

# Evaluating LLM Watermarking for Clinical EHR: Detectability, Style Shifts, and Helpfulness

Rochana Prih Hastuti, Rian Adam Rajagede, Mengxin Zheng, Qian Lou

University of Central Florida, USA

{rochana, rian, mengxin.zheng, qian.lou}@ucf.edu

## Abstract

LLM Watermarking for clinical text requires evaluating detection capability alongside task performance. While existing benchmarks focus on accuracy, we show this misses critical quality degradation in Electronic Health Record (EHR) transformation tasks, where technical notes are simplified for patient understanding. We evaluate five watermarking methods across two EHR transformation tasks spanning five clinical note types, developing a multi-aspect framework that measures accuracy, style, and readability. We adapt linguistic distance of Part-of-Speech tags ( $LD_{POS}$ ) to quantify style shifts and enable mechanistic analysis.

Results reveal a fundamental watermarking-induced helpfulness degradation and its detectability tension: EHR transformation achieves only 7.7%–24.2% detection rates under default parameters (vs. >90% in narrative tasks). Tuning for practical detection degrades helpfulness by 24.6%–32.5%. Mechanistic analysis via  $LD_{POS}$  shows watermarking systematically shifts style toward generic terms and connected structures, degrading helpfulness even when accuracy is preserved.

## 1 Introduction

LLMs are actively deployed in patient-facing clinical workflows, transforming technical clinical notes into patient-readable summaries or atomic facts (Fleming et al., 2024; Munnangi et al., 2025). Tools like ChatGPT Health now allow patients to upload Electronic Health Records (EHRs) and receive simplified summaries directly (OpenAI, 2026), making provenance of AI-generated clinical text a patient safety and regulatory imperative. Unlike visible disclaimers, watermarking embeds verifiable attribution directly into generated text, enabling post-hoc forensic detection when AI-generated content causes harm. Watermarking has thus emerged as

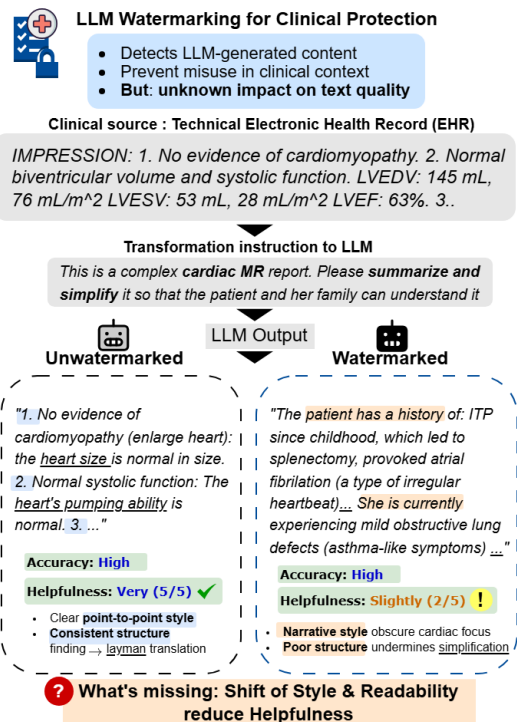


Figure 1: Watermarking hidden quality problem: Despite retained accuracy, watermarked output showed reduced helpfulness from style and structural readability shift, dimensions invisible to standard evaluation.

the primary mechanism to verify AI-generated content (Kirchenbauer et al., 2023; Xue et al., 2025), required under the EU AI Act's Draft Code of Practice (enforced August 2026) (European Commission AI, 2026), which itself acknowledges that text watermarking risks degrading output quality, making its helpfulness cost a live regulatory concern. Yet watermarking's impact on clinical text quality remains poorly understood.

Prior work shows watermarking degrades accuracy in medical narrative tasks (Hastuti et al., 2025; Kong et al., 2025). Yet accuracy alone may not capture quality degradation in EHR transformation tasks. As Figure 1 illustrates, watermarking can preserve accuracy while shifting linguistic style

and reduce readability in ways that make the output less helpful for patient’s clinical understanding. This *helpfulness* dimension is critical for medical LLMs (Yun et al., 2023; Maslej et al., 2023), yet existing frameworks lack quantifiable metrics to measure style shifts and their impact on clinical helpfulness.

We benchmark five watermarking methods representing three categories (*logit-based*, *distortion-free*, and *semantic-based*) across two EHR transformation tasks spanning five clinical note types. We develop a multi-aspect framework evaluating detectability, accuracy, style, readability, and helpfulness. To quantify stylistic deviation, we adapt linguistic distance of Part-of-Speech tags ( $LD_{POS}$ ), enabling mechanistic analysis of how watermarking alters text structure.

Our evaluation reveals a critical watermarking-induced helpfulness degradation. EHR transformation tasks achieve only 7.7%–24.2% true positive rates under default parameters, far below the >90% typical in narrative tasks. Tuning for practical detection degrades helpfulness by 24.6%–32.5%. Mechanistic analysis via  $LD_{POS}$  reveals watermarking shifts style toward generic terminology and connected structures, reducing specificity and readability.

For clinical deployment, we recommend watermarking methods that preserve critical medical terminology and maintain expected Part-of-Speech distributions, with task-specific parameter tuning for transformation versus narrative contexts. Our contributions are:

- A multi-aspect evaluation framework with two helpfulness estimates: interpretable traditional approximation and LLM judge rating, validated against human expert annotation.
- Evidence that watermarking degrades helpfulness through style and readability shifts even when accuracy is preserved, with mechanistic explanation via Part-of-Speech analysis
- Demonstration that EHR transformation poses unique detectability challenges, with parameter analysis and sensitivity ablations<sup>1</sup>

## 2 Related Work

**Watermarking methods for LLMs.** LLM watermarking has evolved from statistical detection (Mitchell et al., 2023; Gehrmann et al., 2019)

<sup>1</sup>Code for all evaluation : <https://github.com/rochanaph/style-eval-wllm>

to embedding patterns during generation. Recent approaches modify token probabilities (Kirchenbauer et al., 2023, 2024), employ domain-specific strategies (Lee et al., 2024), manipulate sampling (Christ et al., 2024; Hu et al., 2023; Wu et al., 2024), or use semantic guidance (Liu et al., 2024a; Hou et al., 2024a,b; Huo et al., 2024). While effective in general domains, their impact on clinical text transformation remains unexplored.

**Benchmarking watermarking methods.** Existing benchmarks evaluate detection-quality trade-offs in general domains, assessing robustness (Tu et al., 2024; Singh and Zou, 2024; Molenda et al., 2024), attack resilience (Liang et al., 2025; Al Ghanim et al., 2025), and deployment readiness (Pang et al., 2024; Dang et al., 2025), but lack domain-specific evaluation (Ajith et al., 2024).

**In medical contexts,** prior work benchmarks factual accuracy in narrative generation (Hastuti et al., 2025; Kong et al., 2025), leaving unexplored EHR transformation tasks (Fleming et al., 2024; Munnangi et al., 2025; Van Veen et al., 2024) where structural compression may amplify watermarking’s impact on patient’s clinical understanding.

## 3 Method

We develop a multi-aspect evaluation framework for watermarking using LLM judgments and traditional metrics to predict helpfulness (Figure 2) validated with human evaluation.

### 3.1 Multi-Aspect Evaluation and Traditional Metric Selection

**Quality aspects.** We define four micro-level aspects: *accuracy*, *lexical quality*, *syntactic quality*, and *readability*, alongside a holistic *helpfulness* rating (Table 1). Micro-level aspects use a 3-point scale as these qualities are difficult to differentiate at finer granularity, while helpfulness uses a 5-point scale to avoid coarse binary judgments and allow nuanced clinical comprehension support. These quality comprehension question are assessed by LLM as judge (Section 4.4) in full experimental scope and human expert as validation (Section 3.3).

**Traditional metric selection.** While LLM-based evaluation is thorough, it is costly for large-scale watermarking studies across methods, parameters, and models. We identify traditional automatic metrics that correlate with human judgments, enabling efficient quality assessment that is sufficient for comparing watermarking methods within tasks.

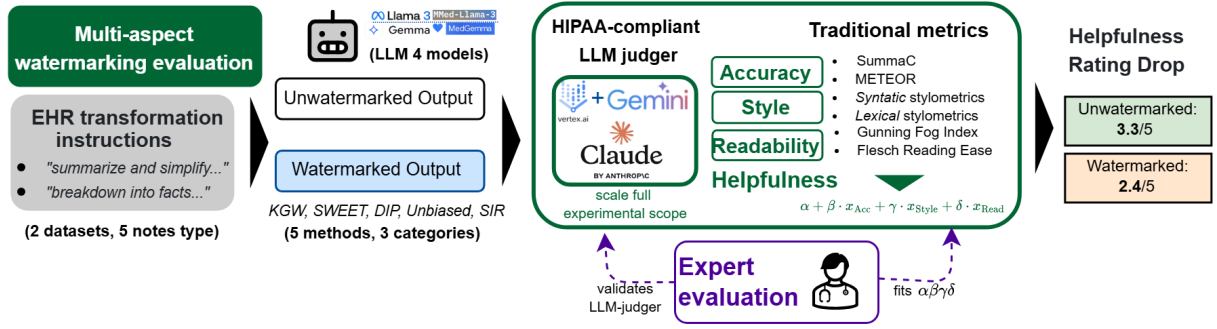


Figure 2: Multi-aspect evaluation framework for watermarking quality assessment. LLM-generated watermarked and unwatermarked EHR transformation outputs are evaluated across four aspects (Accuracy, Style, Readability, Helpfulness). The framework provides two helpfulness estimates: a **traditional approximation** for cost-free interpretable evaluation, and an **LLM judge rating** for thorough assessment. **Expert evaluation** (dashed) fits the formula weights and validates LLM judge reliability, supporting its use at full experimental scope.

| Aspect             | Definition and Rating Scale                                                                                                                                                                                                                                                                                                                                      |
|--------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Accuracy</b>    | Whether the response conveys correct information. 3: <i>Fully correct</i> (no misleading details, accurate terminology); 2: <i>Partially correct</i> (minor errors, generally correct terminology); 1: <i>Incorrect</i> (major errors, misleading information, incorrect terminology).                                                                           |
| <b>Style</b>       | <b>Lexical:</b> Appropriateness of vocabulary and word choice. 3: <i>Good</i> (precise and appropriate); 2: <i>Acceptable</i> (some awkward choices); 1: <i>Poor</i> (confusing terms).<br><b>Syntactic:</b> Grammar and sentence structure. 3: <i>Good</i> (well-formed); 2: <i>Acceptable</i> (minor structural issues); 1: <i>Poor</i> (confusing structure). |
| <b>Readability</b> | Clarity and ease of reading and scanning. 3: <i>Easy to read</i> (clear and concise); 2: <i>Moderate</i> ; 1: <i>Hard to read</i> (unclear and complex).                                                                                                                                                                                                         |
| <b>Helpfulness</b> | Whether the response supports understanding of clinical information in the transformed EHR. 5: <i>Very helpful</i> ; 4: <i>Helpful</i> ; 3: <i>Moderately helpful</i> ; 2: <i>Slightly helpful</i> ; 1: <i>Not helpful at all</i> .                                                                                                                              |

Table 1: Quality aspects and rating scale definitions for LLM judgment. Each micro aspect is rated on a 3-point scale where the Helpfulness with 5-point scale.

For **Accuracy**, we evaluate content similarity metrics including BERTScore (Zhang\* et al., 2020) and METEOR (Banerjee and Lavie, 2005) both used in EHR summarization task to evaluate correctness (Fleming et al., 2024) as well as other factuality metrics like SummaC and QuestEval (Scialom et al., 2021). For **Readability**, we assess established readability indices: Flesch Reading Ease (higher = easier), while Gunning Fog Index estimates years of education needed to understand a text (higher = harder). For **Style**, we consider lexical diversity (Type-Token Ratio, Yule's K) and

syntactic complexity (parse tree depth). Prior work on style quality relies solely on LLM (Wu and Aji, 2025) or classification accuracy of style transfer task (Mir et al., 2019; Tikhonov et al., 2019), which do not quantify stylistic deviation, as required in our setting. We therefore calculate linguistic distance of Part-of-Speech ( $LD_{PoS}$ ) as in Eq. 1. This give an interpretable distance between generated and reference texts, enabling both quantification and mechanistic analysis of watermarking-induced style changes.

$$LD_{PoS} = \frac{1}{2} \sum_{t \in \mathcal{T}_{PoS}} |p(t) - q(t)| \quad (1)$$

$\mathcal{T}_{PoS}$  cover all Part-of-Speech categories such as Noun, PROP, VERB, ADJ, PUNCT, etc. (Toutanova et al., 2003), and  $p(t)$  and  $q(t)$  denote the frequencies of PoS tag  $t$  in the generated and reference texts, respectively.

### 3.2 Helpfulness

While traditional metrics capture individual micro-aspects, **Helpfulness** is a holistic judgment which is difficult to be directly measured by any single metric. We therefore investigate helpfulness approximation by combining these micro-aspects ratings (Eq. 2). We employ sensitivity analysis on how each individual, pairwise, and composite aspects approximating helpfulness (Table 2), with leave-one-out cross validation for the human-annotated samples of FactEHR's nursing and progress notes. This provide varying type of notes, similar protocol on generalizing over unseen dataset (Xu et al., 2023) or domain (Zhao et al., 2020). Detail coeffi-

cient for each ablation model is in Table 5.

$$Help = \alpha + \beta \cdot x_{Acc} + \gamma \cdot x_{Style} + \delta \cdot x_{Read} \quad (2)$$

### 3.3 Human Evaluation

We conducted human evaluation with 3 healthcare professionals with LLM familiarity and medical expertise. The evaluation covered KGW watermarking on Nursing and Progress notes (n=120), chosen as KGW underlies many watermarking approaches and these note types span varying style and length. Each instance was rated on the quality aspects defined in Table 1, with the full evaluation template in Appendix B.2. Results across four validation stages are reported in Table 2, covering traditional metric selection, helpfulness formula weight fitting, detectability–quality correlation, and overall alignment between human judgment, LLM judge, and traditional approximation.

## 4 Experiment

### 4.1 Tasks and Datasets

We assess watermarking performance on two tasks in Electronic Health Records (EHR) covering various transformation instructions. First, **Summarization** using the MedAlign EHR dataset and instructions (Fleming et al., 2024). We use "Retrieve and Summarize" category, which requires models to summarize information from complete EHR. In this task, we use a full EHR document in XML format and truncate it to the last 8192 tokens to fit model context constraints. Second, **Fact Decomposition** using the FactEHR dataset (Munnangi et al., 2025) requires models to break down clinical notes from EHR into atomic, readable facts separated by //. This dataset encompasses four note types: procedure notes, progress notes, nursing notes, and discharge summaries, with experiments conducted separately for each type. For both tasks, we truncate model outputs to the first 200 tokens to maintain consistency in the evaluation following the prior works (Kirchenbauer et al., 2023; Hastuti et al., 2025). The truncation-length sensitivity analysis across all note types is in Appendix A.7. Details on dataset distribution and construction are provided in Appendix B.1.

Throughout this paper, unless stated otherwise, note type names (e.g., "procedure note") refer to the fact decomposition task, and "MedAlign" refers to the MedAlign EHR summarization task.

### 4.2 Watermarking Methods

We evaluate five watermarking methods across three categories (Liu et al., 2024b). The first category is **logit-based** watermarking, which serves as the foundation for most current approaches by modifying token probabilities through vocabulary partitioning. We evaluate two approaches: *KGW* (Kirchenbauer et al., 2023), which partitions the vocabulary into red and green lists based on a hash of the previous token and applies a bias  $\delta$  to green list tokens; and *SWEET* (Lee et al., 2024), an improvement over KGW that addresses low-entropy scenarios common in structured text generation, such as code or medical text (Hastuti et al., 2025).

The second category is **distortion-free** watermarking, theoretically designed to preserve original text quality by maintaining the LLM’s output distribution. We evaluate: *Unbiased* (Hu et al., 2023), which employs reweighting strategies (such as  $\gamma$ -reweight) to ensure the expected watermarked logits equal the original LLM logits; and *DiP-mark* (Wu et al., 2024), which uses complementary reweighting with random token permutations to maintain theoretical distribution preservation.

The third category **semantic-based** watermarking, incorporates semantic information to enhance robustness. Given our focus on style evaluation, this approach is particularly relevant as it may better preserve stylistic properties through semantic awareness. We evaluate *SIR* (Semantic Invariant Robust) (Liu et al., 2024a), which converts text semantics into red-green splits ensuring semantically similar texts get similar watermark pattern.

All methods use MarkLLM (Pan et al., 2024) implementation with default parameters for initial evaluation, followed by optimized parameters for improved detection. Brief descriptions of each scheme and its parameters are in Appendix B.3.

### 4.3 Model and Hardware

We use Llama3 8B Instruct (Grattafiori et al., 2024)<sup>4</sup> as the main model for most experiments, following the prior work (Munnangi et al., 2025). We also present ablation results that evaluate medical-domain adaptations, including Med42<sup>5</sup> (medical fine-tuned Llama3 8B), Gemma 3 4B<sup>6</sup>,

<sup>2</sup>Other evaluated metrics see Appendix A.1

<sup>3</sup>Leave-one-out cross-validation; Model coefficients see Appendix A.2

<sup>4</sup><https://hf.co/meta-llama/Meta-Llama-3-8B-Instruct>

<sup>5</sup><https://hf.co/m42-health/Llama3-Med42-8B>

<sup>6</sup><https://hf.co/google/gemma-3-4b-it>

| (1) Traditional Individual <sup>2</sup> | pearson  | spearman |
|-----------------------------------------|----------|----------|
| Acc - METEOR                            | 0.432**  | 0.422**  |
| Acc - QuestEval                         | 0.345**  | 0.359**  |
| Style - Type Token Ratio                | -0.246   | -0.241   |
| Style - LD <sub>PoS</sub>               | -0.533** | -0.348** |
| Read - Gunning Fog Index                | 0.251    | 0.249    |
| Read - Flesch Reading Ease              | -0.122   | -0.106   |

| (2) Traditional Helpfulness Approximation (LOOCV) <sup>3</sup> | R <sup>2</sup> | spearman |
|----------------------------------------------------------------|----------------|----------|
| Accuracy (Acc)                                                 | 0.630          | 0.622**  |
| Style                                                          | 0.580          | 0.622**  |
| Readability (Read)                                             | 0.630          | 0.621**  |
| Acc + Style                                                    | 0.860          | 0.890**  |
| Acc + Read                                                     | 0.888          | 0.919**  |
| Style + Read                                                   | 0.674          | 0.751**  |
| Acc + Style + Read                                             | 0.911          | 0.940**  |

| (3) Detectability and Quality | pearson | spearman |
|-------------------------------|---------|----------|
| Detectability - Accuracy      | -0.147  | -0.176   |
| Detectability - Style         | -0.284  | -0.358*  |
| Detectability - Readability   | -0.407* | -0.448*  |
| Detectability - Helpfulness   | -0.337  | -0.378*  |

| (4) Helpfulness Ratings  | pearson | spearman |
|--------------------------|---------|----------|
| Human - LLM              | 0.687** | 0.712**  |
| Human - Trad.Helpfulness | 0.500** | 0.502**  |

Table 2: Validation to human ratings across four stages. (1) Traditional Individual: Our LD<sub>PoS</sub> shows strongest correlation within stylometrics. (2) Traditional Helpfulness: combining all three aspects achieves best approximation. (3) Detectability and Quality: **watermarking correlates significantly with Helpfulness and Readability, but not Accuracy**. (4) Helpfulness Ratings: both **LLM judger and Traditional Helpfulness correlate significantly** to human, supporting their use at full experimental scope. \* $p < 0.05$  \*\* $p < 0.001$ .

and MedGemma 4B<sup>7</sup> to assess watermarking performance across different model scales and domain specializations. Model hyperparameters for text generation provided in Table 13. All experiments are conducted on a workstation equipped with an Intel(R) Xeon(R) W-2235 CPU @ 3.80 GHz (12 cores) processor and an NVIDIA GeForce RTX A6000 GPU (49140 MiB).

#### 4.4 LLM Judger

We employ Claude Sonnet 4.6 as our primary LLM judge, accessed through Anthropic’s HIPAA-compliant API. Claude Sonnet 4.6 represents current state-of-the-art capabilities with strong performance on medical evaluation tasks. Judger prompts adapted from previous work (Singh and Zou, 2024) and API costs are provided in Appendix B.5. We validate our evaluation methodology through an ablation study against Gemini 2.5 Pro (Section 5.4).

<sup>7</sup><https://hf.co/google/medgemma-4b-it>

## 5 Results

Our results are organized around four research questions. We first identify aspect-level quality metrics and validate helpfulness approximation against human ratings (RQ1), then reveal watermarking-induced helpfulness degradation and its detectability tension (RQ2). We interpret degradation mechanistically via LD<sub>PoS</sub> stylistic shift analysis (RQ3), and assess robustness across model families and LLM judges (RQ4).

### 5.1 RQ1: Metric Selection and Helpfulness Approximation

Table 2 reports validation across four stages against human ratings. For Traditional Individual metric selection (1), METEOR best captures Accuracy ( $r = 0.432$ ); BERTScore and SummaC yield no improvement (Appendix 4), suggesting a domain-specific ceiling for clinical text. Among stylo-metrics, LD<sub>PoS</sub> shows the strongest correlation ( $r = -0.533$ ), outperforming lexical metrics (TTR, Yule’s K), which indicate surface-level variation alone is insufficient to capture style quality. For Helpfulness approximation (2), combining all three aspects achieves the best performance ( $R^2 = 0.911$ , Spearman  $r = 0.940$ ), with each aspect contributing complementarily.

To investigate which aspect watermarking perturbs, we compare  $R^2$  across subsets. Restricted to watermarked outputs, Accuracy’s predictive power drops ( $R^2: 0.630 \rightarrow 0.503$ ) while Readability’s rises ( $0.630 \rightarrow 0.721$ ), which reflect Accuracy’s collapsed variance under watermarking, leaving Readability as the discriminating dimension. This is also confirmed through Detectability–Quality correlations (3), where watermark strength is non-significant for Accuracy ( $r = -0.147$ ) but significant for Readability ( $r = -0.407$ ,  $p < .05$ ) and Helpfulness ( $r = -0.337$ ,  $p < .05$ ). This point the degradation through readability despite Accuracy being the dominant predictor ( $\beta = 0.574$ , Appendix 5). This dissociation between predictive importance and watermarking sensitivity support our motivation of multi-aspect evaluation framework. Both LLM judger and Traditional Helpfulness correlate significantly with human judgment (4) ( $r = 0.687$ – $0.712$  and  $r = 0.500$ – $0.502$  respectively), supporting their relative ranking ability and complementary use at full experimental scope in the following analyses.

| Watermarking Scheme      | Default Parameters |              |                               |                         | Improved Detectability |              |                               |                         |
|--------------------------|--------------------|--------------|-------------------------------|-------------------------|------------------------|--------------|-------------------------------|-------------------------|
|                          | TPR                | AUROC        | Help Traditional ( $\Delta$ ) | Help LLM ( $\Delta$ )   | TPR                    | AUROC        | Help Traditional ( $\Delta$ ) | Help LLM ( $\Delta$ )   |
| <b>Procedure Note</b>    |                    |              |                               |                         |                        |              |                               |                         |
| KGW                      | 0.070              | 0.864        | $2.85 \pm 0.50$ (-0.04)       | $2.61 \pm 0.84$ (+0.21) | 0.910                  | 0.993        | $2.49 \pm 0.45$ (-0.41)       | $2.14 \pm 0.91$ (-0.65) |
| SWEET                    | 0.085              | 0.844        | $2.86 \pm 0.44$ (-0.03)       | $2.54 \pm 0.73$ (+0.09) | 0.840                  | 0.996        | $2.73 \pm 0.47$ (-0.17)       | $2.02 \pm 0.90$ (-0.83) |
| DIP                      | 0.000              | 0.449        | $2.90 \pm 0.48$ (+0.01)       | $2.61 \pm 0.86$ (+0.16) | -                      | -            | -                             | -                       |
| Unbiased                 | 0.000              | 0.499        | $2.89 \pm 0.45$ (+0.01)       | $2.59 \pm 0.73$ (+0.12) | -                      | -            | -                             | -                       |
| SIR                      | 0.045              | 0.803        | $2.74 \pm 0.47$ (-0.15)       | $2.38 \pm 0.86$ (-0.17) | 0.915                  | 0.979        | $2.45 \pm 0.49$ (-0.45)       | $1.75 \pm 0.88$ (-1.28) |
| <b>Average</b>           | <b>0.040</b>       | <b>0.692</b> | $\Delta -0.04$                | $\Delta +0.08$          | <b>0.888</b>           | <b>0.989</b> | $\Delta -0.34$                | $\Delta -0.92$          |
| <b>Progress Note</b>     |                    |              |                               |                         |                        |              |                               |                         |
| KGW                      | 0.150              | 0.880        | $2.69 \pm 0.38$ (-0.08)       | $2.63 \pm 0.81$ (-0.04) | 0.930                  | 0.992        | $2.63 \pm 0.35$ (-0.08)       | $2.81 \pm 0.92$ (-0.09) |
| SWEET                    | 0.165              | 0.910        | $2.75 \pm 0.39$ (-0.02)       | $2.69 \pm 0.74$ (-0.01) | 0.900                  | 0.995        | $2.64 \pm 0.41$ (-0.07)       | $2.15 \pm 0.95$ (-1.10) |
| DIP                      | 0.000              | 0.497        | $2.74 \pm 0.41$ (-0.03)       | $2.71 \pm 0.81$ (+0.01) | -                      | -            | -                             | -                       |
| Unbiased                 | 0.000              | 0.534        | $2.74 \pm 0.37$ (-0.04)       | $2.69 \pm 0.79$ (-0.01) | -                      | -            | -                             | -                       |
| SIR                      | 0.070              | 0.796        | $2.73 \pm 0.38$ (-0.04)       | $2.54 \pm 0.83$ (-0.25) | 0.742                  | 0.959        | $2.47 \pm 0.48$ (-0.24)       | $2.02 \pm 0.96$ (-1.23) |
| <b>Average</b>           | <b>0.077</b>       | <b>0.723</b> | $\Delta -0.04$                | $\Delta -0.05$          | <b>0.857</b>           | <b>0.982</b> | $\Delta -0.13$                | $\Delta -0.81$          |
| <b>Nursing Note</b>      |                    |              |                               |                         |                        |              |                               |                         |
| KGW                      | 0.270              | 0.867        | $2.69 \pm 0.48$ (+0.05)       | $2.55 \pm 0.92$ (+0.03) | 0.900                  | 0.987        | $2.41 \pm 0.47$ (-0.23)       | $2.30 \pm 0.95$ (-0.44) |
| SWEET                    | 0.250              | 0.809        | $2.63 \pm 0.47$ (-0.01)       | $2.46 \pm 0.82$ (-0.07) | 0.980                  | 0.999        | $2.61 \pm 0.51$ (-0.07)       | $2.04 \pm 0.94$ (-0.87) |
| DIP                      | 0.010              | 0.560        | $2.62 \pm 0.52$ (-0.02)       | $2.48 \pm 0.85$ (-0.08) | -                      | -            | -                             | -                       |
| Unbiased                 | 0.000              | 0.464        | $2.61 \pm 0.54$ (-0.04)       | $2.49 \pm 0.82$ (+0.03) | -                      | -            | -                             | -                       |
| SIR                      | 0.120              | 0.822        | $2.49 \pm 0.54$ (-0.16)       | $2.35 \pm 0.95$ (-0.21) | 0.849                  | 0.984        | $2.30 \pm 0.53$ (-0.34)       | $2.07 \pm 1.08$ (-0.90) |
| <b>Average</b>           | <b>0.130</b>       | <b>0.704</b> | $\Delta -0.04$                | $\Delta -0.06$          | <b>0.910</b>           | <b>0.990</b> | $\Delta -0.20$                | $\Delta -0.74$          |
| <b>Discharge Summary</b> |                    |              |                               |                         |                        |              |                               |                         |
| KGW                      | 0.240              | 0.863        | $2.73 \pm 0.40$ (+0.00)       | $2.54 \pm 0.74$ (+0.14) | 0.990                  | 0.998        | $2.67 \pm 0.43$ (-0.06)       | $2.66 \pm 0.84$ (-0.17) |
| SWEET                    | 0.320              | 0.942        | $2.76 \pm 0.39$ (+0.03)       | $2.54 \pm 0.65$ (+0.20) | 1.000                  | 1.000        | $2.59 \pm 0.38$ (-0.14)       | $1.97 \pm 0.92$ (-1.06) |
| DIP                      | 0.020              | 0.573        | $2.74 \pm 0.36$ (+0.03)       | $2.70 \pm 0.66$ (-0.07) | -                      | -            | -                             | -                       |
| Unbiased                 | 0.000              | 0.505        | $2.74 \pm 0.31$ (+0.01)       | $2.77 \pm 0.92$ (-0.07) | -                      | -            | -                             | -                       |
| SIR                      | 0.180              | 0.887        | $2.68 \pm 0.37$ (-0.04)       | $2.46 \pm 0.83$ (-0.25) | 0.800                  | 0.989        | $2.51 \pm 0.49$ (-0.22)       | $2.10 \pm 0.90$ (-0.97) |
| <b>Average</b>           | <b>0.152</b>       | <b>0.754</b> | $\Delta +0.00$                | $\Delta -0.09$          | <b>0.930</b>           | <b>0.996</b> | $\Delta -0.14$                | $\Delta -0.73$          |
| <b>Medalign - EHR</b>    |                    |              |                               |                         |                        |              |                               |                         |
| KGW                      | 0.556              | 0.972        | $2.75 \pm 0.53$ (-0.25)       | $2.08 \pm 0.94$ (-0.06) | 0.944                  | 0.999        | $2.71 \pm 0.50$ (-0.23)       | $1.78 \pm 1.04$ (-0.82) |
| SWEET                    | 0.646              | 0.969        | $2.92 \pm 0.65$ (-0.08)       | $1.97 \pm 0.87$ (-0.18) | 0.979                  | 0.997        | $2.83 \pm 0.60$ (-0.11)       | $1.82 \pm 0.95$ (-0.68) |
| DIP                      | 0.000              | 0.542        | $2.87 \pm 0.67$ (-0.13)       | $2.20 \pm 0.88$ (+0.04) | -                      | -            | -                             | -                       |
| Unbiased                 | 0.000              | 0.519        | $2.90 \pm 0.69$ (-0.09)       | $2.24 \pm 0.96$ (+0.14) | -                      | -            | -                             | -                       |
| SIR                      | 0.007              | 0.801        | $2.78 \pm 0.59$ (-0.21)       | $1.85 \pm 0.82$ (-0.32) | 0.257                  | 0.986        | $2.82 \pm 0.69$ (-0.12)       | $1.47 \pm 0.84$ (-1.16) |
| <b>Average</b>           | <b>0.242</b>       | <b>0.761</b> | $\Delta -0.15$                | $\Delta -0.08$          | <b>0.727</b>           | <b>0.994</b> | $\Delta -0.15$                | $\Delta -0.89$          |

Table 3: Performance between Detection rates (TPR and AUROC) and Helpfulness rating (5-scale) of Llama 3. **Default Parameters:** Slight Helpfulness rating drop ( $\Delta < 0.25$ ) but poor detectability (average TPR  $< 0.3$ ). **Improved Detectability:** Achieving decent detectability (TPR  $> 0.7$ ) in Procedure Note costs up to 0.92 (32.5%) Helpfulness LLM rating drop. Complete results with quality proxies and additional statistical significance are in Appendix A.3 and A.4, respectively.  $\Delta$ : Watermarked - Unwatermarked rating difference. Helpfulness rating drops:  $< 0.25$  vs  $\geq 0.25$ .

## 5.2 RQ2: Helpfulness Degradation and Watermark Detectability

**Impractical Detectability in EHR Tasks** Table 3 presents the comprehensive evaluation of watermarking methods across all clinical note types. Under default suggested parameters from the MarkLLM (Pan et al., 2024), watermarking methods demonstrate a slight helpfulness drop across most tasks, with average drops ranging from  $\Delta - 0.02$  to  $\Delta - 0.011$  for both our helpfulness rating (Help Traditional) and LLM-judged helpfulness rating (Help LLM). However, a critical limitation existed: extremely poor detectability, with average TPR values between 0.040 and 0.242 across tasks. These results are substantially lower than shown in prior work on the general domain (Kirchenbauer

et al., 2023), code generation (Lee et al., 2024), or even narrative medical text domain (Hastuti et al., 2025) with TPR  $> 90\%$ . We hypothesize this disparity arises from the nature of transformation tasks, where LLMs are constrained to preserve the content’s factual accuracy, limiting the flexibility needed for effective watermark embedding. This hypothesis is supported by the MedAlign summarization task, which permits greater linguistic variation and achieves notably higher TPR.

**Helpfulness Degradation** To achieve practical detectability, we tuned watermarking parameters targeting AUROC  $> 0.95$  (see Appendix B.3.7 for tuning details). In this setting, we exclude results from DIP and Unbiased, as the AUROC score for both methods is still very low (maximum  $\approx 0.6$ ),

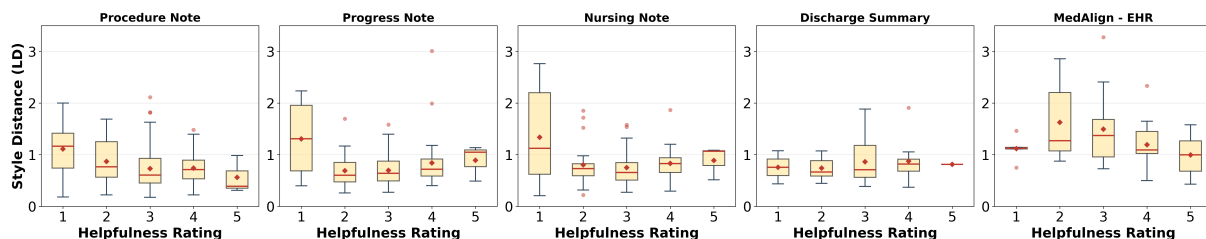


Figure 3:  $LD_{PoS}$  vs. Helpfulness ratings for high-accuracy unwatermarked text (BERTScore > 0.86). Procedure Notes exhibit expected negative correlation. However, other note types show no clear pattern. This hints that other valid and helpful stylistic approaches may exist beyond the reference. Motivating examining PoS shift in Figure 4.

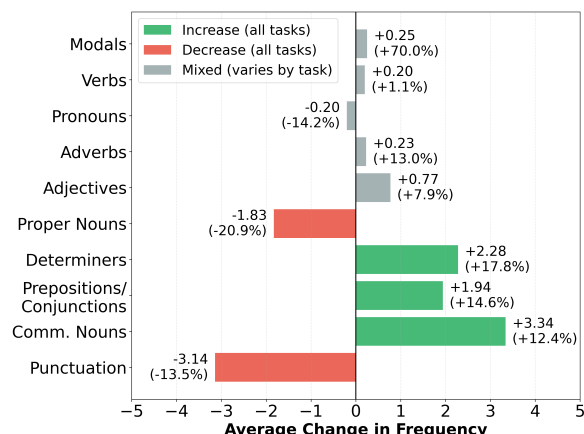


Figure 4: Part-of-speech (PoS) frequency shifts in watermarked transformed EHR text across tasks where helpfulness decreased despite high accuracy. **Watermarked texts shifts to have more generic terminology** (↓ proper nouns) **and dense structure** (↓ punctuation, ↑ prepositions/conjunctions).

showing their impractical detectability. This optimization elevates average TPR to 0.727–0.930 across tasks for KGW, SWEET, and SIR, but it comes at a substantial cost: helpfulness degrades dramatically, with Help Traditional average rating drops ranging from -0.13 to -0.34 (4.7%-11.8%) and Help LLM average rating drops ranging from -0.73 to -0.92 (24.6%-32.5%) across tasks. These results demonstrate watermarking consistently degrades helpfulness in clinical text transformation tasks, which becomes more severe when parameters are tuned for practical detection rates, highlighting a watermarking-induced helpfulness degradation and its detectability tension.

### 5.3 RQ3: Interpreting Helpfulness Degradation via Stylistic Shifts

To better understand the helpfulness degradation, we conduct a post-hoc analysis of watermarked outputs whose helpfulness decreases despite high accuracy (BERTScore > 0.89). Figure 3 shows  $LD_{PoS}$

distributions across helpfulness ratings: while Procedure Notes exhibit a clear correlation between stylistic distance and helpfulness, other note types display flat distributions, suggesting the existence of multiple valid styles beyond the reference. This motivates examining specific PoS category shifts (Figure 4).

**Lexical shifts.** Watermarked outputs show reduced lexical specificity: increased use of generic common nouns (+12.4%) and decreased proper nouns (-20.9%). This shift favors general terms over specific entities, consistent with watermark mechanisms biasing generation toward higher-entropy (Lee et al., 2024), interchangeable tokens, which align to previous work showing low-entropy token characterize medical terminology (Hastuti et al., 2025). **Syntactic shifts.** Watermarked text exhibits higher frequencies of conjunctions and prepositions (+14.6%) and reduced punctuation (-13.5%), promoting longer, more connected sentences. This contrasts with the segmented or enumerative style typical of unwatermarked outputs.

Together, these lexical and syntactic shifts provide a plausible explanation for decreased perceived helpfulness: factual content remains accurate, but reduced specificity and increased structural complexity may hinder readability and ease of information understanding. While the boxplot (Figure 3) shows that style distance alone does not perfectly predict helpfulness, it highlights the existence of alternative stylistic nuance that can still be considered helpful. Sample of watermarked text with impacted structure (Table 6) and complete PoS frequency distributions for all clinical notes are provided in Appendix A.5, which reveal the same proper-to-common-noun shift across all five note types, and in Figure 7 the shift is even clearer in more templated notes like Discharge Summary.

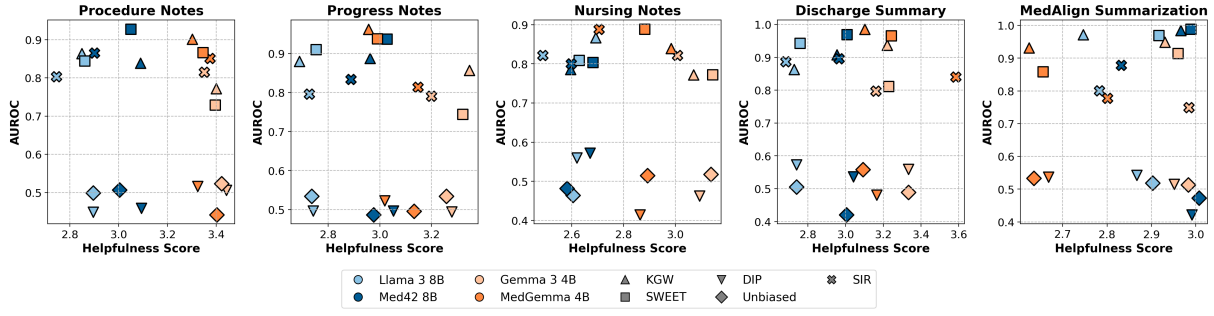


Figure 5: Helpfulness vs detectability (AUROC) across different models and tasks. Gemma 3 4B achieves higher helpfulness than Llama 3 8B, but with reduced detectability. Medical fine-tuning produces task-dependent effects.

#### 5.4 RQ4: Ablation Across Model Domain and LLM-as-a-Judge Consistency

**Model Domain and Size Effects** Figure 5 reveals differences in how watermarking affects model families. Despite being a smaller model (4B vs 8B parameters), Gemma 3 consistently outperforms Llama3 in helpfulness across fact decomposition tasks, though at the cost of lower detection performance. Medical fine-tuning produces divergent effects: MedGemma slightly degrades helpfulness across all fact decomposition tasks, while Med42 shows improvement compared to Llama 3. Across all models, distortion-free strategies (DIP and Unbiased) consistently fail to achieve reliable detection ( $\text{AUROC} \approx 0.5$ ). Complete results can be seen in Appendix A.3, Table 15 (Llama 3), Table 16 (Med42), Table 17 (Gemma 3), and Table 18 (MedGemma).

**LLM-as-a-Judge Reliability** To validate our evaluation methodology, we provide comparison results of our primary judge (Claude Sonnet) against Gemini 2.5 Pro on the Nursing Notes task across all watermarking schemes. Figure 6 demonstrates strong inter-judge agreement (Spearman  $\rho = 0.900$ ), with consistent ranking patterns across watermarking methods. Gemini rates watermarked outputs systematically lower than Claude, but the relative ordering of schemes remains stable (KGW, DIP, Unbiased > SWEET > SIR). This high correlation validates that our helpfulness assessments reflect genuine quality differences rather than judge-specific biases, supporting the reliability of LLM-based evaluation for clinical text quality assessment.

## 6 Conclusion

We present a multi-aspect evaluation framework for LLM watermarking in clinical EHR transformation

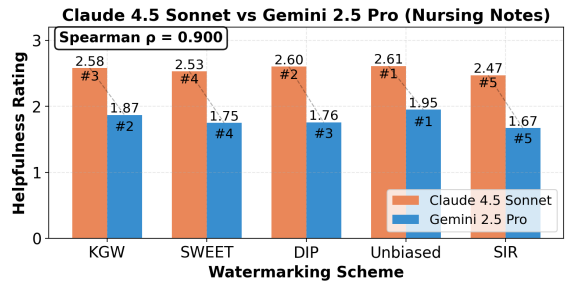


Figure 6: The average of the helpfulness rating between two LLM judges across schemes. Gemini rates are systematically lower than Claude, but relative scheme ordering remains consistent.

tasks, revealing fundamental challenges in medical contexts. Our analysis across five clinical note types shows that achieving practical detectability degrades helpfulness substantially. Through our  $\text{LD}_{\text{PoS}}$  metric and analysis, we provide mechanistic evidence that watermarking shifts text toward generic terminology and connected structures, degrading helpfulness even when accuracy is preserved.

Our findings indicate that clinical deployment requires task-specific parameter optimization, watermarking methods preserving medical terminology and syntactic structure, and evaluation frameworks incorporating domain-specific quality metrics. Future work should develop entropy-adaptive and syntax-preserving approaches specifically designed for transformation tasks where factual accuracy and stylistic structure are both crucial.

## Limitations

This study includes a human evaluation to validate the alignment between GPT-Judger assessments and human judgments, limited to the Nursing and Progress notes. Our objective was not to establish comprehensive human evaluation across all

tasks, but to assess the reliability of our proposed workflow, which we demonstrate through a strong correlation with human judgments. We also identified contributing factors reducing helpfulness such as style shifts and reduce readability, while other potential factors likely exist beyond this work. The connections between these mechanisms and various watermarking approaches represent a promising area for further research toward developing more domain-sensitive watermarking techniques.

## Ethical Statement

This work investigates the implementation of existing LLM watermarking methods in clinical text transformation tasks. We note that applying these methods without consideration of domain-specific quality dimensions may compromise the utility of AI-generated clinical summaries. Deployment in actual clinical settings requires extensive validation with domain experts beyond the dimensions examined here. Healthcare organizations should conduct a thorough evaluation to establish clinically meaningful thresholds for acceptable quality degradation before implementing watermarked AI systems in real-world healthcare settings.

## References

- Anirudh Ajith, Sameer Singh, and Danish Pruthi. 2024. [Downstream trade-offs of a family of text watermarks](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14039–14053, Miami, Florida, USA. Association for Computational Linguistics.
- Mansour Al Ghanim, Jiaqi Xue, Rochana Prih Hastuti, Mengxin Zheng, Yan Solihin, and Qian Lou. 2025. [Evaluating the robustness and accuracy of text watermarking under real-world cross-lingual manipulations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7396–7416, Suzhou, China. Association for Computational Linguistics.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Miranda Christ, Sam Gunn, and Or Zamir. 2024. [Undetectable watermarks for language models](#). In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1125–1139. PMLR.
- Kieu Dang, Phung Lai, NhatHai Phan, Yelong Shen, Ruoming Jin, Abdallah Khreishah, and My T. Thai. 2025. [Sok: Are watermarks in llms ready for deployment?](#) *Preprint*, arXiv:2506.05594.
- European Commission AI. 2026. [Code of practice on transparency of AI-generated content: Second draft](#). European Commission. Draft Code under Article 50 of the EU AI Act; obligations applicable August 2026.
- Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Jenkins, Ethan Steinberg, Ashwin Nayak, Birju Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, and 11 others. 2024. [Medalign: a clinician-generated dataset for instruction following with electronic medical records](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’24/IAAI’24/EAAI’24*. AAAI Press.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. [GLTR: Statistical detection and visualization of generated text](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Rochana Prih Hastuti, Rian Adam Rajagede, Mansour Al Ghanim, Mengxin Zheng, and Qian Lou. 2025. [Factuality beyond coherence: Evaluating LLM watermarking methods for medical texts](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.
- Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2024a. [SemStamp: A semantic watermark with paraphrastic robustness for text generation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4067–4082, Mexico City, Mexico. Association for Computational Linguistics.
- Abe Hou, Jingyu Zhang, Yichen Wang, Daniel Khashabi, and Tianxing He. 2024b. [k-SemStamp](#):

- A clustering-based semantic watermark for detection of machine-generated text. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1706–1715, Bangkok, Thailand. Association for Computational Linguistics.
- Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2023. [Unbiased watermark for large language models](#). *Preprint*, arXiv:2310.10669.
- Mingjia Huo, Sai Ashish Somayajula, Youwei Liang, Ruishi Zhang, Farinaz Koushanfar, and Pengtao Xie. 2024. Token-specific watermarking with enhanced detectability and semantic coherence for large language models. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. [A watermark for large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2024. [On the reliability of watermarks for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Cong Kong, Rui Xu, Weixi Chen, Jiawei Chen, and Zhaoxia Yin. 2025. [Protecting copyright of medical pre-trained language models: Training-free backdoor model watermarking](#). *Preprint*, arXiv:2409.10570.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. [SummaC: Re-visiting NLI-based models for inconsistency detection in summarization](#). *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. 2024. [Who wrote this code? watermarking for code generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4890–4911, Bangkok, Thailand. Association for Computational Linguistics.
- Jiacheng Liang, Zian Wang, Spencer Hong, Shouling Ji, and Ting Wang. 2025. [Watermark under fire: A robustness evaluation of LLM watermarking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21050–21074, Suzhou, China. Association for Computational Linguistics.
- Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2024a. [A semantic invariant robust watermark for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. 2024b. [A survey of text watermarking in the era of large language models](#). *ACM Comput. Surv.*, 57(2).
- Marta M Maslej, Stefan Kloiber, Marzyeh Ghassemi, Joanna Yu, and Sean L Hill. 2023. [Out with ai, in with the psychiatrist: a preference for human-derived clinical decision support in depression care](#). *Translational Psychiatry*, 13(1):210.
- Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. [Evaluating style transfer for text](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. [DetectGPT: Zero-shot machine-generated text detection using probability curvature](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.
- Piotr Molenda, Adian Liusie, and Mark Gales. 2024. [WaterJudge: Quality-detection trade-off when watermarking large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3515–3525, Mexico City, Mexico. Association for Computational Linguistics.
- Monica Munnangi, Akshay Swaminathan, Jason Alan Fries, Jenelle A Jindal, Sanjana Narayanan, Ivan Lopez, Lucia Tu, Philip Chung, Jesutofunmi Omiye, Mehr Kashyap, and Nigam Shah. 2025. [FactEHR: A dataset for evaluating factuality in clinical notes using LLMs](#). In *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298 of *Proceedings of Machine Learning Research*. PMLR.
- OpenAI. 2026. [Introducing ChatGPT health](#). OpenAI Blog. Accessed: March 2026.
- Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuan-dong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. 2024. [MarkLLM: An open-source toolkit for LLM watermarking](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 61–71, Miami, Florida, USA. Association for Computational Linguistics.
- Qi Pang, Shengyuan Hu, Wenting Zheng, and Virginia Smith. 2024. [No free lunch in LLM watermarking: Trade-offs in watermarking design choices](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang,

- and Patrick Gallinari. 2021. [QuestEval: Summarization asks for fact-based evaluation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Karanpartap Singh and James Zou. 2024. [New evaluation metrics capture quality degradation due to LLM watermarking](#). *Transactions on Machine Learning Research*.
- Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan P. Yamshchikov. 2019. [Style transfer for texts: Retrain, report errors, compare with rewrites](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3936–3945, Hong Kong, China. Association for Computational Linguistics.
- Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. [Feature-rich part-of-speech tagging with a cyclic dependency network](#). In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.
- Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. 2024. [WaterBench: Towards holistic evaluation of watermarks for large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1542, Bangkok, Thailand. Association for Computational Linguistics.
- Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. [Adapted large language models can outperform medical experts in clinical text summarization](#). *Nature Medicine*.
- Minghao Wu and Alham Fikri Aji. 2025. [Style over substance: Evaluation biases for large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312, Abu Dhabi, UAE. Association for Computational Linguistics.
- Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. 2024. A resilient and accessible distribution-preserving watermark for large language models. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. [Globem: Cross-dataset generalization of longitudinal human behavior modeling](#). *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(4).
- Jiaqi Xue, Yifei Zhao, Mansour Al Ghanim, Shangqian Gao, Ruimin Sun, Qian Lou, and Mengxin Zheng. 2025. [Pro: Enabling precise and robust text watermark for open-source llms](#). *Preprint*, arXiv:2510.23891.
- Hye Yun, Iain Marshall, Thomas Trikalinos, and Byron Wallace. 2023. [Appraising the potential uses and harms of LLMs for medical systematic reviews](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10122–10139, Singapore. Association for Computational Linguistics.
- Tianyi Zhang\*, Varsha Kishore\*, Felix Wu\*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. 2020. [Domain generalization via entropy regularization](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 16096–16107. Curran Associates, Inc.

## A Additional Results and Analysis

### A.1 Complete Traditional Individual Metrics

In addition to the metrics shown in Table 2, we also evaluate other traditional metrics, such as SummaC(Laban et al., 2022) and BERTScore(Zhang\* et al., 2020), shown in Table 4.

| Metrics                    | Pearson  | Spearman |
|----------------------------|----------|----------|
| Acc - METEOR               | 0.432**  | 0.422**  |
| Acc - QuestEval            | 0.345*   | 0.359*   |
| Acc - SummaC               | 0.133    | 0.105    |
| Acc - BERTScore            | 0.384*   | 0.333*   |
| Style - Type Token Ratio   | -0.246   | -0.241   |
| Style - LD <sub>POS</sub>  | -0.533** | -0.348** |
| Style - Parse-Tree Depth   | -0.391   | -0.370   |
| Style - Yule’s K           | 0.206    | 0.250    |
| Read - Gunning Fog Index   | 0.251    | 0.249    |
| Read - Flesch Reading Ease | -0.122   | -0.106   |

Table 4: Traditional metrics selection for individual aspect.

### A.2 Regression Coefficients

Given Equation 2, the coefficient values for each model are shown in Table 5.

| Model              | $\alpha$ | $\beta$ | $\gamma$ | $\delta$ |
|--------------------|----------|---------|----------|----------|
| Accuracy (Acc)     | 0.093    | 0.880   | -        | -        |
| Style              | 0.124    | -       | 0.657    | -        |
| Readability (Read) | 0.127    | -       | -        | 0.623    |
| Acc + Style        | -0.058   | 0.623   | 0.440    | -        |
| Acc + Read         | -0.054   | 0.607   | -        | 0.428    |
| Style + Read       | 0.076    | -       | 0.324    | 0.389    |
| Acc + Style + Read | -0.078   | 0.574   | 0.215    | 0.282    |

Table 5: Regression coefficient of each linear model conducted on sensitivity analysis.

### A.3 Complete Detectability and Quality Score

This section presents comprehensive evaluation results for all watermarking schemes across the four language models tested in our study. Tables 15 through 18 report detectability metrics (TPR at 0% FPR and AUROC) alongside quality metrics spanning accuracy, lexical diversity, syntactic structure, and readability dimensions. Complete results can be seen in Table 15 (Llama 3), Table 16 (Med42), Table 17 (Gemma 3), and Table 18 (MedGemma). Each table includes results for five clinical note generation tasks and the MedAlign summarization task, with unwatermarked baselines provided for reference.

### A.4 Statistical Significance

Beyond the standard deviations in Table 3, we report one-sample t-tests comparing each watermarking scheme against the unwatermarked condition. In the Improved Detectability setting, 90% of observations degrade significantly ( $p < 0.001$ ); in the Default setting, 31% do ( $p < 0.05$ ). We also report 95% CIs and Cohen’s d for all results. CIs are narrow (width 0.179–0.298) and fully negative for all significant degradations, confirming consistent effect direction. Cohen’s d further corroborates this: 85.2% of significant degradations in the Improved setting are medium-to-large ( $d \geq 0.5$ ), indicating effects are reliable even at 100 samples.

### A.5 PoS Tag Shifts

**Text sample.** PoS shifts for each clinical note shown in Figure 7. All note types exhibit the NNP→NN swap (proper to common nouns), with most showing increased determiners and function words. Other than NNP→NN trends, significant pattern is increased frequency of Coordinating Conjunction (CC) as seen in Table 6, where it turned the text into more dense and connected structure. This plausibly negate the goal of simplification and reduce readability.

**Clinical note-specific PoS shift.** *Discharge Summary* demonstrates large increases in determiners (DT), coordinating conjunctions (CC), and plural nouns (NNS) alongside the NNP→NN swap. *Nursing Note* shows substantial common noun increases and notable decreases in punctuation. *Procedure Note* exhibits the smallest magnitude NNP→NN swap with minimal common noun increases, but demonstrates a dramatic shift from past tense verbs (VBD) to base form verbs (VB) and coordinating conjunctions (CC), representing a transition from past-tense narrative to more connected, present-oriented structures. *Progress Note* demonstrates a strong NNP→NN swap with substantial increases in determiners (DT) and adjectives (JJ), creating generic, determiner-heavy descriptive language. *MedAlign* shows the most narrative transformation with large increases in past tense verbs (VBD), prepositions (IN), adjectives (JJ), and coordinating conjunctions (CC), alongside decreases in personal (PRP) and possessive pronouns (PRP\$), suggesting a shift toward more descriptive, narrative phrasing with generic terminology.

| Unwatermarked                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Watermarked                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Increased Coordinating Conjunctions (CC) (from 1 → 9)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| The patient is a 34-year-old female. // The patient has a history of Wolff-Parkinson-WHITE syndrome. // The patient has a history factor 5 leiden mutation. // The patient has a history right ankle reconstruction. // The pain is located on the anterior ankle. // The swelling is located on lateral aspect ankle. // There is no pain with passive ROM. // The level of pain is 1/10. // There are no signs of radiation of pain to the foot. // There was no numbness or tingling. // There were no signs of weakness. // The injury occurred on April 11, 2018. // The MRI report shows a 15x8mm hypointense lesions in the anterior-lateral subcutaneous tissues. // The lesion is indeterminate. // The differential diagnosis includes postsurgery area of fibroses, thrombosis, complex hemat | // The patient is 34 years old.\n\n// The patient has a history of past medical conditions including WPW, a factor 5Leiden mutation, right ankle reconstruction, and high ankle sprains.\n\n// The patient presents with right ankle swelling and pain after a specific incident on April 11, 2018.\n\n// The swelling and pain started after the patient tripped ascending the stairs, and her foot \"went forward and then turned inwards\" and made a cracking or popping sound.\n\n// The ER Xray did not show any fractures.\n\n// The CAM walker helped with pain reduction, and MRI shows swelling in lateral soft tissues with a 15x8mm hypointense lesion anterior lateral to the ankle, ruptured anterolateral fibular (ATFL), and partial tear calcaneofibular (CFL).\n\n// The patient reports having experienced |

Table 6: Unwatermarked-watermarked text sample where it shows **increased frequency of Coordinating Conjunctions (CC) PoS tag**.

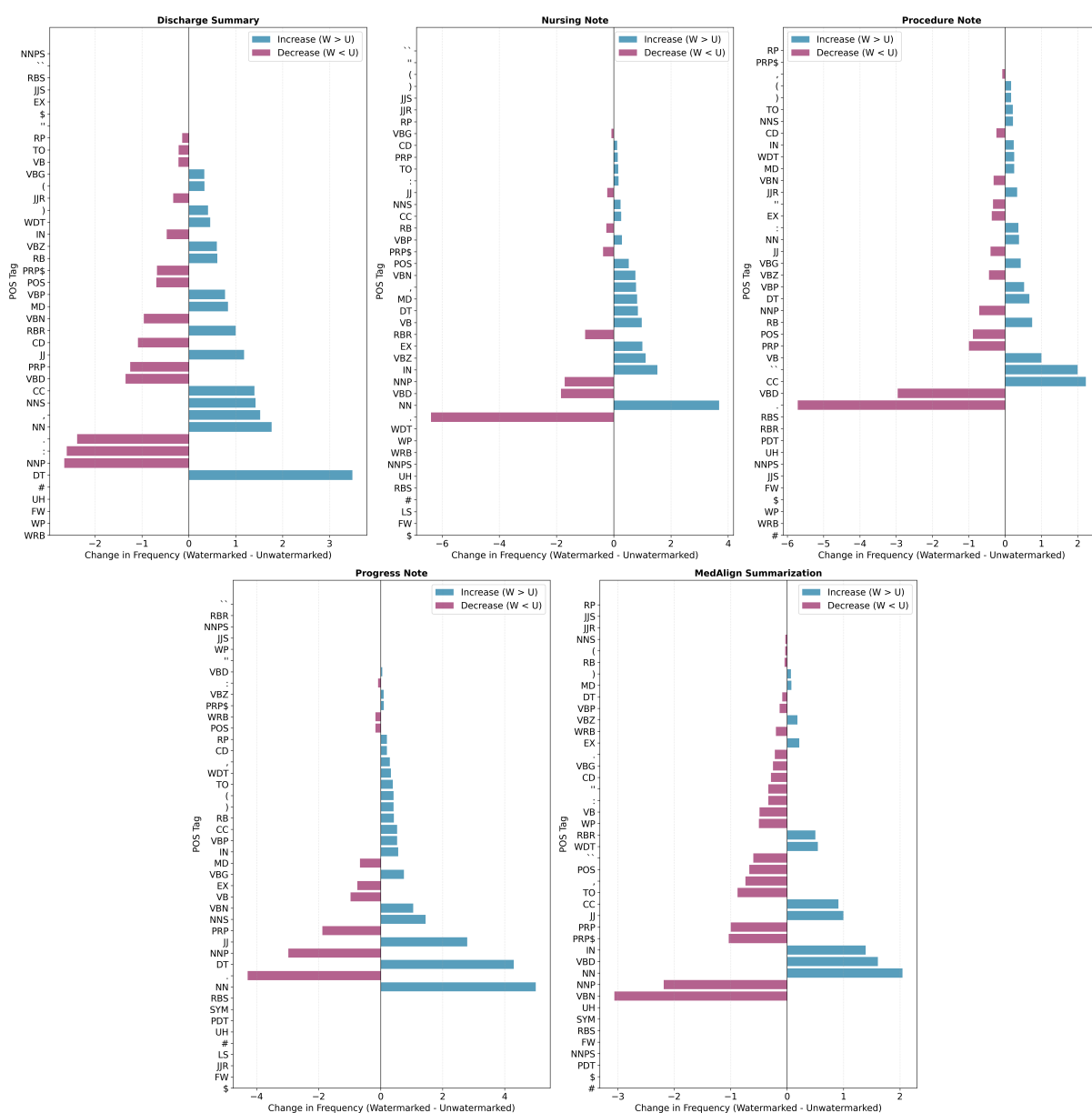


Figure 7: Part of Speech frequency change of Watermarked - Unwatermarked text for each clinical note type. **Major NNP→NN swap trend** seen in all notes.

## A.6 Effect of Watermarking Parameters on PoS Shifts

Table 7 provides ablation of how KGW’s parameter  $\delta$  affects PoS tag distribution shifts. The overall trend is consistent with Figure 4: decreases concentrate on proper nouns (PPN) and punctuation (Punc.), while increases concentrate on common nouns (NN), determiners (Det.), and conjunctions (Con.). This aligns with the conclusion that watermarking systematically shifts style toward generic terminology and connected structures across all parameter settings.

## A.7 Token Length Sensitivity

As discussed in Section 4.1, the 200-token output truncation aligns with established watermarking works (Kirchenbauer et al., 2023; Hastuti et al., 2025), and ensures fair comparison across clinical note types, as nursing notes average  $\sim 209$  tokens and longer thresholds would yield insufficient samples. Table 8 reports KGW detection and helpfulness across truncation lengths for all four note types. In every note type, TPR does not peak at 200 tokens: Nursing Notes peak at 150 tokens (0.280), Progress Notes at 150 tokens (0.275), Discharge Summaries at 150 tokens (0.390), and Procedure Notes at 250 tokens (0.160). Averaged across note types, TPR peaks at 150 tokens (0.275) and drops at 250 tokens (0.206). This confirms detection does not monotonically improve with length and rules out artificial inflation of results at our chosen 200-token cutoff.

## B Implementation Details and Methods Preliminary

### B.1 Tasks and Dataset

#### B.1.1 MedAlign Summarization

We utilize the MedAlign dataset (Fleming et al., 2024), which consists of instruction-EHR pairs where each instruction is a clinician-generated question paired with a patient’s Electronic Health Record (EHR) in XML format. The natural text comprises clinician-written responses to these questions. Following the original paper’s methodology, we apply several filtering criteria to ensure data quality: (1) select only questions from the *Retrieve and Summarize* category, (2) retain only responses with a minimum of 20 words to ensure substantive answers, (3) remove near-duplicate questions by filtering out those with maximum ROUGE-L similarity  $\geq 0.7$  with any other question in the dataset, and (4) restrict to general patient types (excluding disease-specific scenarios). After applying these filters, we obtain 144 data points.

We employ the instruction template from the original paper, as shown in Table 9. Since EHR records can exceed model context windows, we implement model-aware truncation: we calculate the available token budget by subtracting generation buffer (200 tokens) and prompt template overhead from the model’s context window, then truncate the EHR content from the beginning (preserving the most recent medical information) using the model’s tokenizer to ensure the complete prompt fits within the context limit. This strategy also used in the original work (Fleming et al., 2024). Additionally, we truncate clinician responses to a maximum of 200 tokens to standardize output length.

#### B.1.2 FactEHR Fact Decomposition

We utilize the FactEHR dataset (Munnangi et al., 2025), which requires models to decompose clinical notes from Electronic Health Records into atomic, readable facts separated by “//”. This dataset encompasses four distinct clinical note types with varying characteristics: **procedure notes**, **progress notes**, **nursing notes**, and **discharge summaries**. These note types differ substantially in length, structure, and clinical focus, as detailed in the original work (Munnangi et al., 2025).

We sample based on the size of the publicly available data for each note: 200 samples from procedure notes, 200 samples from progress notes,

| Parameter                           | Proper Noun   | Common Noun   | Det., Con.             | Punctuation   |
|-------------------------------------|---------------|---------------|------------------------|---------------|
| <b>KGW</b>                          |               |               |                        |               |
| Default ( $\delta=2$ )              | -28.1%        | +5.1%         | +18.3%, +7.6%          | -4.1%         |
| Improved Detect. ( $\delta=5,6,7$ ) | <b>-20.9%</b> | <b>+12.4%</b> | <b>+17.8%, +14.6%</b>  | <b>-13.5%</b> |
| Higher ( $\delta=8$ )               | -29.3%        | +3.6%         | +20.4%, +14.3%         | -19.2%        |
| <b>SWEET</b>                        |               |               |                        |               |
| Default                             | <b>-12.3%</b> | +5.0%         | +10.2%, <b>+9.0%</b>   | +4.0%         |
| Improved Detect.                    | <b>-10.2%</b> | -8.8%         | -9.0%, +0.3%           | <b>-20.4%</b> |
| <b>SIR</b>                          |               |               |                        |               |
| Default                             | 0.0%          | -8.4%         | -25.2%, -9.9%          | -2.6%         |
| Improved Detect.                    | -1.0%         | -18.5%        | <b>-77.5%</b> , -24.3% | -4.4%         |

Table 7: PoS frequency shifts under different KGW bias settings. Bold values indicate consistent shifts across all tasks.

| Metric            | Token Length |       |       |       |       |
|-------------------|--------------|-------|-------|-------|-------|
|                   | 50           | 100   | 150   | 200   | 250   |
| Procedure Note    |              |       |       |       |       |
| TPR               | 0.015        | 0.050 | 0.155 | 0.070 | 0.160 |
| AUROC             | 0.809        | 0.822 | 0.857 | 0.864 | 0.854 |
| Helpfulness       | 1.88         | 2.35  | 2.72  | 2.85  | 2.86  |
| Progress Note     |              |       |       |       |       |
| TPR               | 0.055        | 0.190 | 0.275 | 0.150 | 0.215 |
| AUROC             | 0.812        | 0.864 | 0.878 | 0.880 | 0.881 |
| Helpfulness       | 1.82         | 2.27  | 2.59  | 2.73  | 2.80  |
| Nursing Note      |              |       |       |       |       |
| TPR               | 0.120        | 0.210 | 0.280 | 0.270 | 0.150 |
| AUROC             | 0.741        | 0.845 | 0.828 | 0.867 | 0.834 |
| Helpfulness       | 1.88         | 2.24  | 2.47  | 2.69  | 2.65  |
| Discharge Summary |              |       |       |       |       |
| TPR               | 0.190        | 0.270 | 0.390 | 0.240 | 0.300 |
| AUROC             | 0.817        | 0.786 | 0.889 | 0.863 | 0.923 |
| Helpfulness       | 1.88         | 2.30  | 2.58  | 2.69  | 2.76  |

Table 8: KGW detection and helpfulness sensitivity to output truncation length across tasks.

| MedAlign Prompt Template                                     |
|--------------------------------------------------------------|
| Instruction: Answer the following question based on the EHR: |
| ###                                                          |
| Question: ""[question]""                                     |
| EHR: ""[ehr_content]""                                       |

Table 9: Prompt template used for MedAlign summarization tasks.

100 samples from nursing notes, and 100 samples from discharge summaries. The natural decomposition text that was chosen as reference is GPT-4o-generated outputs provided in the FactEHR dataset, whose sample achieved 100% accuracy validation with medical experts (Munnangi et al., 2025).

We employ the prompt template from the original paper with two few-shot examples, as shown in Table 10. Following prior work, we truncate both input clinical notes and model outputs to 200 tokens to maintain consistency across experiments.

#### FactEHR Fact Decomposition Prompt Template

Please breakdown the following text into independent facts as a string delimited by "/" to separate the facts.

Example 1:

Note: "There is a dense consolidation in the left lower lobe."

Atomic facts:

There is a consolidation. // The consolidation is dense. // The consolidation is on the left. // The consolidation is in a lobe. // The consolidation is in the lower portion of the left lobe.

Example 2:

Note: "The patient has been having intermittent shortness of breath for the last two years."

Atomic facts:

The patient has been having shortness of breath. // The shortness of breath is intermittent. // The shortness of breath has been present for the last two years.

Do not include any other text, or say "Here is the list..."

Note: {ehr\_content}

Table 10: Prompt template used for FactEHR fact decomposition tasks.

## B.2 Human evaluation

The human evaluation is conducted for the Nursing and Progress notes as they cover varying length and style which typically easier for humans to interpret and rate based on the reference answer, compared to other task like Discharge summary. Sample of questionnaire shown in Figure 8. We collected results from 3 respondents (medical/healthcare professional) to evaluate sample of 120 instances from both Un-watermarked and Watermarked text of the total generated texts, while making sure each item is rated by  $\geq 2$  different respondents. Respondents were recruited through academic social network. All participation was voluntary and was clearly stated in the consent page of the questionnaire. The collected data consist solely of participants' ratings and do not include identifiable information. This study has been approved by IRB under exempt human subject research.

## B.3 Watermarking Schemes

### B.3.1 KGW (Kirchenbauer et al., 2023)

We provide brief preliminaries for KGW watermarking approach. For a given language model  $f_{\text{LM}}$  with vocabulary  $\mathcal{V}$ , the likelihood probability of a token  $y_t$  is calculated as follows:

$$l_t = f_{\text{LM}}(x, y_{[:t]}) \quad (3)$$

$$p_{t,i} = \frac{e^{l_{t,i}}}{\sum_{j=1}^{|\mathcal{V}|} e^{l_{t,j}}} \quad (4)$$

where  $x = x_0, \dots, x_{M-1}$  and  $y_{[:t]} = y_1, \dots, y_{t-1}$  are an  $M$ -length tokenized prompt and the generated token sequence respectively, and  $l_t \in \mathbb{R}^{|\mathcal{V}|}$  is the logit vector.

**Watermarking** In the watermarking procedure, the entire tokens in  $\mathcal{V}$  at each time-step are randomly binned into green ( $\mathcal{G}_t$ ) and red ( $\mathcal{R}_t$ ) groups in proportions of  $\gamma$  and  $1 - \gamma$  ( $\gamma \in (0, 1)$ ), respectively. The method increases the logits of green group tokens by adding a fixed bias  $\delta$ , promoting these tokens to be sampled at each position. Consequently, watermarked LM-generated text is more likely than  $\gamma$  to contain tokens from the green group. In contrast, since humans have no knowledge of the hidden green-red partition rule, the proportion of green group tokens in human-written text is expected to be close to  $\gamma$ .

**Detection** The presence of watermarking in text is detected through a one-sided  $z$ -test by testing the null hypothesis that the text is not watermarked.

The  $z$ -score is calculated using the number of recognized green tokens in the test text. Subsequently, the text is determined to be watermarked if the  $z$ -score exceeds a predetermined threshold  $z_{\text{threshold}}$ .

### B.3.2 SWEET (Lee et al., 2024)

SWEET improve KGW (Kirchenbauer et al., 2023) by distinguishing watermark applicable tokens, meaning it embed and detect watermarks only within tokens with high entropy.

**Watermarking** Given a tokenized prompt  $x = x_0, \dots, x_{M-1}$  and already generated tokens  $y_{[:t]} = y_0, \dots, y_{t-1}$ , a model calculates an entropy value ( $H_t$ ) of the probability distribution for  $y_t$ . The watermarking only applied when  $H_t$  exceeds the threshold,  $\tau$ . Then, randomly split the vocabulary into green and red groups with a fixed green token ratio  $\gamma$ . If a token is selected to be watermarked, a constant  $\delta$  is added to the logits of green tokens, thereby promoting the sampling of these tokens. By limiting the promotion of green tokens only to positions with high entropy, it prevent modifications to the model's logit distribution for tokens where the model has high confidence (and, therefore, low entropy).

**Detection** Given a token sequence  $y = y_0, \dots, y_{N-1}$ , the task is to detect watermarks within  $y$ , thereby determining whether it was generated by the specific language model. As in the generation phase, SWEET computes the entropy values  $H_t$  for each  $y_t$ . Let  $N_h$  denote the number of tokens that have an entropy value  $H_t$  higher than the threshold  $\tau$ , and let  $N_h^G$  denote the number of green tokens among those in  $N_h$ . Finally, with the green list ratio  $\gamma$  used in the generation step, SWEET computes a  $z$ -score under the null hypothesis that the text is not watermarked:

$$z = \frac{N_h^G - \gamma N_h}{\sqrt{N_h \gamma (1 - \gamma)}} \quad (5)$$

The presence of a watermark can be assessed with increasing confidence as the  $z$ -score increases. SWEET sets  $z_{\text{threshold}}$  as a cut-off score. If  $z > z_{\text{threshold}}$  holds, it determines that the watermark is embedded in  $y$  and thus the text was generated by the LLM.

### B.3.3 Unbiased Watermark (Hu et al., 2023)

The Unbiased Watermark method introduces watermarking approaches that preserve the output probability distribution. The key innovation is that when

averaged over all possible watermark keys, the distribution remains identical to the original model, making the watermark undetectable to users without the secret key.

**Watermarking** The  $\gamma$ -reweight variant works by randomly shuffling the vocabulary according to a context-dependent watermark code  $E$ . After shuffling, the method rejects tokens in the bottom half of the cumulative probability distribution and doubles the probabilities of tokens in the top half, then renormalizes. Specifically, for a token  $t$  at position  $i$  in the shuffled order, the method computes a cumulative probability threshold. If this threshold exceeds 0.5, the token receives doubled probability; otherwise, it receives zero probability. The watermark code  $E$  is derived from a hash function that combines the secret key  $k$  with recent context tokens. To ensure unbiasedness across multiple generations, the method maintains a history of previously used context codes to avoid repetition.

**Detection** Detection uses a log-likelihood ratio test that compares how likely the text is under the watermarked model versus the original model. For each token, a score is computed based on the ratio of probabilities under both models. The sum of all token scores is compared against a threshold to determine if watermarking is present. The maximin variant enhances robustness by accounting for potential text modifications or generation parameter mismatches, controlled by a perturbation strength parameter  $d$ . The threshold is set to achieve a desired false positive rate  $\alpha$ .

### B.3.4 DiPmark (Wu et al., 2024)

DiPmark operates on a vocabulary set  $V$  with size  $N = |V|$  and employs a distribution-preserving reweight strategy that modifies token probabilities without changing the original distribution  $P_M(x_{n+1}|x_{1:n})$ .

**Watermarking** During generation, the system derives a texture key  $s_i$  from previously generated tokens and combines it with a secret key  $k$  to produce a cipher  $\theta_i = h(k, s_i)$  using a hash function  $h$ . This cipher represents a permutation of the vocabulary tokens. The core of DiPmark is its DiP-reweight strategy, defined as:

$$P_W(t_i|x, \theta) := (1 - \alpha)P_W^\alpha(t_i|x, \theta) + \alpha P_W^{1-\alpha}(t_i|x, \theta) \quad (6)$$

where  $\alpha \in [0, 1]$  is the reweight parameter and  $P_W^\alpha$  adjusts token probabilities within interval  $[0, \alpha]$

to 0 and scales others by  $\frac{1}{1-\alpha}$ . The next token is then sampled according to this modified distribution. This approach mathematically guarantees that  $E_\theta[P_W(t|x, \theta)] = P_M(t|x)$ , ensuring the original distribution is preserved across multiple generations, which distinguishes DiPmark from other watermarking approaches.

**Detection** The detection process uses a statistical test based on the "green token ratio" to identify watermarked text. For a given text sequence  $x_{1:n}$ , the detector computes the cipher  $\theta_i$  for each position using the same hash function and secret key. Each permutation  $\theta_i$  is divided using a green list separator  $\gamma \in [0, 1]$ , with the last  $(1 - \gamma)N$  tokens in the permutation designated as "green tokens."

The detector counts the number of green tokens  $L_G(\gamma)$  in the sequence and calculates the green token ratio:

$$\Phi(\gamma, x_{1:n}) := \frac{L_G(\gamma)}{n} - (1 - \gamma) \quad (7)$$

This ratio is compared against a threshold  $z$ ; if  $\Phi > z$ , the text is classified as watermarked.

### B.3.5 SIR (Liu et al., 2024a)

The Semantic Invariant Robust (SIR) watermarking method generates watermark logits based on the *semantics* of preceding tokens rather than their token IDs. Unlike traditional methods such as KGW that use hash functions of specific token sequences, SIR's semantic-based approach ensures that paraphrases or synonym substitutions—which preserve meaning but alter tokens—produce similar watermark logits and remain detectable.

**Watermarking** SIR generates watermark logits based on the *semantics* of all preceding tokens rather than their token IDs. The method uses an embedding language model (Compositional-BERT) to extract semantic embeddings from preceding tokens, then transforms these embeddings into watermark logits via a trained watermark model. The watermark model is a four-layer fully connected residual network trained to ensure that semantically similar texts produce similar watermark logits. The final watermark logits are processed to produce bimodal values close to  $\pm 1$ , analogous to KGW's red-green token concept, which are then added to the LLM's original logits with strength  $\delta$ .

**Detection** Detection uses a z-score test similar to KGW. For each token in the text, the corresponding watermark logit value is computed based on its preceding context, and these values are averaged.

The text is classified as watermarked if the average exceeds a threshold. Since semantically similar text modifications produce similar embeddings and thus similar watermark logits, the method maintains high detection accuracy even after paraphrasing or synonym substitution attacks.

### B.3.6 Default Watermarking Parameters

The default parameters used for each watermarking scheme are the default parameters used in Mark-LLM (Pan et al., 2024). The default parameters are as follows:

- **KGW**:  $\gamma = 0.5$ ,  $\delta = 2.0$
- **SWEET**:  $\gamma = 0.5$ ,  $\delta = 2.0$ , entropy threshold  $\tau = 0.9$
- **DiPmark**:  $\alpha = 0.45$ , green list ratio  $\gamma = 0.5$
- **Unbiased**: We use the  $\gamma$ -reweight variant.
- **SIR**:  $\delta = 1.0$ . We use the English-adapted implementation with Compositional-BERT-Large-Uncased as the embedding model.

### B.3.7 Watermarking Parameters for Increased Detectability

To increase detectability toward AUROC  $> 0.95$ , we tune parameters over the search spaces in Table 11, then select the best-performing values per document type (Table 12). We evaluate only KGW, SWEET, and SIR, as Unbiased and DIPMark remain insufficiently detectable even after tuning.

| Method   | Parameter Search Space                                              |
|----------|---------------------------------------------------------------------|
| KGW      | $\delta \in \{2, 3, 4, 5, 6, 7, 8\}, \gamma \in \{0.5, 0.25, 0.1\}$ |
| SWEET    | $\delta \in \{2, 3, 4, 5, 6, 7, 8\}, \gamma \in \{0.5, 0.25, 0.1\}$ |
| SIR      | $\delta \in \{1.0, 2.0, 3.0\}$                                      |
| DIP      | $\alpha \in \{0.35, 0.4, 0.45, 0.5\}$                               |
| Unbiased | $\gamma$ -variant, $\delta$ -variant                                |

Table 11: Parameter search spaces used for reproducibility tuning.

| Scheme | Nursing         | Discharge       | Progress        | Procedure       |
|--------|-----------------|-----------------|-----------------|-----------------|
| KGW    | $\delta = 6$    | $\delta = 5$    | $\delta = 5$    | $\delta = 7$    |
| SWEET  | $\delta = 8$    | $\delta = 8$    | $\delta = 8$    | $\delta = 8$    |
|        | $\gamma = 0.25$ | $\gamma = 0.25$ | $\gamma = 0.25$ | $\gamma = 0.25$ |
| SIR    | $\delta = 3$    | $\delta = 3$    | $\delta = 3$    | $\delta = 3$    |

Table 12: Selected parameter values for KGW, SWEET, and SIR to maximize detectability.

### B.4 Model Hyperparameters for Text Generation

Table 13 shows the hyperparameters used for all models.

| Hyperparameter   | Llama3 Med42 | Gemma 3 MedGemma |
|------------------|--------------|------------------|
| Quantization     | 4-bit NF4    | 4-bit NF4        |
| Compute dtype    | bfloat16     | bfloat16         |
| Attention impl.  | SDPA         | SDPA             |
| Vocabulary size  | 128,256      | 262,208          |
| Temperature      | 0.7          | 0.7              |
| Top-p            | 0.95         | 0.95             |
| No-repeat n-gram | 4            | 4                |
| Sampling         | True         | True             |
| Min new tokens   | 20           | 20               |
| Max new tokens   | 200          | 200              |

Table 13: Hyperparameters for text generation.

### B.5 LLM Judger Cost and Prompt

We report the total API cost for all experiments shown in Table 14.

The prompt used for LLM-Judger referred to the template of Singh and Zou (2024), where the aspects are tailored for helpfulness evaluation designed in Table 1. The example of prompt and response of LLM-Judger is available in Table 19.

| LLM Judge                  | API Cost (USD) |
|----------------------------|----------------|
| claude-sonnet-4-5-20250929 | 91.99          |

Table 14: API Cost for all experiments

### B.6 Use of Generative AI

We used LLMs for language editing to improve readability and minor stylistic refinement. We didn’t use LLMs for developing scientific content, ideas, or results, and their use aligns with standard manuscript preparation practices.

| LLAMA3                 | Detectability |             | Quality     |             |               |             |                   |             |              |              |
|------------------------|---------------|-------------|-------------|-------------|---------------|-------------|-------------------|-------------|--------------|--------------|
| Scheme                 | TPR           | AUROC       | Accuracy    |             | Lexical       | Syntactic   |                   | Readability |              |              |
|                        |               |             | BERT        | MTR         | YUL           | TTR         | LD <sub>Pos</sub> | PrD         | FRE          | GFI          |
| Procedure Note         |               |             |             |             |               |             |                   |             |              |              |
| Unwatermarked          | -             | -           | 0.88        | 0.40        | 264.77        | 0.65        | 0.92              | 1.52        | 56.59        | 10.53        |
| KGW                    | 0.07          | <b>0.86</b> | <b>0.88</b> | <b>0.40</b> | <b>286.21</b> | 0.64        | <b>0.77</b>       | 1.49        | <b>57.53</b> | 10.35        |
| SWEET                  | <b>0.09</b>   | 0.84        | <b>0.88</b> | 0.39        | 224.90        | <b>0.67</b> | 0.91              | 1.44        | 55.88        | 10.46        |
| DIP                    | 0.00          | 0.45        | <b>0.88</b> | 0.39        | 239.19        | <b>0.67</b> | 0.90              | 1.44        | 54.70        | 10.70        |
| Unbiased               | 0.00          | 0.50        | <b>0.88</b> | <b>0.40</b> | 240.42        | 0.66        | 0.88              | 1.47        | 55.29        | 10.65        |
| SIR                    | 0.05          | 0.80        | 0.87        | 0.36        | 247.88        | <b>0.67</b> | 0.95              | <b>1.37</b> | 55.27        | <b>10.16</b> |
| Progress Note          |               |             |             |             |               |             |                   |             |              |              |
| Unwatermarked          | -             | -           | 0.87        | 0.38        | 226.47        | 0.65        | 0.89              | 1.54        | 56.74        | 9.89         |
| KGW                    | 0.15          | 0.88        | <b>0.87</b> | 0.36        | <b>251.98</b> | 0.63        | <b>0.83</b>       | 1.55        | 57.04        | 10.07        |
| SWEET                  | <b>0.17</b>   | <b>0.91</b> | <b>0.87</b> | <b>0.37</b> | 221.57        | 0.65        | 0.91              | 1.66        | 57.21        | 9.85         |
| DIP                    | 0.00          | 0.50        | <b>0.87</b> | <b>0.37</b> | 229.15        | 0.65        | 0.91              | 1.59        | <b>58.03</b> | 9.91         |
| Unbiased               | 0.00          | 0.53        | <b>0.87</b> | 0.36        | 216.04        | 0.66        | 0.96              | 1.56        | 55.89        | 10.12        |
| SIR                    | 0.07          | 0.80        | <b>0.87</b> | 0.36        | 216.56        | <b>0.68</b> | 0.97              | <b>1.43</b> | 54.05        | <b>9.83</b>  |
| Nursing Note           |               |             |             |             |               |             |                   |             |              |              |
| Unwatermarked          | -             | -           | 0.88        | 0.38        | 301.20        | 0.64        | 0.92              | 1.33        | 65.85        | 7.89         |
| KGW                    | <b>0.27</b>   | <b>0.87</b> | <b>0.88</b> | 0.36        | <b>341.81</b> | 0.61        | 0.90              | 1.43        | <b>68.00</b> | <b>7.72</b>  |
| SWEET                  | 0.25          | 0.81        | <b>0.88</b> | 0.36        | 275.64        | <b>0.65</b> | 0.94              | <b>1.40</b> | 63.23        | 8.48         |
| DIP                    | 0.01          | 0.56        | <b>0.88</b> | <b>0.38</b> | 294.09        | 0.64        | <b>0.89</b>       | 1.42        | 66.15        | 7.89         |
| Unbiased               | 0.00          | 0.46        | <b>0.88</b> | 0.37        | 317.23        | 0.62        | 0.90              | 1.40        | 67.53        | 7.90         |
| SIR                    | 0.12          | 0.82        | 0.87        | 0.33        | 314.58        | <b>0.66</b> | 1.12              | <b>1.38</b> | 65.91        | 7.37         |
| Discharge Summary      |               |             |             |             |               |             |                   |             |              |              |
| Unwatermarked          | -             | -           | 0.85        | 0.32        | 237.52        | 0.65        | 1.04              | 1.53        | 49.55        | 11.39        |
| KGW                    | 0.24          | 0.86        | <b>0.85</b> | <b>0.34</b> | <b>239.85</b> | 0.63        | <b>0.92</b>       | 1.57        | <b>52.22</b> | <b>11.09</b> |
| SWEET                  | <b>0.32</b>   | <b>0.94</b> | <b>0.85</b> | 0.33        | 234.49        | 0.65        | 1.02              | 1.55        | 49.99        | 11.43        |
| DIP                    | 0.02          | 0.57        | 0.84        | 0.31        | 212.70        | <b>0.67</b> | 1.06              | 1.66        | 47.15        | 11.84        |
| Unbiased               | 0.00          | 0.51        | 0.84        | 0.30        | 215.39        | 0.66        | 1.08              | 1.61        | 46.70        | 12.03        |
| SIR                    | 0.18          | 0.89        | 0.84        | 0.29        | 239.15        | <b>0.67</b> | 1.22              | <b>1.40</b> | 45.64        | 11.13        |
| Medalign Summarization |               |             |             |             |               |             |                   |             |              |              |
| Unwatermarked          | -             | -           | 0.82        | 0.18        | 141.22        | 0.70        | 1.55              | 3.79        | 29.34        | 18.82        |
| KGW                    | 0.56          | <b>0.97</b> | <b>0.83</b> | <b>0.19</b> | 148.50        | 0.69        | <b>1.48</b>       | 3.24        | <b>39.46</b> | <b>15.61</b> |
| SWEET                  | <b>0.65</b>   | 0.97        | 0.82        | 0.18        | 147.09        | 0.68        | 1.51              | 3.82        | 32.40        | 18.02        |
| DIP                    | 0.00          | 0.54        | 0.82        | 0.17        | <b>156.06</b> | 0.68        | 1.50              | 3.65        | 33.64        | 17.61        |
| Unbiased               | 0.00          | 0.52        | <b>0.83</b> | <b>0.19</b> | 147.02        | 0.68        | 1.49              | 3.77        | 33.99        | 17.52        |
| SIR                    | 0.01          | 0.80        | 0.82        | 0.17        | 125.59        | <b>0.72</b> | 1.62              | <b>3.19</b> | 35.84        | 16.27        |

Table 15: Watermarking scheme evaluation on LLAMA3 across tasks. Metrics: **TPR** (True Positive Rate), **AUROC** (Area Under ROC Curve), **BERT** (BERTScore), **MTR** (METEOR), **YUL** (Yule’s K), **TTR** (Type-Token Ratio), **LD<sub>Pos</sub>** (Linguistic Distribution Distance of PoS), **PrD** (Parse tree Depth), **FRE** (Flesch Reading Ease), **GFI** (Gunning Fog Index).

Please rate the Generated Answer quality based on the EHR note and the Original answer (reference) below.

EHR Note #1:

Health Maintenance Due Topic Date Due 1 IPV IMMUNIZATIONS (1 of 4 - All-IPV series) 12/22/2006 1 MMR IMMUNIZATIONS (1 of 2 - Standard series) 10/22/2007 1 VARICELLA IMMUNIZATIONS (1 of 2 - 2-dose childhood series) 10/22/2007 1 DTap/Tdap/Td Immunizations (2 - Td) 02/17/2018 1 HPV IMMUNIZATIONS (2 of 2 - Male 2-dose series) 07/20/2018

Original Answer (reference):

IPV immunizations are due. // IPV immunizations are part of a 4-dose All-IPV series. // IPV immunizations are due on 12/22/2006. // MMR immunizations are due. // MMR immunizations are part of a 2-dose Standard series. // MMR immunizations are due on 10/22/2007. // Varicella immunizations are due. // Varicella immunizations are part of a 2-dose childhood series. // Varicella immunizations are due on 10/22/2007. // DTap/Tdap/Td immunizations are due. // DTap/Tdap/Td immunizations are part of a Td series. // DTap/Tdap/Td immunizations are due on 02/17/2018. // HPV immunizations are due. // HPV immunizations are part of a 2-dose Male series. // HPV immunizations are due on 07/20/2018.

Generated Answer A (evaluated text):

IPV Immunizations are due on 12/28/2006. // IPV Immunization is part of an all-IPV 4-series. // MMR Immunization is part of a standard series of 2. // Mmr Immunization date is 10/28/07. // VARICELLA Immunization was part of a 2-doses childhood series. // VARCELLA Immunization occurred on 10/29/2007. // DTap/TD Immunization is part a series of 3. // DTap/Tdap/TD Immunization dates are 02/28/2018. // HPV Immunization series is a 2-series for males. // HPV Immunization date is 07/28/18.

Quality ratings for A

Accuracy

Does the response **correctly convey the information** in the reference?

1 - Incorrect (major errors, misleading information, incorrect terminology)

2 - Partially Correct (minor errors, generally correct terminology)

3 - Fully Correct (accurate terminology, no misleading details)

Syntactic

Is the **grammar and sentence structure** well-formed and aligned with the reference style?

1 - Poor (confusing structure, different from reference)

2 - Acceptable (minor structural issues)

3 - Good (well-formed, structurally aligned to reference)

Readability

How easy is the response to **read and scan** while maintaining the clarity found in the reference?

1 - Hard to read (complex or unclear)

2 - Moderate

3 - Easy to read (clear, concise)

Helpfulness

Overall, how helpful is this generated response for **understanding the clinical information** as presented in the reference (the intended transformation of the EHR)?

1 - Not helpful at all

2 - Slightly helpful

3 - Moderately helpful

4 - Helpful

5 - Very helpful

Figure 8: Survey page for human (expert) evaluation comparing both watermarked and un-watermarked text.

58635

| MED42                  | Detectability |       | Quality  |      |         |           |                   |             |       |       |
|------------------------|---------------|-------|----------|------|---------|-----------|-------------------|-------------|-------|-------|
| Scheme                 | TPR           | AUROC | Accuracy |      | Lexical | Syntactic |                   | Readability |       |       |
|                        |               |       | BERT     | MTR  | YUL     | TTR       | LD <sub>Pos</sub> | PrD         | FRE   | GFI   |
| Procedure Note         |               |       |          |      |         |           |                   |             |       |       |
| Unwatermarked          | -             | -     | 0.86     | 0.28 | 98.03   | 0.80      | 1.08              | 2.16        | 37.71 | 14.51 |
| KGW                    | 0.04          | 0.84  | 0.86     | 0.27 | 99.25   | 0.79      | 1.07              | 2.43        | 36.18 | 14.80 |
| SWEET                  | 0.17          | 0.93  | 0.86     | 0.28 | 96.75   | 0.80      | 1.06              | 2.18        | 40.19 | 13.85 |
| DIP                    | 0.01          | 0.46  | 0.86     | 0.28 | 89.55   | 0.80      | 1.07              | 2.21        | 39.30 | 14.38 |
| Unbiased               | 0.00          | 0.51  | 0.86     | 0.26 | 100.58  | 0.81      | 1.10              | 2.24        | 40.47 | 14.03 |
| SIR                    | 0.05          | 0.87  | 0.85     | 0.27 | 96.94   | 0.82      | 1.18              | 1.53        | 43.45 | 12.76 |
| Progress Note          |               |       |          |      |         |           |                   |             |       |       |
| Unwatermarked          | -             | -     | 0.85     | 0.26 | 93.85   | 0.77      | 1.05              | 2.56        | 36.05 | 14.22 |
| KGW                    | 0.26          | 0.89  | 0.85     | 0.24 | 109.73  | 0.77      | 1.08              | 2.58        | 36.23 | 14.39 |
| SWEET                  | 0.55          | 0.94  | 0.85     | 0.26 | 98.78   | 0.76      | 1.09              | 2.72        | 40.59 | 14.38 |
| DIP                    | 0.01          | 0.50  | 0.85     | 0.25 | 90.39   | 0.78      | 1.12              | 2.81        | 33.84 | 15.09 |
| Unbiased               | 0.00          | 0.49  | 0.85     | 0.25 | 88.96   | 0.78      | 1.09              | 2.78        | 39.33 | 14.32 |
| SIR                    | 0.15          | 0.83  | 0.84     | 0.26 | 116.76  | 0.79      | 1.20              | 1.89        | 41.40 | 12.86 |
| Nursing Note           |               |       |          |      |         |           |                   |             |       |       |
| Unwatermarked          | -             | -     | 0.87     | 0.28 | 143.10  | 0.79      | 1.26              | 1.84        | 53.16 | 10.16 |
| KGW                    | 0.03          | 0.79  | 0.87     | 0.26 | 185.76  | 0.78      | 1.16              | 1.66        | 55.22 | 9.83  |
| SWEET                  | 0.15          | 0.80  | 0.87     | 0.26 | 172.40  | 0.79      | 1.16              | 1.88        | 54.20 | 10.55 |
| DIP                    | 0.00          | 0.57  | 0.87     | 0.26 | 164.75  | 0.80      | 1.20              | 1.72        | 54.23 | 10.37 |
| Unbiased               | 0.00          | 0.48  | 0.87     | 0.25 | 185.68  | 0.79      | 1.16              | 1.70        | 55.66 | 9.83  |
| SIR                    | 0.10          | 0.80  | 0.87     | 0.27 | 205.88  | 0.79      | 1.13              | 1.46        | 54.45 | 9.45  |
| Discharge Summary      |               |       |          |      |         |           |                   |             |       |       |
| Unwatermarked          | -             | -     | 0.83     | 0.21 | 75.95   | 0.79      | 1.14              | 2.63        | 32.45 | 15.46 |
| KGW                    | 0.60          | 0.91  | 0.82     | 0.19 | 89.65   | 0.77      | 1.26              | 2.83        | 33.23 | 16.19 |
| SWEET                  | 0.62          | 0.97  | 0.82     | 0.20 | 90.79   | 0.77      | 1.15              | 3.00        | 31.11 | 16.47 |
| DIP                    | 0.02          | 0.54  | 0.82     | 0.19 | 68.23   | 0.80      | 1.18              | 3.02        | 25.01 | 17.23 |
| Unbiased               | 0.00          | 0.42  | 0.82     | 0.19 | 66.81   | 0.81      | 1.24              | 2.65        | 27.87 | 16.67 |
| SIR                    | 0.10          | 0.90  | 0.82     | 0.20 | 95.77   | 0.80      | 1.39              | 2.51        | 33.49 | 15.90 |
| Medalign Summarization |               |       |          |      |         |           |                   |             |       |       |
| Unwatermarked          | -             | -     | 0.83     | 0.18 | 125.50  | 0.74      | 1.53              | 3.46        | 35.16 | 16.42 |
| KGW                    | 0.73          | 0.98  | 0.82     | 0.18 | 122.08  | 0.71      | 1.44              | 3.61        | 36.19 | 16.34 |
| SWEET                  | 0.93          | 0.99  | 0.82     | 0.18 | 129.18  | 0.70      | 1.55              | 3.52        | 34.39 | 16.63 |
| DIP                    | 0.01          | 0.42  | 0.82     | 0.18 | 113.68  | 0.73      | 1.52              | 3.49        | 32.83 | 16.64 |
| Unbiased               | 0.00          | 0.47  | 0.82     | 0.18 | 115.02  | 0.73      | 1.45              | 3.42        | 32.17 | 16.80 |
| SIR                    | 0.00          | 0.88  | 0.81     | 0.16 | 108.35  | 0.75      | 1.75              | 2.96        | 34.29 | 15.31 |

Table 16: Watermarking scheme evaluation on Med42 across tasks. See Table 15 for metric definitions.

| GEMMA 3                |      |       | Detectability |      | Quality |           |                   |             |       |       |
|------------------------|------|-------|---------------|------|---------|-----------|-------------------|-------------|-------|-------|
| Scheme                 | TPR  | AUROC | Accuracy      |      | Lexical | Syntactic |                   | Readability |       |       |
|                        |      |       | BERT          | MTR  | YUL     | TTR       | LD <sub>Pos</sub> | PrD         | FRE   | GFI   |
| Procedure Note         |      |       |               |      |         |           |                   |             |       |       |
| Unwatermarked          | -    | -     | 0.90          | 0.48 | 241.19  | 0.66      | 0.76              | 1.39        | 56.17 | 10.66 |
| KGW                    | 0.12 | 0.77  | 0.90          | 0.48 | 243.79  | 0.65      | 0.75              | 1.51        | 57.15 | 10.60 |
| SWEET                  | 0.01 | 0.73  | 0.90          | 0.47 | 239.98  | 0.66      | 0.78              | 1.37        | 56.62 | 10.61 |
| DIP                    | 0.03 | 0.51  | 0.90          | 0.48 | 230.07  | 0.66      | 0.76              | 1.40        | 56.01 | 10.76 |
| Unbiased               | 0.00 | 0.52  | 0.90          | 0.48 | 224.61  | 0.67      | 0.77              | 1.37        | 55.89 | 10.68 |
| SIR                    | 0.04 | 0.81  | 0.90          | 0.46 | 238.99  | 0.66      | 0.82              | 1.39        | 57.87 | 10.57 |
| Progress Note          |      |       |               |      |         |           |                   |             |       |       |
| Unwatermarked          | -    | -     | 0.88          | 0.42 | 184.31  | 0.67      | 0.81              | 1.62        | 52.13 | 10.93 |
| KGW                    | 0.12 | 0.86  | 0.88          | 0.41 | 188.82  | 0.66      | 0.83              | 2.00        | 48.92 | 12.27 |
| SWEET                  | 0.06 | 0.74  | 0.89          | 0.43 | 183.15  | 0.67      | 0.82              | 1.72        | 51.42 | 11.47 |
| DIP                    | 0.00 | 0.49  | 0.88          | 0.43 | 188.42  | 0.66      | 0.80              | 1.69        | 52.85 | 11.01 |
| Unbiased               | 0.00 | 0.53  | 0.88          | 0.43 | 181.43  | 0.67      | 0.77              | 1.64        | 53.20 | 10.89 |
| SIR                    | 0.07 | 0.79  | 0.88          | 0.41 | 198.11  | 0.65      | 0.83              | 1.68        | 52.26 | 10.98 |
| Nursing Note           |      |       |               |      |         |           |                   |             |       |       |
| Unwatermarked          | -    | -     | 0.90          | 0.47 | 285.64  | 0.63      | 0.77              | 1.31        | 69.24 | 7.76  |
| KGW                    | 0.10 | 0.77  | 0.89          | 0.46 | 311.05  | 0.62      | 0.75              | 1.30        | 69.12 | 7.80  |
| SWEET                  | 0.15 | 0.77  | 0.89          | 0.47 | 286.94  | 0.63      | 0.76              | 1.35        | 68.20 | 8.10  |
| DIP                    | 0.02 | 0.46  | 0.89          | 0.46 | 307.19  | 0.62      | 0.77              | 1.34        | 68.49 | 7.81  |
| Unbiased               | 0.00 | 0.52  | 0.89          | 0.47 | 297.75  | 0.62      | 0.79              | 1.35        | 68.84 | 7.78  |
| SIR                    | 0.01 | 0.82  | 0.89          | 0.43 | 295.93  | 0.62      | 0.84              | 1.44        | 68.00 | 8.09  |
| Discharge Summary      |      |       |               |      |         |           |                   |             |       |       |
| Unwatermarked          | -    | -     | 0.86          | 0.36 | 179.64  | 0.68      | 0.97              | 1.62        | 42.28 | 13.12 |
| KGW                    | 0.09 | 0.94  | 0.86          | 0.36 | 178.23  | 0.68      | 1.02              | 1.50        | 42.78 | 12.73 |
| SWEET                  | 0.22 | 0.81  | 0.86          | 0.36 | 173.81  | 0.68      | 0.95              | 1.53        | 42.52 | 12.99 |
| DIP                    | 0.00 | 0.56  | 0.86          | 0.37 | 177.05  | 0.68      | 0.96              | 1.66        | 41.75 | 13.56 |
| Unbiased               | 0.00 | 0.49  | 0.86          | 0.37 | 177.57  | 0.68      | 0.97              | 1.67        | 41.72 | 13.55 |
| SIR                    | 0.20 | 0.80  | 0.86          | 0.34 | 179.36  | 0.67      | 1.03              | 1.62        | 43.84 | 12.78 |
| Medalign Summarization |      |       |               |      |         |           |                   |             |       |       |
| Unwatermarked          | -    | -     | 0.82          | 0.18 | 123.00  | 0.75      | 1.42              | 2.76        | 33.23 | 16.09 |
| KGW                    | 0.38 | 0.95  | 0.82          | 0.18 | 128.87  | 0.73      | 1.36              | 2.88        | 33.66 | 16.27 |
| SWEET                  | 0.36 | 0.91  | 0.82          | 0.18 | 124.38  | 0.74      | 1.44              | 2.75        | 32.75 | 16.40 |
| DIP                    | 0.00 | 0.52  | 0.82          | 0.17 | 134.87  | 0.74      | 1.46              | 2.77        | 32.80 | 16.40 |
| Unbiased               | 0.00 | 0.51  | 0.82          | 0.18 | 119.92  | 0.75      | 1.45              | 2.82        | 32.25 | 16.71 |
| SIR                    | 0.00 | 0.75  | 0.82          | 0.18 | 142.58  | 0.72      | 1.43              | 2.99        | 33.08 | 16.74 |

Table 17: Watermarking scheme evaluation on Gemma 3 across tasks. See Table 15 for metric definitions.

| MEDGEMMA               | Detectability |       | Quality  |      |         |      |                   |      |             |       |
|------------------------|---------------|-------|----------|------|---------|------|-------------------|------|-------------|-------|
| Scheme                 | TPR           | AUROC | Accuracy |      | Lexical |      | Syntactic         |      | Readability |       |
|                        |               |       | BERT     | MTR  | YUL     | TTR  | LD <sub>Pos</sub> | PrD  | FRE         | GFI   |
| Procedure Note         |               |       |          |      |         |      |                   |      |             |       |
| Unwatermarked          | -             | -     | 0.88     | 0.42 | 270.81  | 0.66 | 0.88              | 2.11 | 49.98       | 13.14 |
| KGW                    | 0.39          | 0.90  | 0.88     | 0.41 | 315.77  | 0.64 | 0.84              | 1.61 | 54.39       | 11.76 |
| SWEET                  | 0.24          | 0.87  | 0.88     | 0.40 | 281.42  | 0.66 | 0.84              | 1.90 | 51.31       | 12.70 |
| DIP                    | 0.00          | 0.52  | 0.88     | 0.39 | 272.20  | 0.67 | 0.87              | 1.79 | 51.40       | 12.90 |
| Unbiased               | 0.00          | 0.44  | 0.88     | 0.41 | 265.89  | 0.66 | 0.88              | 1.87 | 50.51       | 13.01 |
| SIR                    | 0.17          | 0.85  | 0.88     | 0.37 | 291.53  | 0.63 | 0.97              | 2.00 | 49.86       | 13.91 |
| Progress Note          |               |       |          |      |         |      |                   |      |             |       |
| Unwatermarked          | -             | -     | 0.87     | 0.39 | 245.21  | 0.62 | 0.89              | 1.78 | 58.56       | 10.36 |
| KGW                    | 0.45          | 0.96  | 0.87     | 0.37 | 259.94  | 0.63 | 0.90              | 1.58 | 58.98       | 9.58  |
| SWEET                  | 0.47          | 0.94  | 0.87     | 0.36 | 253.43  | 0.61 | 0.92              | 1.78 | 59.55       | 10.34 |
| DIP                    | 0.01          | 0.52  | 0.87     | 0.37 | 225.95  | 0.64 | 0.95              | 1.81 | 57.05       | 10.43 |
| Unbiased               | 0.00          | 0.50  | 0.87     | 0.37 | 224.13  | 0.63 | 0.95              | 2.11 | 54.60       | 11.71 |
| SIR                    | 0.18          | 0.81  | 0.87     | 0.34 | 265.61  | 0.61 | 1.02              | 2.50 | 52.06       | 12.84 |
| Nursing Note           |               |       |          |      |         |      |                   |      |             |       |
| Unwatermarked          | -             | -     | 0.87     | 0.37 | 257.20  | 0.67 | 1.04              | 1.32 | 66.12       | 8.24  |
| KGW                    | 0.15          | 0.84  | 0.88     | 0.41 | 303.02  | 0.63 | 0.88              | 1.33 | 66.60       | 8.50  |
| SWEET                  | 0.30          | 0.89  | 0.88     | 0.39 | 284.17  | 0.62 | 0.92              | 1.28 | 65.25       | 8.12  |
| DIP                    | 0.01          | 0.41  | 0.87     | 0.37 | 218.04  | 0.68 | 1.09              | 1.28 | 63.48       | 8.40  |
| Unbiased               | 0.00          | 0.52  | 0.87     | 0.38 | 228.02  | 0.68 | 1.07              | 1.34 | 64.99       | 8.44  |
| SIR                    | 0.21          | 0.89  | 0.87     | 0.34 | 269.45  | 0.64 | 1.06              | 1.39 | 66.78       | 7.79  |
| Discharge Summary      |               |       |          |      |         |      |                   |      |             |       |
| Unwatermarked          | -             | -     | 0.86     | 0.34 | 232.09  | 0.65 | 1.11              | 2.42 | 41.64       | 14.88 |
| KGW                    | 0.79          | 0.98  | 0.86     | 0.35 | 285.69  | 0.61 | 1.03              | 1.75 | 52.16       | 11.79 |
| SWEET                  | 0.61          | 0.96  | 0.85     | 0.30 | 236.99  | 0.64 | 1.22              | 2.44 | 42.71       | 14.99 |
| DIP                    | 0.00          | 0.48  | 0.84     | 0.29 | 212.96  | 0.64 | 1.23              | 2.43 | 44.58       | 14.37 |
| Unbiased               | 0.00          | 0.56  | 0.85     | 0.32 | 225.39  | 0.65 | 1.19              | 2.01 | 48.38       | 12.80 |
| SIR                    | 0.41          | 0.84  | 0.85     | 0.28 | 250.79  | 0.62 | 1.47              | 3.59 | 33.13       | 18.99 |
| Medalign Summarization |               |       |          |      |         |      |                   |      |             |       |
| Unwatermarked          | -             | -     | 0.83     | 0.16 | 200.43  | 0.73 | 1.46              | 2.96 | 46.17       | 13.93 |
| KGW                    | 0.31          | 0.93  | 0.83     | 0.15 | 178.20  | 0.73 | 1.46              | 3.10 | 47.83       | 13.73 |
| SWEET                  | 0.33          | 0.86  | 0.83     | 0.15 | 202.15  | 0.74 | 1.38              | 3.02 | 46.12       | 14.14 |
| DIP                    | 0.01          | 0.54  | 0.83     | 0.15 | 173.69  | 0.76 | 1.46              | 3.12 | 45.64       | 14.16 |
| Unbiased               | 0.00          | 0.53  | 0.83     | 0.16 | 178.85  | 0.75 | 1.42              | 2.90 | 46.49       | 13.69 |
| SIR                    | 0.06          | 0.78  | 0.83     | 0.17 | 218.18  | 0.70 | 1.34              | 3.20 | 43.08       | 15.21 |

Table 18: Watermarking scheme evaluation on MedGemma across tasks. See Table 15 for metric definitions.

Table 19: LLM-Judger prompt and example response set for Fact Decomposition on Nursing Note.

| PROMPT                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | JUDGER RESPONSE                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Please act as an impartial judge and evaluate the quality of the LLM-generated response that transform medical notes into atomic facts. Assess each response according to the criteria outlined, using a 1-3 Likert scale for Accuracy, Lexical, Syntactic, Redability, and 1-5 Likert score for Helpfulness.</p> <p>Some response are truncated: Due to length limitations, you will only see approximately 100 words max of each machine-generated output. This means some facts from the original EHR note may not appear in the visible portion.</p> <p>Criteria:</p> <ol style="list-style-type: none"> <li>1. Accuracy. Does the response correctly convey the information in the reference? [1-incorrect, 2-partially correct, 3-fully correct]</li> <li>2. Lexical. Does the wording choice follow the vocabulary in the reference? [1-Poor, 2-acceptable, 3-good]</li> <li>3. Syntactic. Is the grammar and sentence structure well-formed and aligned with the reference style? [1-Poor, 2-acceptable, 3-good]</li> <li>4. Readability. How easy is the response to read and scan while maintaining the clarity found in the reference? [1-hard to read, 2-moderate, 3-easy to read]</li> <li>5. Helpfulness. Overall, how helpful is this generated response for understanding the clinical information as presented in the reference (the intended transformation of the EHR) [1-not helpful at all, 2-slightly helpful, 3-moderate helpful, 4-helpful, 5-very helpful]</li> </ol> <p>After scoring each criterion, provide a short summary for each response, including specific examples that influenced your scoring. Additionally, don't let the length of the responses influence your evaluation. Be as objective as possible and ensure that the order in which the responses are presented does not affect your decision.</p> <p>Start with a brief statement about which response you think is superior. Then, for each response and criterion, provide a score, followed by a brief justification for that score. At the very end of your response, declare your verdict by choosing one of the choices below, strictly following the given format:</p> <p>[[A]]: [list of scores for LLM A output, in order of Accuracy, Lexical, Syntactic, Readability, Helpfulness]<br/> [[B]]: [list of scores for LLM B output, in order of Accuracy, Lexical, Syntactic, Readability, Helpfulness]</p> <p>[Medical Note]<br/> [LLM A's Answer]<br/> [LLM B's Answer]</p> | <p>Looking at these two responses, Response A is superior as it transforms the medical note into readable, complete atomic facts, while Response B appears to be a fragmented parsing that's difficult to interpret.</p> <p>Response A Evaluation:</p> <p>Accuracy: 2 (Partially correct) - Contains one significant error: Blood pressure is recorded as 140/80 instead of 140/86 - All other clinical values are accurately transcribed (pulse 69, SpO2 97%, fall risk negative) - Correctly captures chief complaint elements</p> <p>Lexical: 3 (Good) - Uses appropriate medical terminology consistently - Expands abbreviations appropriately (e.g., "SpO2 is 97 percent") - Vocabulary choices align well with the reference note</p> <p>Syntactic: 3 (Good) - Complete, grammatically correct sentences throughout - Clear structure: "Blood Pressure is 140/80", "Pulse is 69" - Natural sentence flow while maintaining clinical precision</p> <p>Readability: 3 (Easy to read) - Facts are clearly delineated with "/" separators - Each statement is a complete, standalone fact - Easy to scan and comprehend quickly</p> <p>Helpfulness: 4 (Helpful) - Despite the BP error, the response successfully transforms the note into digestible atomic facts - Clinical information is accessible and understandable - Format facilitates quick information retrieval</p> <p>Response B Evaluation: Accuracy: 2 (Partially correct) - Blood pressure correctly recorded as 140/86 - Has a typo: "irst Site" instead of "first Site" - Values are present but context is stripped away, making verification difficult . . .</p> <p>:</p> <p>[[A]]: [2, 3, 3, 3, 4]<br/> [[B]]: [2, 1, 1, 1, 2]</p> |