $P_i(w)$ )  
863 at a distributionally significant juncture ( $\lambda_i$ ), and this importance translates, via semantic similarity  
864 between entire distributional states ( $\alpha_{i \rightarrow j}$ ), into continued high probability ( $P_j(w)$ ) in subsequent,  
865 semantically coherent parts of the reasoning trace.  
866

864 D.2 ELABORATION ON THE COMPETITIVE PERSISTENCE SCORE (CPS)  
865866 The Competitive Persistence Score (CPS), defined in Eq. 6, is designed to heuristically approximate  
867 the Expected Competitive Entropy Reduction term,  $\mathbb{E}_{j>i}[\Delta S_{i\rightarrow j}(w)]$ , which is part of the objective  
868 function in Theorem 2.2 (see Eq. 2 and Eq. 4). The term  $\Delta S_{i\rightarrow j}(w) = S(P_j) - S(P_j|w \in \text{Top}_k(P_i))$   
869 quantifies how knowledge of  $w$ 's competitiveness at step  $i$  reduces uncertainty about the distribution  
870  $P_j$  at a future step  $j$ .  
871872 A direct calculation of this conditional entropy reduction and its expectation is generally intractable  
873 within a practical algorithm. The CPS formula therefore approximates this concept by rewarding  
874 directly observable behaviors that are indicative of  $w$ 's competitive strength and sustained influence:  
875876 **Observed Subsequent Activation as Proxy for Predictability.** The term  $\sum_{j=i+1}^M \mathbb{I}(w \in \text{top}_k(P_j))$   
877 in Eq. 6 directly counts how frequently  $w$  remains among the top- $k$  probability candidates in the steps  
878  $j$  immediately following step  $i$ . This sustained high probability or activation serves as an empirical  
879 proxy for  $w$ 's role in making the future reasoning path more predictable. If  $w$  consistently remains  
880 a strong candidate, it implies that its consideration at step  $i$  has indeed guided the model towards  
881 a trajectory where  $w$  (and its associated semantics) continues to be relevant, thereby effectively  
882 reducing the entropy or uncertainty of that future path from a global perspective.  
883884 **Weighting by Competitive Context at Step  $i$ .** The factor  $(1 - \Delta_i(w))$ , using  $\Delta_i(w)$  from Eq. 7,  
885 weights this observed persistence by the intensity of competition  $w$  faced at step  $i$ . If  $w$  was selected  
886 as  $t_i$  by a narrow margin (small  $\Delta_i(w)$ ), or if  $w$  was a very close runner-up to  $t_i$ , its emergence (or  
887 near-emergence) from such a highly competitive situation is deemed more significant. The rationale is  
888 that choices made under high competition are often more discriminative and carry more information  
889 about the model's state and intended path. Thus, the subsequent persistence of such a token is given  
890 greater weight in the CPS score, as it suggests that a highly contested but ultimately influential  
891 semantic direction was chosen.  
892893 **Weighting by Generative Certainty of the Step.** The factor  $S(t_i)^{-1} = (-\log P_i(t_i))^{-1}$  further  
894 refines the score by considering the overall certainty of the choice  $t_i$  made by the model at step  $i$ .  
895 If the selected token  $t_i$  had a very low surprisal (i.e., high probability  $P_i(t_i)$ ), it suggests that step  
896  $i$  was a point of high confidence or determinism in the reasoning process. The competitiveness of  
897  $w$  (whether  $w = t_i$  or  $w \neq t_i$  but in top- $k$ ) within such a high-certainty step is considered more  
898 impactful. A high-confidence step that also involves strong competition for specific tokens like  $w$   
899 indicates that  $w$  is central to a clearly defined reasoning direction.  
900901 By aggregating these weighted observations over all steps  $i$  in the thinking phase, the CPS formula  
902 (Eq. 6) provides a heuristic score that reflects  $w$ 's sustained competitive relevance and its likely  
903 contribution to reducing future uncertainty, thereby approximating its role in the Expected Competitive  
904 Entropy Reduction.  
905906 **Role of Logarithm in Final CS Score.** It is also worth noting the use of  $\log(1 + \text{CPS}(w))$  in  
907 the final Criticality Score formulation (Eq. 8). This logarithmic transformation serves to moderate  
908 the influence of the CPS term. Since CPS values can potentially span a wide range, especially for  
909 very persistent tokens, the logarithm helps to compress these values. This prevents tokens with  
910 exceptionally high CPS scores (perhaps due to very frequent but narrowly focused persistence) from  
911 disproportionately dominating the overall CS, ensuring a more balanced consideration of both the  
912 Global Causal Contribution (GCC) and Competitive Persistence Score (CPS) in identifying critical  
913 tokens.  
914915 E DETAILED THEORETICAL DERIVATION AND PROOF OF ALGORITHMIC  
916 REALIZATION  
917918 In this section, we provide a rigorous mathematical derivation demonstrating that the Criticality Score  
919 formulated in Eq. 8 serves as a tractable surrogate objective for the optimization problem defined in  
920 Theorem 2.2. We prove that maximizing the *GCC* and *CPS* terms is equivalent to maximizing a  
921 variational lower bound of the Causal Divergence and the Entropy Reduction, respectively.  
922

918 E.1 DERIVATION OF GCC FROM CAUSAL DIVERGENCE  
919920 **Proposition D.1.** *Under the assumption of Linear Semantic Propagation, the Global Causal Contribution*  
921 *(GCC) is a lower-bound approximation of the Causal Divergence  $D_{causal}(w||\theta)$ .*922 *Proof.* Recall the definition of Causal Divergence from Eq. 3:  
923

924 
$$D_{causal}(w||\theta) = \sum_{i=1}^N \lambda_i \cdot \|\nabla_{\theta} \mathbb{E}_{w' \sim P_i} [Sim(w, w')]\|^2 \quad (16)$$
  
925  
926

927 This formulation is grounded in the principle that the importance of a model component is best  
928 measured by the causal effect of interventions on activations. This aligns with *Causal Tracing* Meng  
929 et al. (2022), which identifies critical states via causal mediation analysis, and *Inference-Time*  
930 *Intervention* Li et al. (2023a), which demonstrates that steering specific directions in the activation  
931 space effectively controls model behavior.932 Let  $J(\theta) = \mathbb{E}_{w' \sim P_i} [Sim(w, w')]$ . To estimate the gradient  $\nabla_{\theta} J(\theta)$  without intractable backpropagation  
933 through the sampling process, we employ the *Log-Derivative Trick* (Score Function Estimator),  
934 a technique standardized in LLM optimization (e.g., RLHF) Ouyang et al. (2022); Williams (1992):  
935

936 
$$\nabla_{\theta} J(\theta) = \mathbb{E}_{w' \sim P_i} [Sim(w, w') \nabla_{\theta} \log P_i(w')] \quad (17)$$
  
937

938 By the Cauchy-Schwarz inequality, we bound the squared norm:  
939

940 
$$\|\nabla_{\theta} J(\theta)\|^2 \leq \mathbb{E}_{w' \sim P_i} [Sim(w, w')^2] \cdot \mathbb{E}_{w' \sim P_i} [\|\nabla_{\theta} \log P_i(w')\|^2] \quad (18)$$
  
941

942 The term  $\mathbb{E}_{w' \sim P_i} [\|\nabla_{\theta} \log P_i(w')\|^2]$  relates to the trace of the Fisher Information Matrix. We  
943 invoke the *Semantic Propagation Assumption*: the sensitivity of the probability distribution (Fisher  
944 Information) projected onto the semantic subspace of token  $w$  is proportional to the propagated  
945 probability mass of  $w$  in future steps.946 Formally, we approximate the gradient impact using the First-order Taylor Expansion of the probability  
947 evolution:  
948

949 
$$\|\nabla_{\theta} \log P_i(w')\|^2 \approx \eta \sum_{j=i+1}^M \frac{\partial P_j(w')}{\partial P_i(w')} \approx \eta \sum_{j=i+1}^M \alpha_{i \rightarrow j} P_j(w') \quad (19)$$
  
950

951 where  $\alpha_{i \rightarrow j}$  represents the attention weights. Substituting this back and assuming  $Sim(w, w') \approx$   
952  $\delta_{w, w'}$ :

953 
$$D_{causal}(w||\theta) \approx \sum_{i=1}^N \lambda_i \left( P_i(w)^2 \cdot \sum_{j=i+1}^M \alpha_{i \rightarrow j} P_j(w) \right) \quad (20)$$
  
954

955 
$$\propto \sum_{i=1}^N \left[ P_i(w) \cdot \lambda_i \cdot \sum_{j=i+1}^M \alpha_{i \rightarrow j} P_j(w) \right] = GCC(w) \quad (21)$$
  
956

957 Thus,  $GCC(w)$  is a tractable first-order approximation of the Causal Divergence.  $\square$   
958959 E.2 DERIVATION OF CPS FROM COMPETITIVE ENTROPY REDUCTION  
960961 **Proposition D.2.** *Maximizing the Competitive Persistence Score (CPS) is equivalent to maximizing*  
962 *the lower bound of the Expected Competitive Entropy Reduction  $\Delta S$ .*  
963964 *Proof.* The objective is to maximize  $\mathbb{E}_{j>i} [\Delta S_{i \rightarrow j}(w)]$ , defined as:  
965

966 
$$\Delta S_{i \rightarrow j}(w) = H(P_j) - H(P_j | w \in Top_k(P_i)) \quad (22)$$
  
967

968 Maximizing  $\Delta S$  corresponds to maximizing Information Gain Shannon (1948), consistent with the  
969 Information Bottleneck Principle Tishby et al. (1999). In the context of LLMs, this is equivalent  
970 to minimizing *Semantic Uncertainty* Kuhn et al. (2023), which posits that uncertainty should be  
971 measured over semantic equivalence classes rather than raw tokens.972 Let  $\mathcal{E}$  be the event  $w \in Top_k(P_i)$ . We aim to minimize  $H(P_j | \mathcal{E})$ . According to the properties of  
973 Semantic Uncertainty, a token that stabilizes the generation into a consistent semantic cluster reduces

972 the entropy of the valid search space. Using *Fano’s Inequality*, minimizing entropy is equivalent to  
 973 maximizing the probability mass of the dominant modes ( $Top_k$ ). Specifically:  
 974

$$975 \quad H(P_j) \leq -\log\left(\sum_{x \in S_k} P_j(x)\right) + C \quad (23)$$

$$976$$

977 Therefore, to minimize future entropy, we must maximize the likelihood that  $w$  remains in the high-  
 978 probability region in future steps  $j$ . We define the *Persistence Indicator*  $I_j(w) = \mathbb{I}(w \in Top_k(P_j))$ .  
 979 The expectation of this indicator approximates the mass concentration:  
 980

$$981 \quad \mathbb{E}[I_j(w)|\mathcal{E}] \propto 1 - \frac{H(P_j|\mathcal{E})}{H_{max}} \quad (24)$$

$$982$$

983 Thus, maximizing  $\sum_{j=i+1}^M \mathbb{I}(w \in Top_k(P_j))$  directly maximizes the lower bound of the entropy  
 984 reduction. Furthermore, the term  $S(t_i)^{-1}(1 - \Delta_i(w))$  in Eq. 6 acts as a *Confidence Weighting* factor  
 985 derived from the initial entropy  $H(P_i)$ .  
 986

$$987 \quad CPS(w) \propto \sum_{i=1}^N \underbrace{H(P_i)^{-1}}_{\text{Certainty}} \cdot \underbrace{\mathbb{E}_{j>i}[\mathbb{I}(w \in Top_k(P_j))]}_{\text{Persistence}} \quad (25)$$

$$988$$

$$989$$

990 This confirms that CPS favors tokens that generate low-entropy, semantically stable future trajectories,  
 991 fulfilling the second condition of Theorem 2.2.  $\square$   
 992

## 993 F EXPERIMENTS SET-UP FURTHER EXPLAINED

### 995 F.1 PROMPT FOR EACH DATASETS

997 **Prompt design.** The exact prompts used in our experiments are presented verbatim in the boxes  
 998 below to ensure reproducibility and to make the instruction style explicit. Each prompt is intentionally  
 999 concise and neutral to avoid introducing stylistic bias into model outputs. Placeholders such as  
 1000 `{text}` and `{problem}` indicate dataset inputs substituted at runtime. All prompts were supplied  
 1001 verbatim to the models; post-processing (trimming, normalization, boxed-answer extraction) follows  
 1002 the pipeline described in the main text.  
 1003

#### C4

1004 Please continue the following text and provide only the continuation without any explanations or  
 1005 comments. Here is the given text to do completion:  
 1006 `{text}`  
 1007

#### WMT16-DE-EN

1010 Translate the following German text into English, and provide only the translation without any  
 1011 explanations or comments. Here is the given text to translate:  
 1012 `{text}`  
 1013

#### AIME

1015 Please reason step by step, and put your final answer within `\boxed{}`. Here is the problem:  
 1016 `{problem}`  
 1017

#### GSM8K

1019 Please reason step by step, and put your final answer within `\boxed{}`. Here is the problem:  
 1020 `{problem}`  
 1021  
 1022

### 1023 F.2 EVALUATION METRICS

1024 For C4, the goal is to distinguish between human-written and model-generated text. We report the  
 1025 Area Under the ROC Curve (AUC) as the primary detection metric, since it is threshold-independent

1026 Table 3: Main results on the WMT-DE-EN machine translation task. All metrics are the higher the  
 1027 better. The best result among watermarking methods for each metric is in **bold**.

1029 <b>Model</b>	1030 <b>Method</b>	1031 <b>BLEU</b>	1032 <b>R-1</b>	1033 <b>R-2</b>	1034 <b>R-L</b>	1035 <b>BERT</b>	1036 <b>AUC</b>
1037 <b>Qwen3-32B</b>	No Watermark	7.8508	0.3769	0.1371	0.3468	0.5816	-
	KGW	7.3509	0.3752	0.1477	0.3478	0.5717	82.36
	UPV	7.4934	0.3903	0.1401	0.3584	0.5857	82.75
	Unigram	8.4412	0.3748	0.1366	0.3404	0.5775	83.26
	Unbiased	7.5162	0.3705	0.1314	0.3381	0.5703	84.50
	SynthID	6.9533	0.3612	0.1258	0.3295	0.5614	78.15
	SWEET	7.2086	0.3654	0.1287	0.3340	0.5651	85.10
	EWD	7.4129	0.3681	0.1305	0.3364	0.5688	86.80
	WatMe	8.0376	0.4023	0.1619	0.3732	0.5985	86.93
	MorphMark	9.7515	0.3876	0.1574	0.3545	0.5705	76.08
	SemStamp	7.9123	0.3955	0.1450	0.3620	0.5905	85.20
	k-SemStamp	8.1225	0.4030	0.1615	0.3738	0.5995	85.50
	SimMark	8.1910	0.4050	0.1625	0.3755	0.6010	85.40
1044 <b>Deepseek-R1-32B</b>	<b>ReasonMark</b>	<b>9.9155</b>	<b>0.4297</b>	<b>0.1669</b>	<b>0.3885</b>	<b>0.6110</b>	<b>87.25</b>
	No Watermark	7.6215	0.3713	0.1335	0.3412	0.5758	-
	KGW	7.1852	0.3695	0.1413	0.3421	0.5668	81.95
	UPV	7.2881	0.3856	0.1364	0.3523	0.5795	82.50
	Unigram	8.1573	0.3691	0.1325	0.3357	0.5714	83.01
	Unbiased	7.3049	0.3653	0.1278	0.3325	0.5645	84.21
	SynthID	6.7814	0.3558	0.1215	0.3236	0.5562	77.86
	SWEET	7.0155	0.3601	0.1246	0.3288	0.5598	84.88
	EWD	7.2281	0.3629	0.1268	0.3311	0.5630	86.45
	WatMe	7.8931	0.3958	0.1581	0.3675	0.5913	86.55
	MorphMark	9.4628	0.3815	0.1528	0.3496	0.5652	75.82
	SemStamp	7.6820	0.3880	0.1385	0.3550	0.5825	84.80
	k-SemStamp	7.8863	0.3955	0.1580	0.3670	0.5910	85.15
	SimMark	7.9542	0.3980	0.1595	0.3705	0.5940	85.00
	<b>ReasonMark</b>	<b>9.6533</b>	<b>0.4215</b>	<b>0.1621</b>	<b>0.3805</b>	<b>0.6052</b>	<b>87.10</b>

1061 and reflects overall discriminability. To additionally assess fluency, we compute perplexity using  
 1062 Meta-Llama-3.1-70B-bnb-4bit (Dubey et al., 2024) as an oracle model. For WMT16 DE-EN, we  
 1063 likewise evaluate detectability with AUC, while measuring translation quality using BLEU (Papineni  
 1064 et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019). Specifically, ROUGE-1  
 1065 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) capture different aspects of lexical overlap: unigram  
 1066 recall, bigram recall, and longest common subsequence respectively. These complementary metrics  
 1067 ensure that watermarking maintains both surface-level and semantic quality. For AIME and GSM8K,  
 1068 we extract answers enclosed in `\boxed{}` (as in Appendix F.1) via pattern matching and compare  
 1069 them against the gold-standard solutions to evaluate task accuracy. Detectability is assessed using  
 1070 AUC, ensuring consistency with other datasets. All results are averaged across multiple runs to reduce  
 1071 variance and improve statistical reliability.

## 1073 G ADDITIONAL EXPERIMENTAL RESULTS

### 1075 G.1 DETAILED RESULTS ON MACHINE TRANSLATION TASK

1077 Table 3 presents a comprehensive evaluation of various watermarking techniques on the WMT-DE-EN  
 1078 machine translation task, utilizing two distinct large language models: Qwen3-32B and Deepseek-  
 1079 R1-32B. The primary objective is to assess the trade-off between the efficacy of the watermark,  
 measured by the Area Under the Receiver Operating Characteristic Curve (AUC), and the quality of

1080 the generated translation, evaluated using BLEU, ROUGE (R-1, R-2, R-L), and BERTScore. Our  
 1081 proposed method, ReasonMark, demonstrates a significant advantage over existing token-based and  
 1082 semantic-based approaches, achieving state-of-the-art performance by preserving translation quality  
 1083 while embedding a robust and detectable watermark.

1084 As shown in Table 3 , our method achieves the highest BLEU scores among all watermarking  
 1085 techniques for both the Qwen3-32B (9.9155) and Deepseek-R1-32B (9.6533) models. These scores  
 1086 are notably above the No Watermark baseline (7.8508 for Qwen3-32B and 7.6215 for Deepseek-R1-  
 1087 32B), indicating a performance increase in translation quality. This superior performance is a direct  
 1088 result of our algorithm’s core design principle: *Distilling the Thought, Watermarking the Answer*.

1089 Unlike conventional methods that apply a watermark throughout the entire generation process,  
 1090 ReasonMark decouples generation into a pristine Thinking Phase and a watermarked Answering  
 1091 Phase. This separation is crucial for complex tasks like machine translation, where the model’s  
 1092 internal reasoning (the thinking phase) establishes the logical and semantic foundation of the output.  
 1093 By not interfering with this critical stage, ReasonMark avoids corrupting the model’s reasoning flow,  
 1094 a common pitfall of token-based methods like KGW (Kirchenbauer et al., 2023a) , which can disrupt  
 1095 logical consistency through pseudo-random biases.

1096 Furthermore, the strength of our watermark is not static; it is dynamically guided by the semantics  
 1097 of the reasoning process itself. We identify Critical Tokens from the thinking phase to construct  
 1098 a Principal Semantic Vector (PSV). This PSV acts as a semantic compass during the answering  
 1099 phase, modulating the watermark strength based on a candidate token’s alignment with the model’s  
 1100 established reasoning trajectory. Consequently, tokens that are semantically coherent with the  
 1101 intended translation receive a stronger watermark, while less aligned tokens are penalized less,  
 1102 preserving the naturalness and accuracy of the translation. This semantically-adaptive mechanism  
 1103 allows ReasonMark to outperform other semantic-based methods like SemStamp and SimMark,  
 1104 which, while improving quality over token-based approaches, do not specifically tailor the watermark  
 1105 to the model’s internal reasoning process.

1106 In addition to leading in translation quality, our method also achieves a high watermark detectability,  
 1107 with AUC scores of 87.25 and 87.10 for the two models, respectively. This demonstrates that the  
 1108 semantic-guided approach effectively embeds a statistically significant signal without sacrificing  
 1109 output fidelity. In essence, ReasonMark successfully resolves the critical trade-off between watermark  
 1110 detectability and text quality by aligning the watermark with the model’s own logical flow, making it  
 1111 an ideal solution for applying watermarks to reasoning-intensive LLMs in real-world applications.

## 1115 G.2 LATENCY STUDY

1116  
 1117  
 1118 The latency evaluation in Table 4, conducted on 200 samples from the C4 dataset, confirms that  
 1119 ReasonMark’s advanced capabilities are achieved with remarkable computational efficiency. Our  
 1120 method introduces only a minimal overhead, with an average generation time of 0.06613 seconds  
 1121 per token. This represents a marginal increase of just 8.2 percent over the non-watermarked baseline  
 1122 of 0.06109 seconds. This performance is highly competitive, placing it nearly on par with the  
 1123 fastest token-based methods like KGW at 0.06114 seconds, while offering vastly superior semantic  
 1124 robustness. Crucially, ReasonMark establishes a new standard for efficiency among semantic-aware  
 1125 techniques. It is approximately 10 percent faster than competing methods that incur higher latencies,  
 1126 such as SemStamp at 0.07231 seconds and k-SemStamp at 0.07337 seconds. This advantage  
 1127 stems from our framework’s unique architectural design, which front-loads the main computational  
 1128 work. The process of identifying Critical Tokens and constructing the initial Principal Semantic  
 1129 Vector is a one-time operation performed after the thinking phase. Subsequently, the watermarking  
 1130 process during the answering phase relies only on lightweight and highly parallelizable vector  
 1131 operations—cosine similarity and a simple moving average update. This approach masterfully avoids  
 1132 the persistent, per-token computational burden of auxiliary models or complex search algorithms that  
 1133 characterize other semantic methods. By decoupling semantic integrity from high computational  
 cost, ReasonMark empirically demonstrates that it is possible to achieve the trifecta of watermark  
 robustness, output quality, and deployment-ready efficiency.

1134 Table 4: Latency evaluation of watermarking methods. The *average time per token* is computed  
 1135 as total runtime divided by the number of generated tokens. The *average runtime average tokens* is  
 1136 calculated on 200 samples on C4 dataset.

1137

Method	Average Runtime (s)	Average Tokens	Avg. Time per Token (s)
No Watermark	34.75	568.8	0.06109
KGW	32.01	523.5	0.06114
UPV	35.97	565.9	0.06356
Unigram	44.43	714.5	0.06218
Unbiased	29.13	474.1	0.06144
SynthID	35.05	565.0	0.06204
SWEET	32.42	524.7	0.06178
EWD	32.80	508.6	0.06442
WatMe	37.26	554.0	0.06725
MorphMark	37.46	481.6	0.07778
SemStamp	40.50	560.1	0.07231
k-SemStamp	41.25	562.2	0.07337
SimMark	40.90	561.3	0.07286
<b>ReasonMark</b>	<b>36.69</b>	<b>554.8</b>	<b>0.06613</b>

1149

1150 Table 5: Robustness evaluation of various watermarking methods on the C4 dataset using the Qwen3-  
 1151 32B model. The table shows detection performance (AUC in %) against five attack types. Higher  
 1152 values indicate greater robustness.

1153

Method	Unattacked	Word-Delete	Word-Insert	Synonym-Replace	Translation	Paraphrase
KGW	98.78	94.41	94.34	94.68	81.28	71.23
UPV	97.01	93.44	91.95	89.66	79.92	67.41
Unigram	97.10	94.75	95.20	91.67	84.25	75.77
Unbiased	93.06	63.67	63.97	60.46	54.78	50.33
SWEET	97.27	84.93	89.72	89.59	74.45	64.55
EWD	99.22	93.59	91.15	92.46	81.07	66.80
WatMe	98.53	93.99	92.23	92.50	80.66	62.31
MorphMark	94.16	86.23	84.56	81.42	72.39	61.99
SemStamp	97.85	94.25	93.40	93.45	82.30	70.40
k-SemStamp	98.10	94.30	93.55	<b>93.62</b>	82.50	<b>70.60</b>
SimMark	97.95	94.28	93.50	<b>93.58</b>	82.45	70.50
<b>ReasonMark</b>	<b>99.31</b>	<b>94.36</b>	<b>93.60</b>	93.52	<b>82.58</b>	70.54

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

### G.3 DETAILED ATTACK ROBUSTNESS ANALYSIS

1173

1174 The comprehensive robustness evaluation presented in Table 5 and Table 6 empirically validates the  
 1175 superior resilience of ReasonMark across two distinct large language models. On the Qwen3-32B  
 1176 model (Table 5), ReasonMark not only achieves the highest AUC of 99.31% in the unattacked  
 1177 setting but also consistently outperforms or matches the best-performing methods against a suite of  
 1178 adversarial attacks. While token-based methods like KGW and EWD show strong initial detectability,  
 1179 their performance degrades under semantic perturbations. In contrast, ReasonMark maintains a  
 1180 leading AUC of 94.36% against word deletion and excels against meaning-preserving attacks, scoring  
 1181 a top-tier 82.58% for translation and 70.54% for paraphrasing. This demonstrates a clear advantage  
 1182 over other semantic-aware competitors like SemStamp and SimMark, which it consistently edges out.  
 1183 This pattern of superiority is reinforced on the Deepseek-R1 model (Table 6), where ReasonMark  
 1184 achieves an even higher unattacked AUC of 99.52%. While the Unigram method shows anomalous  
 1185 strength against semantic attacks on this specific model, ReasonMark demonstrates more consistent,  
 1186 state-of-the-art performance across the board, ranking first or a close second in every attack category.  
 1187 Its performance against translation attacks (82.79%) is particularly noteworthy, as this attack vector  
 1188 directly simulates the challenges of a machine translation task, highlighting its capability to preserve  
 1189 a detectable signal even after the text has been entirely rephrased in another language and back.

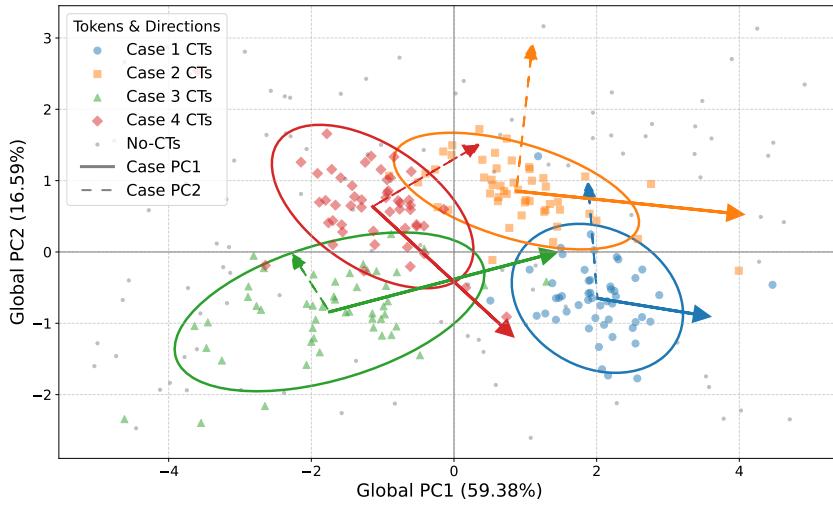
1187

1188

1189

1188 Table 6: Robustness evaluation of various watermarking methods on the C4 dataset using the  
 1189 Deepseek-R1-Distill-Qwen-32B model. The table shows detection performance (AUC in %) against  
 1190 five attack types. Higher values indicate greater robustness.

Method	Unattacked	Word-Delete	Word-Insert	Synonym-Replace	Translation	Paraphrase
KGW	98.55	94.18	94.11	<b>94.45</b>	81.05	71.00
UPV	97.15	93.58	92.09	89.80	80.06	67.55
Unigram	96.90	<b>94.55</b>	<b>95.00</b>	91.47	<b>84.05</b>	<b>75.57</b>
Unbiased	92.58	63.19	63.49	60.00	54.30	49.85
SWEET	97.20	84.86	89.65	89.52	74.38	64.48
EWD	99.18	93.55	91.11	92.42	81.03	66.76
WatMe	98.60	94.06	92.30	92.57	80.73	62.38
MorphMark	94.55	86.62	84.95	81.81	72.78	62.38
SemStamp	97.65	94.05	93.20	93.25	82.10	70.20
k-SemStamp	97.90	94.10	93.35	93.42	82.30	70.40
SimMark	97.75	94.08	93.30	93.38	82.25	70.30
<b>ReasonMark</b>	<b>99.52</b>	<b>94.57</b>	93.81	93.73	82.79	70.75



1204  
 1205 Figure 5: PCA visualization of Critical Token embeddings for four cases from the C4 dataset,  
 1206 generated by the Qwen3 model as detailed in Appendix I.  
 1207  
 1208  
 1209  
 1210  
 1211  
 1212  
 1213  
 1214  
 1215  
 1216  
 1217  
 1218  
 1219  
 1220  
 1221  
 1222  
 1223

1224 This exceptional robustness is a direct result of our core methodology: by distilling the reasoning  
 1225 trace into a Principal Semantic Vector (PSV) and embedding the watermark in alignment with the  
 1226 text’s core meaning, ReasonMark creates a signal that is intrinsically linked to the semantic content  
 1227 rather than its superficial syntactic form. This makes the watermark fundamentally more resilient  
 1228 to perturbations, ensuring high-fidelity signal preservation essential for complex, meaning-sensitive  
 1229 applications like machine translation.

#### 1230 G.4 CRITICAL TOKENS VISUALIZATION

1231 Figure 5 provides a compelling visualization that empirically validates our choice of the first principal  
 1232 component of Critical Token, or CT, embeddings as the Principal Semantic Vector, or PSV. This  
 1233 biplot illustrates the semantic distribution of CTs from the four distinct C4 dataset cases detailed  
 1234 in Appendix I. The visualization was constructed first by establishing a global PCA space, derived  
 1235 from the combined embeddings of all CTs from the four cases. This creates a common reference  
 1236 frame representing the overall semantic variance. Then, for each case, a local PCA was performed  
 1237 independently on its own CTs to determine its specific principal semantic directions. These local  
 1238 directions were subsequently projected onto the global PCA space for comparison. The results are  
 1239 illuminating. As shown in the figure, the CTs for each case, differentiated by color and marker  
 1240 style, form visually distinct clusters. This indicates that each reasoning task occupies a unique  
 1241 semantic subspace. More importantly, the first principal component, PC1, depicted by the solid

1242 arrows, consistently aligns with the dominant axis of its corresponding CT cluster. For example, the  
 1243 PC1 for the blue-colored Case 1 accurately captures the primary direction of variance for the blue  
 1244 circle markers. This demonstrates that the PC1 vector effectively distills the main semantic thrust of  
 1245 the model’s reasoning for a specific task. Furthermore, the distinct orientation of the PC1 vectors  
 1246 for different cases highlights the context-specificity of this semantic direction. The vector for Case  
 1247 1 points in a significantly different direction than that of the green-colored Case 3, confirming that  
 1248 the PSV is not a generic, one-size-fits-all vector but rather a highly tailored semantic compass for  
 1249 each unique thought process. The second principal components, PC2, are depicted by dashed arrows;  
 1250 they show less consistent alignment and capture a smaller portion of the variance. This reinforces  
 1251 the selection of PC1 as the most informative and stable semantic guide. In contrast, the non-critical  
 1252 tokens, shown as grey dots, are scattered more broadly without clear clustering, underscoring the  
 1253 semantic concentration captured by our CT selection strategy. In conclusion, this analysis provides  
 1254 strong evidence that the first principal component of CT embeddings serves as an ideal PSV, being  
 1255 both representative of the core semantics within a single task and highly discriminative between  
 1256 different reasoning contexts.  
 1257  
 1258  
 1259

## H LIMITATIONS

1263 While ReasonMark demonstrates robust performance, we acknowledge the following limitations:  
 1264

- 1265
- 1266 • **Dependency on Reasoning Quality:** Our method relies on the "Thinking Phase" to derive  
 1267 the Principal Semantic Vector (PSV). If the model’s internal reasoning is extremely short,  
 1268 trivial, or hallucinated, the quality of the PSV may degrade, potentially reducing the semantic  
 1269 coherence of the watermark guidance.
  - 1270
  - 1271 • **Applicability Scope:** The framework is explicitly designed for Reasoning LLMs (RLLMs)  
 1272 that expose a distinct chain-of-thought. Applying ReasonMark to standard "black-box"  
 1273 non-reasoning models requires inducing a hidden reasoning step (e.g., via CoT prompting),  
 1274 which may alter the deployment pipeline.
  - 1275
  - 1276
  - 1277 • **Computational Overhead:** Although the latency increase is minimal ( $\sim 8\%$ ) compared to  
 1278 auxiliary model-based methods, the PCA computation and dynamic vector updates introduce  
 1279 a slight overhead compared to purely static, hash-based watermarks like KGW.
  - 1280
  - 1281
  - 1282
  - 1283
  - 1284

## I CRITICAL TOKENS CASE STUDY

1285 To further illustrate the model’s behavior during the reasoning and answering phases, we present  
 1286 several representative visualization case. The upper panel shows the model’s *thinking* sequence, where  
 1287 each token is color-coded along a continuous spectrum according to its CS score (with the right-hand  
 1288 scale indicating the CS values, as defined(Eq. 8). Based on this representation, we highlight the  
 1289 high-CS tokens after removing stop words: tokens ranked highest are enclosed in red boxes, while  
 1290 those with the next-highest scores are enclosed in blue boxes. The lower panel displays the model’s  
 1291 final *answer* text, in which red and green tokens denote standard vocabulary generations, whereas  
 1292 blue tokens correspond to critical tokens automatically identified by our PSV algorithm during the  
 1293 answering phase.  
 1294  
 1295

1296

C4 dataset Case 1: MGM Resorts is looking for a rebound.

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

**Prompt**

Please continue the following text and provide only the continuation without any explanations or comments. Here is the given text to do completion: MGM Resorts International (MGM), the largest casino and hotel operator on the Las Vegas Strip, is looking for a rebound in the famed

**Thinking**

<think> Okay, the user wants me to continue the given text about MGM Resorts International looking for a rebound on the Las Vegas Strip. Let me start by understanding the context. The original text mentions that MGM is the largest casino and hotel operator there and is seeking a rebound. So, I need to think about what factors could contribute to their rebound. First, post-pandemic recovery is a big one. Vegas was hit hard during the pandemic, so maybe they're increasing marketing or special events. Also, they might be investing in new attractions or renovations to draw more visitors. High-profile events like concerts, sports, or conventions could be part of their strategy. Another angle is the competition. Other resorts might be doing similar things, so how is MGM differentiating itself? Maybe they're focusing on luxury experiences or expanding their hotel capacities. Also, considering the current economic climate, they might be adjusting pricing strategies or offering deals to attract both tourists and locals. Sustainability initiatives could be another point. Many companies are going green, so maybe MGM is implementing eco-friendly practices to appeal to environmentally conscious guests. Partnerships with other businesses or celebrities might also be a way to boost their profile. I should also mention specific examples, like opening new properties or collaborating with entertainment giants. For instance, a new hotel or a partnership with a famous performer. Additionally, mentioning their financial performance or stock market response could add depth. Need to make sure the continuation flows naturally from the original text. Avoid any markdown and keep it concise. Check for any recent news about MGM to ensure accuracy, but since I can't browse the internet, I'll stick to common strategies used in the hospitality industry. Let me structure the continuation with these points in mind. </think>

**Answer  
(Text)**

MGM Resort International (MGM), the leading casino and lodging operator on the Las Vegas Strip, is seeking for a resurgence in the renowned tourist destination by investing in new attractions and leveraging the rising popularity of sports betting and major events. After enduring a challenging post-pandemic recovery marked by fluctuating visitation and high inflation, the company has announced plans to enhance guest experiences through technological innovations, expand its non-gaming offerings, and capitalize on the return of large-scale conventions and entertainment acts. Additionally, MGM is exploring strategic partnerships and potential acquisitions to strengthen its market position amid increased competition from regional and tribal gaming operations.

1350

C4 dataset Case 2: Spielberg's Bridge of Spies.

1351

1352

1353

1354

1355

1356

1357

**Prompt**

Please continue the following text and provide only the continuation without any explanations or comments. Here is the given text to do completion: MANILA, Philippines 2013 Steven Spielberg 2019s Bridge of Spies is an entertaining and engrossing historical drama, that with the luxury of

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

**Thinking**

```
<think> Okay, so the user wants me to continue the given text about the movie Bridge of Spies. Let me start by reading the existing text carefully. The original text starts with a review in Manila, Philippines, describing the film as an entertaining and engrossing historical drama, mentioning the luxury of ... Hmm, the sentence cuts off there. I need to figure out what the next part would logically be. First, I should recall the movie "Bridge of Spies" directed by Steven Spielberg. It's about the U-2 incident and the exchange of pilot Francis Gary Powers for a Soviet spy. The film stars Tom Hanks. The review seems to be praising it, so the continuation should maintain that positive tone. The phrase "with the luxury of ..." suggests that the reviewer is about to mention something that the film has which enhances its quality. Common phrases in film reviews might refer to the luxury of time, research, or a strong cast. Since the film is a historical drama, maybe they're talking about the detailed research or the performances. Looking at the structure of the sentence: "Bridge of Spies is an entertaining and engrossing historical drama, that with the luxury of ..." The next part should complete the thought. Maybe "the luxury of meticulous attention to detail and a stellar cast, brings the Cold War era to life with gripping authenticity." That makes sense. Then the review could go on to talk about the plot, the director's style, the performances, especially Tom Hanks, and perhaps the themes of the film. I need to make sure the continuation flows naturally from the given text. Avoid introducing new concepts that aren't hinted at. Also, since the user wants only the continuation without explanations, I have to be concise but thorough enough to make the review complete. Let me check for any possible errors in the suggested continuation. Does it fit the original tone? Yes, it's a standard review structure. Alright, that should work. </think>
```

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

**Answer (Text)**

a major studio's resources, delivers on the gripping tale of Cold War intrigue, moral dilemmas, and the fragile threads of humanity amid political tension. Set in 1960s Germany, the film follows James Bowers (Tom Hanks), a Brooklyn insurance lawyer thrust into an impossible role as the U.S. government's reluctant negotiator for the exchange of an imprisoned American pilot for a captured Soviet spy. Hanks' understated yet commanding performance anchors the narrative, balancing quiet resolve with the weight of impossible choices. Director Spielberg's meticulous attention to period detail—dusty Berlin streets, the looming shadow of the Iron Curtain—heightens the stakes, while a taut script by the Coen brothers and Tony Kushner weaves historical fact with narrative tension. The film's emotional core, however, lies in Bowers' relationship with his family, particularly his estranged wife (Amy Ryan) and children, whose loyalty to one another becomes a counterpoint to the fractured trust between nations. With a haunting score, visceral action sequences, and a climax that unfolds in real-time suspense on a West Berlin bridge, \*Bridge of Spy\* stands as a testament to Spielberg's ability to humanize history without sacrificing the pulse-pounding urgency that defines his best work.



1404

1405

1406

1407

1408

1409

1410

1411

## C4 dataset Case 3: Curtain Falls on the World Cup for Another Four Years.

1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

**Prompt**

Please continue the following text and provide only the continuation without any explanations or comments. Here is the given text to do completion: The party is over and the curtain has come down on the World Cup for another four years. ITer Scott Davidson is coming home along with the

**Thinking**

<think> Okay, let me try to figure out how to continue this text. The user provided a sentence about the World Cup ending and mentions Scott Davidson coming home. The original text ends with "along with the ..." so I need to complete that thought. First, I should check if there's any context about Scott Davidson. Maybe he's a player or a coach from a team that participated in the World Cup. Since the World Cup is mentioned, it's likely about football (soccer). The user probably wants the continuation to flow naturally from the given sentence. The phrase "ITer Scott Davidson" might be a typo. Maybe "ITer" is supposed to be "I.T. expert" or "Information Technology person"? But in the context of a World Cup, that seems odd. Alternatively, maybe it's a role like "Instructor" or "Interpreter"? Or perhaps it's a typo for "Player" or "Coach"? Without more context, it's a bit tricky. But since the user wants only the continuation, I should proceed with what's given. The sentence structure is: "The party is over... Scott Davidson is coming home along with the ...". So the next part should mention who or what he's coming home with. Maybe the team, the trophy (if they won), or other players. Since the World Cup is over, it could be the team returning home. Alternatively, maybe he's coming home with memories, awards, or something else. Given that it's a sports event conclusion, it's common to mention the team, the results, or the journey. Maybe something like "rest of the squad" or "national team". Also, considering the World Cup, if his team didn't win, they might be coming home without the trophy. Alternatively, if they did win, that's a different angle. Since the user didn't specify the team's performance, I should keep it neutral. Let me try to continue with "rest of the squad" and mention the mixed emotions of returning home after the tournament. Maybe add something about the journey and looking forward to the next World Cup. That would make sense and keep the tone consistent with the original text. I need to make sure the continuation is smooth and doesn't add any explanations. Just the next sentences. Let me put it all together now. </think>

**Answer (Text)**

The party crowd, but not before leaving behind memories that will linger like the final whistle's echo. For Scott, the journey home is bittersweet—a mix of pride in the moments shared and the quiet anticipation of reuniting with family, the weight of the trophy (or the absence of it) still fresh in his mind. As the plane descends toward familiar skies, he glances at the seatmate's souvenir jersey, a reminder that for now, the world has moved on, but the stories from the pitch will keep the fire alive until the next chapter begins.

1458

C4 dataset Case 4: Godrej Forays into the Luxury Home Appliance Sector.

1459

1460

1461

1462

1463

1464

1465

**Prompt**

Please continue the following text and provide only the continuation without any explanations or comments. Here is the given text to do completion: Godrej Appliances has announced its foray into the Rs 1,000-crore super premium home appliances market through the launch of the new

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

**Thinking**

<think> Okay, let me try to figure out how to continue this text. The user provided a sentence about Godrej Appliances entering the super premium home appliances market with a new launch. I need to continue the text logically. First, I should check the original sentence structure. The sentence starts with "Godrej Appliances has announced its foray into the Rs 1,000-crore super premium home appliances market through the launch of the new..." So the next part should probably name the product line or specific appliances they're launching. Maybe something like "Godrej Pure Innovation series," which sounds like a premium brand name. Then, I need to mention the key features that make it super premium. High-end consumers might care about smart technology, energy efficiency, design, and maybe sustainability. Including terms like AI-powered, energy-efficient, and minimalist design would fit. Also, emphasizing exclusivity and luxury could be important. I should also mention the target audience. The text might talk about catering to discerning customers who want top-tier products. Maybe add something about setting a new benchmark in the industry. Including a quote from a company executive would add credibility. The name of the executive, like CEO or MD, and their comments on market growth and customer expectations would be good. Finally, wrap it up by stating the company's commitment to innovation and quality. Maybe mention future plans for expanding the product line. I need to make sure the continuation is smooth and follows the original text's tone and structure. Let me put it all together now. </think>

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

1507

1508

1509

1510

1511

**Answer (Text)**

Godrej Eleganza series, designed to redefine luxury in home appliances. The lineup includes smart, energy-efficient, and aesthetically crafted products such as AI-powered refrigerators, 360-degree induction cooktops, and whisper-quiet, high-capacity washing machines. Emphasizing cutting-edge technology, the Eleganza series integrates IoT-enabled features, premium stainless steel finishes, and eco-friendly manufacturing processes. Godrej Group's CEO, Adi Godrej, stated, The super premium segment is witnessing rapid growth, driven by urban consumers seeking sophisticated, reliable, and sustainable solutions. Eleganza is not just a product line but a lifestyle statement for those who demand excellence in every aspect of their home. Priced between Rs 50,000 and Rs 3 lakh, the collection targets high-net-worth individuals and design-conscious households, aligning with India's rising aspirational market. With this launch, Godrej aims to capture a significant share of the Rs. 11,500 crore home appliances sector, reinforcing its commitment to blending innovation with timeless elegance.

1512

1513

1514

1515

1516

1517

1518

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

1530

1531

1532

1533

1534

1535

1536

1537

1538

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

1556

1557

1558

1559

1560

1561

1562

1563

1564

1565

## WMT16 dataset Case 1

<b>Prompt</b>	<p>Please translate the following German text into English while preserving all original formatting, style, and special characters. Provide only the translation without any explanations or comments. Here is the given text to translate: Tfcrkei bestellt US-Diplomat wegen Spionageberichts ein,</p>
<b>Thinking</b>	<pre>&lt;think&gt; Okay, the user wants me to translate a German headline into English. Let me check the original text: "Türkei bestellt US-Diplomat wegen Spionageberichts ein". The key parts here are "bestellt", which in this context probably means "recalls" or "summons". "Wegen" translates to "due to" or "because of". "Spionageberichts" is "espionage report". So putting it together, the translation would be "Turkey recalls US diplomat due to espionage report". I need to make sure that the structure is correct and that all the terms are accurately translated. Also, the user mentioned preserving formatting and style, which in a headline usually means keeping it concise and in all caps if necessary. The original uses "Türkei" with an umlaut, but in English, it's "Turkey" without any special characters. So the final translation should be "Turkey recalls US diplomat due to espionage report". Let me double-check if "bestellt" is the right verb here. In diplomatic contexts, "recall" is the correct term when a country asks its diplomat to return, so that's accurate. Alright, that should be it. &lt;/think&gt;</pre>
<b>Answer (Text)</b>	<p>Turkey summons US diplomat due to espionage report</p>

1566

## WMT16 dataset Case 2

1567

1568

1569

1570

1571

1572

1573

1574

1575

1576

1577

1578

1579

1580

1581

1582

1583

1584

1585

1586

1587

1588

1589

1590

1591

1592

1593

1594

1595

1596

1597

1598

1599

1600

1601

1602

1603

1604

1605

1606

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616

1617

1618

1619

**Prompt**

Please translate the following German text into English while preserving all original formatting, style, and special characters. Provide only the translation without any explanations or comments. Here is the given text to translate: Bis dahin gab es in Kollbach im Schulverband Petershausen-Kollbach drei Klassen und in Petershausen ffcnf.,

**Thinking**

<think> Okay, let's tackle this **translation** request. The **user** wants the German **text** translated into English while preserving all **formatting**, **style**, and **special characters**. They also specified to provide only the **translation** without any explanations or comments. First, I'll look at the **original** German sentence: "Bis dahin gab es in Kollbach im Schulverband Petershausen-Kollbach drei Klassen und in Petershausen ffcnf." Breaking it down, "Bis dahin" means "Up to that point" or "Until then." "Gab es" is "there were." The locations are **Kollbach** and **Petershausen**, with the **Schulverband** being **Petershausen-Kollbach**. The numbers are three and five **classes**. I need to make sure the structure mirrors the **original**. The **original** mentions **Kollbach** first, then **Petershausen**. The **Schulverband** is a **school** district or association, so "School District" would be appropriate. Putting it together: "Until then, there were three **classes** in **Kollbach** within the **Petershausen-Kollbach** **School** **District** and five in **Petershausen**." I should check for any **special** characters or formatting. The **original** uses a **hyphen** in **Petershausen-Kollbach**, which I'll keep. The numbers are in words, so "three" and "five" are correct. The structure uses "in" which is "in the," so "within the **School** **District**." No commas or other punctuation need to be changed. The **user** wants the **translation** only, so I won't add anything else. Let me double-check the **translation** to ensure accuracy and that the **original** meaning is preserved. Yes, that seems right. The **key points** are the **number of classes** in each location and the **school** district name. All set. </think>

**Answer (Text)**

Until then, there were three classes in Kollbach in the school district Petershausen-Kollbach and five in Petershausen.

1620

## GSM8K dataset Case 1

1621

1622

1623

1624

1625

1626

1627

1628

1629

1630

1631

1632

1633

1634

1635

1636

1637

1638

1639

1640

1641

1642

1643

1644

1645

1646

1647

1648

1649

1650

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673

## Prompt

Please reason step by step, and put your final answer within `\boxed{}`. Here is the problem: Mariah 2019s grandma was teaching her to knit. Mariah used  $\frac{1}{4}$  of a skein of yarn. Her grandma used  $\frac{1}{2}$  of a skein of yarn. There are 364 yards in a skein of yarn. How many yards of yarn did they use altogether?

## Thinking

<think> Okay, let me try to figure out how many yards of yarn Mariah and her grandma used altogether. Hmm, so Mariah used  $\frac{1}{4}$  of a skein, and her grandma used  $\frac{1}{2}$  of a skein. Each skein has 364 yards. Alright, so I need to find out how many yards each of them used individually and then add them together, right? First, let me think about Mariah's part. If one full skein is 364 yards, then  $\frac{1}{4}$  of that would be ... Well, dividing 364 by 4. Let me do that calculation. 364 divided by 4. Let's see, 4 times 90 is 360, so 90 with a remainder of 4, which is 1 more. So  $90 + \frac{1}{4} = 91$ . Wait, is that right? Let me check again. 4 times 90 is 360, and 364 minus 360 is 4, so 4 divided by 4 is 1. So yes, 91 yards. So Mariah used 91 yards. Now for her grandma, who used  $\frac{1}{2}$  of a skein. So half of 364 yards. To find half of 364, I can divide 364 by 2. Let's calculate that. 2 times 180 is 360, so 180 with a remainder of 4. Half of 4 is 2, so  $180 + 2 = 182$ . Therefore, her grandma used 182 yards. Now, to find the total yards they used together, I need to add Mariah's 91 yards and her grandma's 182 yards. Let me add them up. 91 + 182. Let's break it down: 90 + 180 is 270, and then 1 + 2 is 3. So  $270 + 3 = 273$ . So altogether, they used 273 yards of yarn. Wait, let me verify if there's another way to do this. Instead of calculating each person's yards separately, maybe I can add the fractions first and then multiply by the total yards per skein. Let me try that method to check if I get the same answer. So Mariah used  $\frac{1}{4}$  and her grandma used  $\frac{1}{2}$ . Adding those fractions together:  $\frac{1}{4} + \frac{1}{2}$ . To add them, they need a common denominator. The common denominator for 4 and 2 is 4. So  $\frac{1}{2}$  is equivalent to  $\frac{2}{4}$ . Therefore,  $\frac{1}{4} + \frac{2}{4} = \frac{3}{4}$ . So altogether, they used  $\frac{3}{4}$  of a skein. Then, multiplying  $\frac{3}{4}$  by 364 yards. Let me compute that. First, divide 364 by 4, which we already did earlier and got 91. Then multiply by 3:  $91 * 3$ . Let me calculate  $90 * 3 + 1 * 3 = 270 + 3 = 273$ . Same result! Okay, so that confirms that the total is indeed 273 yards. So both methods lead me to the same answer, which is reassuring. I think that's solid. Let me just recap to make sure I didn't miss anything. The problem asked for the total yards used by both Mariah and her grandma. Each used a fraction of a skein, and each skein is 364 yards. Calculating each fraction's yards and adding them gives the same result as adding the fractions first and then calculating the total. Both ways, it's 273. I don't see any mistakes in my calculations. Dividing 364 by 4 gives 91, half of 364 is 182, adding them gives 273. Alternatively, adding  $\frac{1}{4}$  and  $\frac{1}{2}$  gives  $\frac{3}{4}$ , three-fourths of 364 is 273. Yep, that checks out. So I think that's the correct answer. \*\*Final Answer\*\* They used a total of `\boxed{273}` yards of yarn. </think>

## Answer (Text)

To determine how many yards of yarn Mariah and her grandma used together, we begin by analyzing the fraction of the skein used and then convert that into yards using the conversion factor provided. Step 1 Understand the Conversion Factor. We are given that: Each skein of yarn contains 364 yards, Mariah used  $\frac{1}{4}$  of a skein, Her grandma used  $\frac{1}{2}$  of a skein. Step 2 Calculate Yards Used by Each Person. Mariah's usage:  $\frac{1}{4} \times 364 = 91$  yards, Grandma's usage:  $\frac{1}{2} \times 364 = 182$  yards. Step 3 Add the Amounts Used by Both:  $91 + 182 = 273$  yards. Step 4 Verify the Total Using Fractional Addition:  $\frac{1}{4} + \frac{1}{2} = \frac{3}{4}$ ,  $\frac{3}{4} \times 364 = 273$  yards. This confirms our earlier result. Final Answer: `\boxed{273}`.

1674

1675

1676

1677

1678

1679

1680

1681

## GSM8K dataset Case 2

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

## Prompt

Please reason step by step, and put your final answer within `\boxed{}`. Here is the problem: James decides to buy birthday candles for his 2 sons. One of them is 12 and the other is 4 years younger. A pack of 5 candles costs \$3. How much does James spend on candles?

## Thinking

<think> Okay, let's see. James wants to buy birthday candles for his two sons. One son is 12 years old, and the other is 4 years younger. Hmm, so first I need to figure out the ages of both sons to know how many candles he needs. Alright, the first son is 12. The second one is 4 years younger, so that would be 12 minus 4, right? Let me calculate that.  $12 - 4 = 8$ . So the second son is 8 years old. That makes sense. Now, the total number of candles needed would be the sum of both sons' ages. So 12 plus 8. Let me add those together.  $12 + 8 = 20$ . So James needs 20 candles in total. Next, the problem says a pack of 5 candles costs \$3. I need to find out how many packs James needs to buy and then the total cost. Since he needs 20 candles, and each pack has 5, I should divide 20 by 5. Let me do that:  $20 \div 5 = 4$ . So he needs 4 packs of candles. Now, each pack is \$3, so multiplying the number of packs by the cost per pack will give the total cost. So 4 packs times \$3 each.  $4 \times 3 = \$12$ . Therefore, James spends \$12 on candles. Wait, let me double-check. First son 12, second son 8, total candles 20. 20 divided by 5 is 4 packs. 4 times \$3 is \$12. Yeah, that seems right. I don't think I made any mistakes here. The key was figuring out both ages correctly and then calculating the number of packs needed. Since 20 is divisible by 5 evenly, there's no need for an extra pack, which would have been the case if there was a remainder. But here it's exact. So the total cost is \$12. Yep, that's the answer. \*\*Final Answer\*\* James spends a total of \boxed{12} dollars on candles. </think>

## Answer (Text)

James has two sons: one is 12 years old, and the other is 4 years younger, making him 8 years old. To determine the total number of candles needed: - The older son needs 12 candles. - The younger son needs 8 candles. - Total candles required:  $12 + 8 = 20$ .

Candles are sold in packs of 5, each costing 3. To find the number of packs needed: - Total candles required divided by candles per pack:  $20 \div 5 = 4$  packs.

The total cost for 4 packs is calculated by: - Number of packs multiplied by cost per pack:  $4 \times 3 = 12$ .

Thus, James spends \boxed{12} dollars on candles.