Marking the Wrong Symptoms: Evaluating LLM Watermarks in Medical Texts

Melanie Rieff[1,*] **Robin Staab**[1] **Thibaud Gloaguen**[1]
Stefan Hegselmann[2,3] **Martin Vechev**[1]

[1]Department of Computer Science, ETH Zurich, Switzerland
[2]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Center of Digital Health, Berlin, Germany
[3]Deutsches Herzzentrum der Charité – Medical Heart Center of Charité and German Heart Institute Berlin, Berlin, Germany
[*]Corresponding author: mrieff@student.ethz.ch

Abstract

Large language models (LLMs) are increasingly integrated into clinical workflows, stressing the need for reliable traceability of model-generated output with watermarking. Yet, most watermarks are evaluated on general-purpose benchmarks, leaving domains like medicine, where small token-level perturbations can result in significant semantic changes, under-explored. In this work, we present the first rigorous study of how LLM watermarks affect medical performance, benchmarking 5 watermarking schemes across 11 LLMs and 7 VLMs on various tasks spanning unimodal and multimodal clinical reasoning. Importantly, we complement existing evaluations by introducing a human-expert-validated pipeline for systematically auditing medical reasoning quality, terminological precision, and induced hallucinations. Our results reveal that watermarking can induce substantial degradation across multiple failure modes, including lexical corruption, hallucinated terminology, and amplified misattribution or omission of image findings. Notably, we find that the absence of domain-specific analyses, combined with aggregate metrics that miss failures inherent to clinical text, can systematically obscure practical watermark-induced degradations. Our findings establish domain-specific evaluation as a prerequisite for the safe deployment of watermarked models in medicine, where current benchmarks can otherwise mask clinically consequential failures.

1 Introduction

Large language models (LLMs) are increasingly integrated into clinical workflows: 81% of U.S. physicians now report using AI professionally [1], and LLM-based tools are actively explored for report generation, clinical decision support, and patient communication [2, 3]. In particular, LLM outputs inform downstream judgement, from summarising patient histories and drafting reports to suggesting differential diagnoses. As model capabilities improve, the scope of clinical reliance is likely to grow, making the quality of generated text a matter of patient safety. In parallel, emerging [4] regulation is pushing toward provenance marking of AI-generated content, with watermarking becoming the leading compliance mechanism for tracing LLM outputs [5]. Crucially, this implies that LLMs used in clinical settings are increasingly likely to be using a watermark.

LLM Watermarking Watermarking embeds a statistically detectable signal by modifying the LLM sampling procedure at every generation step, which enables traceability but, by design, directly influences the model's output distribution. Importantly, prior works evaluate the impact of watermarks on output quality with general-domain benchmarks, and coarse quality proxies such as perplexity. In

Preprint.

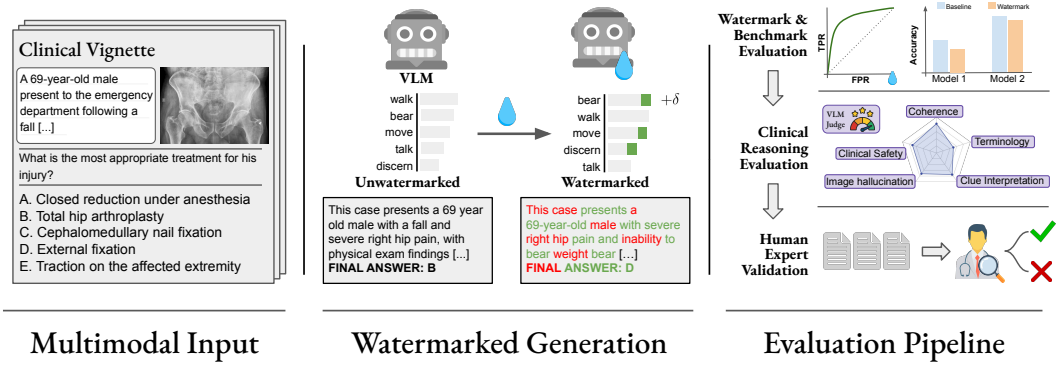


Figure 1: **Overview of our evaluation pipeline.** Given a multimodal benchmark instance, we prompt VLMs (using chain-of-thought prompting) with and without watermarks for reference. We first evaluate the detectability of watermarks in the answers and their accuracy. We then examine the reasoning traces in more depth using three specific LLM-as-judges, validated by clinical experts.

medicine, where correctness depends jointly on precise reasoning, terminology, qualifiers, and numerical values, even small token-level perturbations can alter clinical meaning significantly, suggesting that results from those simple evaluations might not transfer [6–8]. This makes understanding how watermarks alter LLMs' clinical performance a critical, yet under-explored issue.

This work In this work, we propose the first systematic study of how text watermarking affects medical LLM performance, illustrated in Figure 1. In particular, we evaluate 5 prominent watermarks on two medical benchmarks, mixing clinical knowledge and visual understanding (Figure 1 left). To measure the impact of watermarking on reasoning, we introduce (in Sec. 3.2) three specific LLM-as-judges validated against clinical expert annotations (Figure 1 right).

Our results reveal that watermarking can leave benchmark accuracy largely intact while substantially degrading the underlying reasoning: the rate of correct answers backed by flawed reasoning more than doubles on several models (e.g. 11.4%→26.3% on PHI-4-14B with distortionary SynthID), and fabricated medical entities increase by up to +39.2 pp. Among questions where both watermarked and unwatermarked outputs select the correct answer, up to 40% rely on mutually exclusive diagnostic statements. On multimodal inputs, watermarks increasingly corrupt visual interpretation, and for reasoning models, watermarking the reasoning trace inflates output length by up to 69% while degrading reasoning quality, whereas restricting the watermark to the final answer preserves both.

Key Contributions Our main contributions are:

- A systematic evaluation of five watermarking schemes across 11 LLMs and 7 VLMs on medical benchmarks (MedQA, MedXpertQA-MM), with *physician-validated* LLM-as-a-judge audits that assess reasoning quality, terminological precision, and hallucinations beyond letter accuracy.
- The first analysis of how text-decoding watermarks interact with multimodal medical input, revealing that watermarking during text generation degrades visual grounding, e.g., suppressing correct image-based claims and amplifying contradicted ones, despite accuracy remaining stable.
- A study of reasoning model watermarking, showing that restricting watermarks to the user-visible final answer preserves reasoning quality, while watermarking the thinking degrades terminological precision and induces circular deliberation.

2 Related Work

We review current evaluation practices for LLMs in medicine, the main families of text watermarking, and existing work on watermark quality assessment, highlighting the gap that motivates our study.

LLMs in medicine. General-purpose and domain-specialized LLMs are commonly evaluated on USMLE-style benchmarks such as MedQA [9, 10], while multimodal benchmarks [11–13] extend

evaluation to joint interpretation of images, labs, and patient records. Crucially, benchmark accuracy alone is an insufficient measure of clinical quality: models may arrive at the correct answer through flawed reasoning [14], produce confident hallucinations [15], or misuse medical terminology while maintaining a fluent narrative. These observations have motivated richer evaluation frameworks, from structured reasoning rubrics [16] to physician-graded benchmarks [17], and underscore that even small prompt variations can substantially change medical output quality [8]. Our work builds on this line by introducing 3 LLM-as-judges, validated against two board-certified physicians, specifically designed to audit watermark-induced degradation along the quality axes that matter in clinical practice.

Watermarking. Generation-time text watermarking modifies the sampling procedure to embed statistically detectable signals tied to a secret key, enabling traceability of AI-generated content [18]. At each token position, a context-dependent hash seeds a pseudorandom score assignment over the vocabulary, and the sampling distribution is modified accordingly. *Distortionary* methods such as KGW [19] and PPL [20] bias next-token distributions toward high-scoring tokens, introducing a systematic shift away from the model's original distribution. *Distortion-free* methods such as DipMark [21], AAR at zero strength [22], and SynthID-Text with two leaves per tournament [23] instead resample tokens in a way that preserves the original model distribution in expectation over the random watermark key. However, this guarantee is an *expectation* over keys, and under the assumption of no hash collisions. We find that, in practice, even distortion-free schemes can induce some small reasoning degradation (Sec. 4.2).

Evaluating watermark quality. Watermarking studies typically report detection power alongside coarse quality proxies such as perplexity or brief LLM ratings [19, 23, 24]. Yet, these proxies are poorly aligned with human preference [6] and mask substantial, task-dependent degradation: instruction-following quality drops considerably under matched watermark strength [6], smaller models suffer disproportionately [25], and reasoning-intensive tasks degrade more severely than open-ended generation [26, 27]. These findings establish that the impact of watermarking is model- and task-specific, yet most evaluations are confined to general-domain benchmarks, leaving a twofold gap: (i) the tasks studied do not capture the demands of clinical text, where a single-token substitution can alter a diagnosis, and (ii) the quality proxies employed cannot detect failure modes such as terminology corruption, hallucinated findings, or inconsistency between reasoning and conclusion. Separately, recent work has shown that watermarks can be applied to reasoning traces [28, 27], yet no study has assessed whether such application preserves clinical reasoning quality. One prior study attempts to evaluate watermarking on medical text [29]. Yet, we find that it suffers from fundamental methodological flaws, which makes its findings unreliable. Namely, it uses instruction-tuned models without chat templates, caps generation at 25 tokens, too short for reliable watermark detection, and fails to establish a paired unwatermarked baseline from which to measure hallucinations.

Our work introduces, for the first time, a rigorous and systematic evaluation of watermarks in medicine, spanning five watermarking schemes, multiple model scales and domain specializations, and text-decoding watermarks applied to models that process multimodal medical inputs. Importantly, we go beyond accuracy and introduce 3 LLM-as-judges, validated by clinical experts, to probe reasoning quality, lexical precision, and diagnostic consistency.

3 Our Evaluation Pipeline

In this section, we introduce our evaluation pipeline illustrated in Figure 1. We describe the models, watermarks, and data we evaluate (Sec. 3.1), and the metrics we use (Sec. 3.2). We defer additional details on the clinician validation of our LLM judges to App. D.2.

3.1 Setup

Our evaluation spans two medical benchmarks probing complementary aspects of clinical knowledge, 18 language and vision–language models covering general-purpose and medical-specialised architectures at 4B–70B scale, and five prominent watermarking schemes. Together these define the (model, watermark, strength) configurations we evaluate throughout the next sections.

Benchmarks. We anchor our evaluation with two medical benchmarks, each of them chosen to probe distinct aspects of medical knowledge and reasoning.

MedQA [9] contains 2,000 four-option, multiple-choice questions drawn from the United States medical licensing examinations. A meta-analysis by Kim and Yoon [10] identifies MedQA as the benchmark most predictive of real-world clinical performance among existing medical QA datasets.

MedXpertQA-MM [13] comprises 2,000 expert-curated, five-option questions spanning 17 medical specialties. Each question is paired with one to six images (on average 1.43, across 10 modalities including radiology, pathology, and schematic diagrams), requiring joint interpretation alongside clinical vignettes and structured patient data such as laboratory results.

Models. We evaluate three groups of models: On *MedQA*, we include six instruction-tuned LLMs: 4 general-purpose models (LLAMA-3.1-8B, LLAMA-3.1-70B [30], GEMMA-3-12B [31], and PHI-4-14B [32]), and 2 medical-specialised models (OPENBIOLLM-70B [33], and ULTRAMEDICAL-70B [34]), as well as 4 reasoning models (DEEPSEEK-R1-DISTILL-LLAMA-8B, DEEPSEEK-R1-DISTILL-LLAMA-70B, DEEPSEEK-R1-DISTILL-QWEN-32B [35], and QWEN-3.5-27B [36]). On *MedXpertQA-MM*, we evaluate 7 VLMs spanning 4B–32B parameters: QWEN-3-VL-8B and QWEN-3-VL-32B [37], GEMMA-4-31B [38], LINGSHU-7B and LINGSHU-32B [39], and MEDGEMMA-4B and MEDGEMMA-27B [40]. In particular, LINGSHU-7B, LINGSHU-32B, MEDGEMMA-4B, and MEDGEMMA-27B are medical-specialised VLMs and were not trained on MedXpertQA.

Watermarking schemes. Five prominent generation-time watermarking schemes are evaluated, spanning the major algorithmic families: KGW [19] (soft logit bias), DipMark [21] (CDF reweighting), AAR [22] (Gumbel-max sampling), PPL [20] (perplexity-budgeted argmax), and SynthID [23] (tournament sampling). All five share a common structure: at each token position, a context-dependent hash assigns pseudorandom scores to vocabulary entries, and the sampling distribution is modified accordingly. The schemes fall into two categories. *Distortion-based* methods (KGW, PPL at high strength) bias next-token distributions toward high-scoring tokens, introducing a systematic shift away from the model's original distribution. *Distortion-free* methods (DipMark, AAR at zero strength, SynthID with binary tournaments) preserve the original distribution in expectation over the random watermark key. For each scheme, we sweep multiple strength levels to evaluate their detectability–quality trade-off. We describe each scheme in-depth in App. A.1.

3.2 Evaluation Dimensions

Each (model, watermark, strength) configuration is scored along two complementary fronts: (1) baseline metrics that capture the watermark strength and benchmark utility, and (2) structured audits that probe whether watermarking corrupts underlying medical reasoning. The audits are necessary as benchmark accuracy fails to detect fabricated entities, misapplied terminology, or distorted reasoning.

Baseline metrics. We sample each (model, watermark, strength) configuration on the full 2,000-question benchmark using chain-of-thought prompting (with prompts in App. F.1). For each completion, we measure both the *accuracy* and the watermark *detectability*. For watermark detectability, we measure the average true positive rate (TPR) at a 1% false positive rate (FPR) over the 2,000 answers.

Reasoning-quality LLM-Judge. A model may select the correct answer while relying on fabricated entities or misapplied terminology, a failure mode documented in prior work on medical LLMs [14, 15] and one that standard accuracy metrics cannot detect. Letter accuracy alone therefore cannot show whether watermarking degrades medical reasoning. We complement accuracy with three structured LLM-as-judges, using GEMINI-3-FLASH as the judge for its balance of cost and reasoning capability.

1. *Single-response audit* evaluates each completion in isolation: the judge sees one response at a time, blind to both watermark condition and gold answer, and flags eight axes (fabricated entities, corrupted spellings, and misapplied real terms (terminology), confident error, vignette fabrication, clue misread, within-response contradiction, linguistic corruption, and format/termination failure) plus an abstention indicator. For MedXpertQA-MM, the judge additionally classifies each perceptual image-grounding claim as SUPPORTED, CONTRADICTED, or UNVERIFIABLE.
2. *Pairwise divergence audit* presents the watermarked and unwatermarked completions for the same question side by side to the judge, with response order randomised and condition

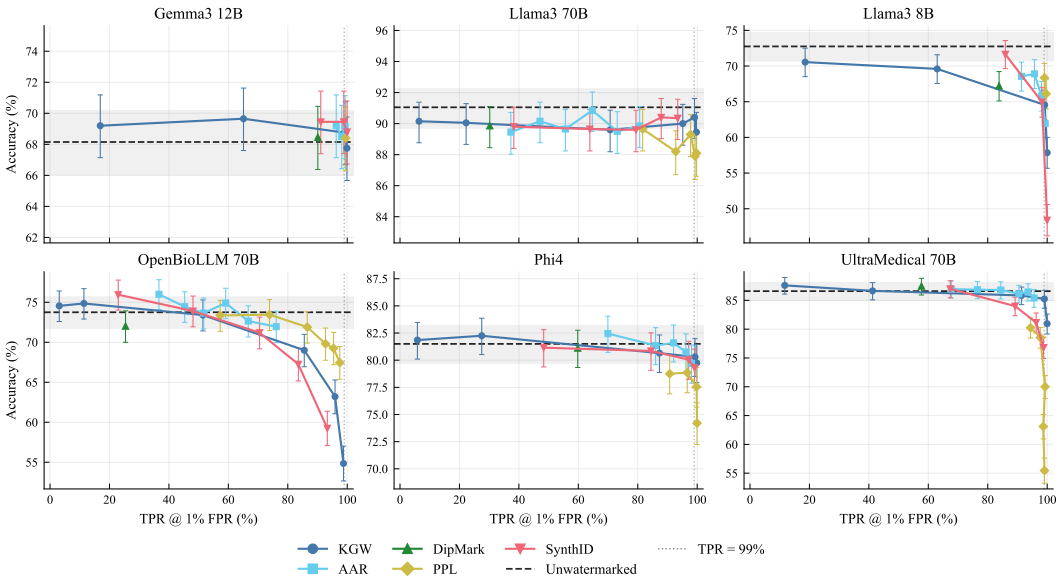


Figure 2: MedQA accuracy vs. detectability (TPR@1% FPR), non-reasoning LLMs across five schemes. Dashed line and grey band: unwatermarked mean and 95% CI; dotted vertical: TPR = 99%. LLAMA-3.1-70B, GEMMA-3-12B, and PHI-4-14B stay within the band across all schemes; the two specialised 70Bs collapse under complementary schemes; LLAMA-3.1-8B degrades broadly.

labels stripped. The judge records the correctness configuration of the pair, the level of diagnostic interpretation divergence (NONE, MINOR, MAJOR), and any scenario fabrication.

3. *Reasoning-trace audit* targets thinking models, whose hidden chain-of-thought may be independently distorted by the watermark. We use a two-block judging protocol. First, without seeing the gold answer, the judge rates four aspects of reasoning quality: reasoning efficiency, clinical coherence, alignment between the reasoning and the final answer, and engagement with distractors. Second, with the gold answer, the judge rates two factual aspects: terminological precision and diagnostic accuracy. The judge also provides two overall clinical-deployment ratings, *clinical safety* and *supervisory burden*, as well as the primary degradation pattern, defined in App. F.2. To separate degradation in the reasoning trace from degradation in the final answer, we run two paired comparisons: (i) *full watermark vs. final-answer-only watermark* and (ii) *final-answer-only watermark vs. no watermark*.

Importantly, we validate our judges with physician agreement in Sec. 4.3 and App. D. For each configuration, we conduct judge audits at the watermark hyperparameters that yield the lowest strength while still achieving $\geq 99\%$ TPR at 1% FPR, which prior work has considered the operating point required for regulatory deployment. We defer the remaining judge parameters (e.g., full prompts, axis definitions, calibration examples, and decoding configurations) to App. F.2.

4 Results

In this section, we evaluate how prominent watermarks affect the clinical performance of LLMs and validate our judges. We show that even though pure benchmark accuracy may remain high (Sec. 4.1), watermarks can still degrade reasoning quality (Sec. 4.2). We then validate our judges in Sec. 4.3. Lastly, in Sec. 4.4, we evaluate watermarks in the context of reasoning models.

4.1 Accuracy Looks Mostly Stable, But the Picture is Misleading

Figures 2 and 3 show that, for most schemes and models, accuracy remains within the 95% confidence interval (CI) of the unwatermarked baseline. On MedQA (Figure 2), three configurations show a visible drop: the medical-specialised 70Bs (ULTRAMEDICAL-70B and OPENBIOLLM-70B) at high TPR@1%FPR, and LLAMA-3.1-8B more broadly. That two equally large medical-specialised

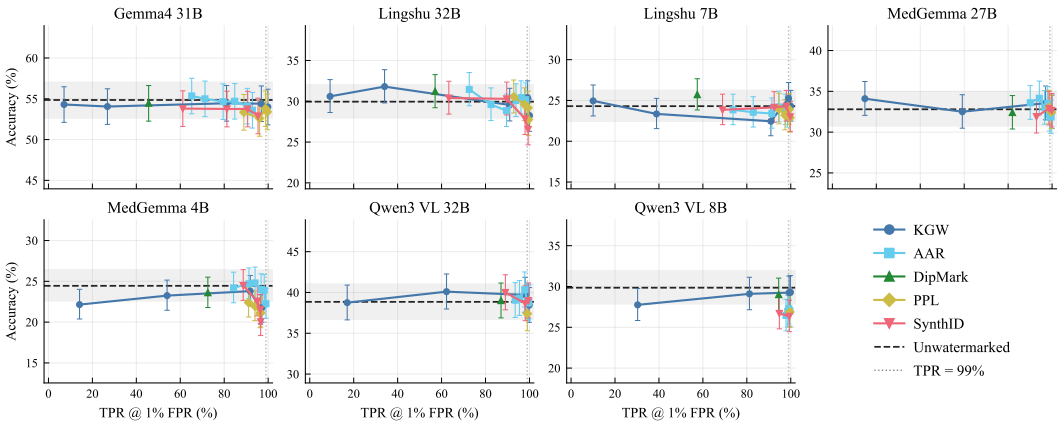


Figure 3: MedXpertQA-MM accuracy vs. detectability, seven VLMs; conventions as in Figure 2. Models consistently stay near the baseline band despite reasoning varying noticeably (Sec. 4.2).

models break at high TPR@1%FPR, while the equally large general-purpose LLAMA-3.1-70B remains consistently stable, rules out scale or domain specialisation as sufficient explanations.[1] On MedXpertQA-MM (Figure 3), all VLMs stay within or near the CI. We find in App. B.1 that although perplexity rises monotonically with watermark strength on every model, it does not correlate with drops in accuracy or reasoning degradation. Hence, a standard accuracy benchmark would conclude that watermarking is safe for medical use. Our audits in Sec. 4.2 show that this conclusion is incorrect: even where accuracy holds, watermarking corrupts the reasoning that clinicians would actually read.

4.2 Reasoning Quality Degrades Even When Accuracy Holds

We score every (model, scheme) configuration at the $\geq 99\%$ TPR operating point using the judge audits of Sec. 3.2. We bootstrap all our results, and our p-values quantify the significance between watermarked and unwatermarked results. We show the two most-affected models per benchmark in Table 2, and defer full results to App. C.1.

Correct answers hide broken reasoning. Watermarking increases the rate at which models reach the right MCQ letter through flawed reasoning. The *faulty-but-correct* rate (F&C) (i.e., correct letter with at least one flagged defect) climbs from 11.4% to 26.3% on PHI-4-14B under SynthID while accuracy drops only -3.2 pp, and similarly on OPENBIOLLM-70B F&C more than doubles ($12.9\% \rightarrow 27.5\%$, *). However, F&C only points out the existence of inconsistencies within the reasoning, but does not explain exactly what the inconsistencies are.

Watermarking silently corrupts medical language. Even when accuracy and F&C barely move, watermarks inject three categories of medical-language errors that standard metrics miss: *fabricated entities*, *misapplied real terms*, and *corrupted spellings* (Table 1). On MEDGEMMA-4B under SynthID, ΔF&C is only $+0.4$, yet the model fabricates $+12.6$ extra entities per 100 questions and misapplies $+24.3$ real terms; on LINGSHU-7B under SynthID, fabrications reach $+39.2$ per 100 questions. Fabricated and misapplied terms are especially dangerous: they produce fluent, confident text that requires domain expertise to catch and that no spell-checker or final-letter check would flag.

Each scheme damages a different axis. We find in Table 7 that watermarking is not generic noise: each scheme damages a different axis. PPL drives surface corruption. SynthID and KGW instead inflate fabrications and term misuse, with vignette-fabrication ratios up to $18\times$ (PHI-4-14B/PPL) and $7.6\times$ (PHI-4-14B/SynthID). AAR produces broad small-to-moderate effects. DipMark causes the least damage, consistent with its distortion-free guarantee [21]. Yet distortion-free does not mean benign: the guarantee holds in expectation over the random key and without hash collisions. We find that, at a fixed key, PHI-4-14B under DipMark still shows a statistically significant increase in F&C.

[1] BIOMISTRAL-7B also collapses (KGW -37.5, SynthID -14.2 pp), but the signal is confounded by poor instruction tuning on the CoT-MCQ template (median ~ 6 words per response vs. 250–450 for other models); we exclude it from the headline claims and report it only in the appendix grids.

Table 1: **Illustrative watermark-induced terminology errors** drawn from our judge audits and paired qualitative reviews of watermarked outputs. Matched unwatermarked completions on the same questions use the correct terms.

Error type	Watermarked Term	Correct Term / Explanation
Fabricated entity	*Streptococcus epidermidis*	*Staphylococcus epidermidis*
	Pentaprazole	Pantoprazole
	Naftifloxacin	No such antibiotic.
	plutonium channels	potassium channels; plutonium is a radioactive element.
Misapplied real term	Eosinophilic granuloma applied to the brown tumors of hyperparathyroidism	Eosinophilic granuloma denotes Langerhans cell histiocytosis–unrelated to the lytic lesions of hyperparathyroidism.
	"β-blockers preferred over CCBs in COPD"	Reversed: β-blockers risk bronchospasm; non-dihydropyridine CCBs are safe in COPD.
	Benzodiazepines "increase the *duration* of GABA-channel opening"	Benzodiazepines increase the channel-opening *frequency*; *duration* is the mechanism of barbiturates.
Corrupted spelling	*Clostridioides diffidile*	Misspelling of *C. difficile*.
	meningritis	Misspelling of meningitis.
	bradiceardia	Misspelling of bradycardia.

Table 2: **Reasoning damage at** 99% **TPR**, for two most-affected models per benchmark with full results in App. C.1. We report the accuracy (ΔAcc), the percentage of fabricated facts (Fab/100), the percentage of misapplied medical terms (Misapp/100), the percentage of corrupted spellings (Spell/100), the vignette fabrication ratio (VigFab), the faulty-but-correct rate (F&C), and the percentage of paired responses with major diagnostic divergence (Major%). For the unwatermarked row we report the actual value, whereas for watermarked rows we report the difference/ratio with the unwatermarked baseline. We **bold** entries that are statistically significant (p-value below 0.01).

Model	Scheme	ΔAcc	Fab/100	Misapp/100	Spell/100	VigFab	F&C	Major%
MedQA								
PHI-4-14B	Unwatermarked	81.4	0.3	4.1	0.1	1.0×	11.4	–
	AAR	-2.5	**+1.8**	**+4.0**	**+1.5**	**14.0**×	**+6.5**	1.9
	DipMark	-0.9	+0.1	+2.6	+0.3	0.8×	**+3.6**	2.8
	KGW	-0.7	**+1.5**	**+3.6**	**+2.2**	**3.0**×	**+9.5**	3.6
	PPL	**-8.6**	**+3.8**	+4.3	**+69.6**	**18.0**×	**+9.6**	2.3
	SynthID	**-3.2**	**+6.6**	**+8.0**	**+10.6**	**7.6**×	**+14.9**	3.3
OPENBIOLLM-70B	Unwatermarked	73.6	0.0	4.5	0.6	1.0×	12.9	–
	AAR	-0.2	**+0.9**	+1.5	+0.9	0.9×	**+4.4**	4.4
	DipMark	-1.1	+0.4	+0.8	+0.3	1.1×	+1.5	3.2
	KGW	**-17.4**	**+1.9**	**+4.4**	**+5.2**	**3.5**×	**+4.2**	5.6
	PPL	**-7.7**	**+2.1**	+1.8	**+51.5**	1.9×	**+6.7**	6.6
	SynthID	**-11.4**	**+13.6**	**+6.9**	**+25.9**	**6.6**×	**+14.6**	8.7
MedXpertQA-MM								
MEDGEMMA-4B	Unwatermarked	24.1	1.9	11.1	1.0	1.0×	16.5	–
	AAR	-1.0	**+6.5**	+8.1	**+3.4**	**3.1**×	+1.2	38.3
	DipMark	+0.2	+0.7	+2.0	+0.2	1.2×	+1.6	35.0
	KGW	-2.6	**+8.9**	+16.2	**+15.4**	**3.1**×	+0.6	31.7
	PPL	-1.7	**+6.0**	+5.0	**+43.6**	**2.1**×	0.0	34.7
	SynthID	-3.1	**+12.6**	**+24.3**	**+18.0**	**4.2**×	+0.4	36.0
LINGSHU-7B	Unwatermarked	23.9	1.5	12.1	0.4	1.0×	13.7	–
	AAR	-0.5	**+12.9**	+18.1	**+5.4**	**2.8**×	+3.7	33.0
	DipMark	+1.0	+0.3	+2.7	+0.1	1.0×	+1.4	22.6
	KGW	-0.8	**+13.2**	+2.4	**+16.2**	**3.6**×	+6.0	42.9
	PPL	-0.6	**+3.9**	+11.5	**+37.9**	**2.4**×	+2.3	20.0
	SynthID	-2.1	**+39.2**	**+20.2**	**+18.9**	**7.0**×	**+5.5**	30.5

Watermarks corrupt visual reporting. Here, we find that watermark-induced degradation may change what models claim to see. On MedXpertQA-MM, we track supported and contradicted visual claims per question against the provided image (Figure 4). SUPPORTED claims fall and CONTRADICTED claims rise simultaneously in most configurations with medical VLMs. For the larger general-purpose VLMs, we find in App. C no significant change.

Domain-specialised and smaller models are most vulnerable. At matched scale and TPR@1%FPR, general-purpose models are less degraded by watermarking than medical-specialised

or small models. LLAMA-3.1-70B is stable under all five schemes while the two same-size medical 70Bs each degrade on at least one scheme. We find in App. C.1 a similar split among 30B VLMs (GEMMA-4-31B vs. LINGSHU-32B). Yet, at higher watermark distortion, even LLAMA-3.1-70B's F&C nearly doubles and GEMMA-4-31B's vignette fabrication reaches 6×.

Harm grows with strength, but the trajectory is scheme-specific. We find in App. C and Figure 7 that the quality-detectability trajectory (i.e., how medical degradation varies with each scheme's detectability) is scheme-specific. For instance, KGW degrades smoothly with detectability whereas SynthID's quality matches the unwatermarked baseline up to 90% TPR@1%FPR and then degrades.

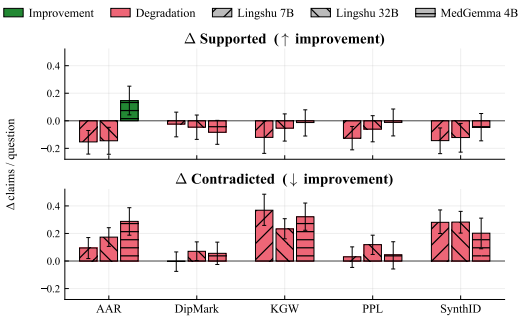


Figure 4: **Impact of watermarking on visual understanding.** We show the difference between watermarked and unwatermarked reasoning in the average number of supported (top) and contradicted (bottom) visual claims on MedXpertQA.

Understanding why watermarks degrade reasoning We observed that the failures above share a common underlying cause: when the watermark perturbs the sampling distribution at a token where the model commits to a clinical interpretation, it can redirect the entire downstream reasoning chain. We isolate this with a controlled prefix-seeding experiment on three MedXpertQA-MM questions (App. E) and two schemes (KGW and PPL): watermarking amplifies distractor prefixes the clean model dismisses (accuracy drops of 14–25 pp) and suppresses rescue cues it would otherwise exploit.

4.3 Judge Validation

A board-certified physician annotated 650 completions blinded to model and watermark configuration, drawn from four (model, scheme) configurations spanning the full error-density range: PHI-4-14B/SynthID and LLAMA-3.1-70B/DipMark on MedQA, LINGSHU-7B/SynthID and GEMMA-4-31B/DipMark on MedXpertQA-MM (250 MedQA + 250 MedXpertQA-MM single-response completions, plus 150 pairwise pairs). As a second automated baseline, CLAUDE-SONNET-4.6 scores the same 650 samples independently.

On the four axes that carry our headline claims, judge–physician agreement is substantial: fabricated-entity counts reach κ=0.75 / ρ=0.75 (n=500), misapplied-term counts κ=0.42 / ρ=0.47, image-contradicted claims ρ=0.74 (n=250, MM only), and pairwise diagnostic divergence κ_w=0.65 / binary κ=0.67 (n=150). Agreement between GEMINI-3-FLASH and CLAUDE-SONNET-4.6 on the medical-content axes (fabrications, misapplied terms, pairwise divergence) matches or exceeds the corresponding judge–physician values (0.80 / 0.81, 0.50 / 0.52, κ_w=0.66 / κ=0.72, respectively), confirming that two independently trained LLM-judges recover the same signal. The full per-axis agreement grid (Table 10) shows that lower κ values concentrate on low-prevalence stylistic flags (format/termination, linguistic corruption, spelling), where judge–physician and judge–judge agreement drop jointly–consistent with base-rate sensitivity of κ rather than systematic bias. We therefore anchor claims on these axes to paired Δs between watermarked and unwatermarked conditions rather than to absolute error levels.

4.4 Reasoning Models: Full vs. Answer-Only Watermarking

The two highest-scoring models on MedQA in our setup are reasoning models (QWEN-3.5-27B (94.6%) and DEEPSEEK-R1-DISTILL-LLAMA-70B (92.3%)), narrowly above the strongest instruction-tuned baseline (LLAMA-3-70B (91.0%)). Unlike instruction-tuned models, they generate a reasoning trace (usually hidden) inside `<thinking>` tags before producing the final answer; a watermark can therefore be applied in two distinct ways: *answer-only* (FA-only), which watermarks only the post-`<thinking>` answer tokens, and *full watermark* (Full WM), which watermarks both the reasoning trace and the final answer. We compare both approaches on four reasoning models–

DEEPSEEK-R1-DISTILL-LLAMA-8B and DEEPSEEK-R1-DISTILL-LLAMA-70B, DEEPSEEK-R1-DISTILL-QWEN-32B, and QWEN-3.5-27B with KGW at 3 strength levels ($\delta \in \{1, 3, 5\}$).

Full WM degrades reasoning quality on every axis except the final answer. Across the six judge dimensions, Full WM drives substantial negative shifts on *reasoning efficiency* and *terminological precision*, with smaller drops on *clinical coherence* and *distractor engagement* for each of the four models (Figure 11). Two dimensions change negligibly: *answer–reasoning alignment* and *diagnostic accuracy*. The model thus still selects the correct letter through reasoning that is internally consistent with its conclusion, but its output becomes substantially harder to use as a clinical reference due to imprecise or fabricated terminology, redundant deliberation, and distractor engagement that the unwatermarked response dismisses. We also rate how harmful the reasoning is (NO_HARM_RISK/LOW_RISK/HARM_RISK_PRESENT). Even though both small and large models similarly degrade the reasoning, only small models' reasoning is judged to be significantly harmful.

Watermarking only the final answer is essentially free. Scoring Full WM against FA-only as the reference reproduces the same per-dimension pattern within bootstrap intervals (Figure 11, orange vs. blue): FA-only is equivalent to Base on every axis. HARM_RISK_PRESENT likewise stays at the No WM baseline under FA-only on every model (Table 9): the watermark, applied only to post-`<thinking>` answer tokens, leaves the model's chain of thought unperturbed.

Trace watermarking inflates verbosity and induces recursive self-correction loops. Inspection of Full WM traces reveals a consistent pathology across the four models: revisiting already established (and corrected) premises. A representative excerpt is shown in Table 3. Quantitatively, Full WM at $\delta = 5$ inflates mean total output length (reasoning trace plus final answer) by $+36\,\%$ to $+69\,\%$ across the four models, while FA-only does not (Table 6). This means that Full WM's increase in detectability (at a given distortion level) is partly a length artefact rather than a stronger per-token signal.

Baseline (No WM)	**Full WM (KGW $\delta = 3$)**
Linear: "The patient has low Vitamin D but normal Ca^{++}, suggesting secondary hyperparathyroidism."	*Circular:* "Wait, let me rethink. The Ca^{++} is normal. Actually, the Ca^{++} is not low. Hmm. Let me check the Ca^{++} again."

Table 3: Excerpt of DEEPSEEK-R1-DISTILL-QWEN-32B reasoning traces, baseline vs. Full WM.

5 Conclusion and Limitations

We introduce a rigorous and systematic evaluation of watermark-induced degradation in medicine, with three specific LLM-as-judges validated by clinicians. We find that, with most watermarks and models tested, the accuracy on medical benchmarks barely drops even at high watermark strength. Nonetheless, our analysis of reasoning traces tells a different story. Watermarks may severely degrade the reasoning of the models (e.g., hallucinate facts, misinterpret data) and may even harm their visual understanding. For reasoning models in particular, we suggest watermarking only the final answer, which we show prevents such degradation.

Limitations Both benchmarks are multiple-choice and single-turn, so they do not capture open-ended or interactive clinical workflows such as report writing, patient communication, multi-turn dialogue, or retrieval-augmented and tool-augmented reasoning. In such settings, each watermarked response becomes context for subsequent generation, and it remains unclear whether watermarking-induced degradation compounds or attenuates across turns. Future work should extend the evaluation to these settings. Additionally, our evaluation is restricted to open-weight models (watermarking closed-source models is, in most cases, not possible). Because proprietary models are often substantially larger, they may behave differently under watermarking. Finally, our judge pipeline, while validated against two board-certified physicians on a subset of the data, must reliably detect even subtle failure modes (e.g., hallucinated image findings), which requires a judge (LLM or VLM) substantially more capable than the models being evaluated.

Future work: improving watermark design. The watermarking methods we test are model- and domain-agnostic and therefore do not account for the clinical importance of specific tokens. It remains unclear whether domain-adapted schemes would outperform generic ones or how such

schemes should be designed in practice. Given the societal impacts, we believe future work should explore watermarking methods that better preserve clinically critical content.

Acknowledgments and Disclosure of Funding

Robin Staab has been supported by the SERI grant SAFEAI (Certified Safe, Fair and Robust Artificial Intelligence, contract no. MB22.00088). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them. The work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI) (SERI-funded ERC Consolidator Grant). We thank Dr. Andrea Radebold (Internal Medicine, Quinnipiac Medical of Branford/Yale School of Medicine) for her contribution to data annotation.

References

[1] American Medical Association. 2026 physician survey on augmented intelligence. https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf, 2026. Accessed: 2026-04-30.

[2] Anirudh S. Gowda, Meer C. Chhabria, and Ryan C. Lam. Use of large language models on radiology reports: A scoping review. *Journal of the American College of Radiology*, 23(3): 437–454, 2026. doi: 10.1016/j.jacr.2025.10.005. Published online 2025.

[3] Gary E. Weissman, Toni Mankowitz, and Genevieve P. Kanter. Unregulated large language models produce medical device-like output. *npj Digital Medicine*, 8(1):148, 2025. doi: 10.1038/s41746-025-01544-y.

[4] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). Official Journal of the European Union, L 2024/1689, 2024.

[5] European Commission. Second draft code of practice on marking and labelling of AI-generated content. https://digital-strategy.ec.europa.eu/en/library/commission-publishes-second-draft-code-practice-marking-and-labelling-ai-generated-content, 2026. Published: 2026-03-05. Accessed: 2026-04-30.

[6] Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. 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)*, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.83/.

[7] Johannes Moll, Markus Graf, Tristan Lemke, Nicolas Lenhart, Daniel Truhn, Jean-Benoit Delbrouck, Jiazhen Pan, Daniel Rueckert, Lisa C. Adams, and Keno K. Bressem. Evaluating reasoning faithfulness in medical vision-language models using multimodal perturbations. In *Proceedings of Machine Learning for Health (ML4H)*, volume 297 of *Proceedings of Machine Learning Research*, 2025. arXiv:2510.11196v2.

[8] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.

[9] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[10] Siun Kim and Hyung-Jin Yoon. Questioning our questions: How well do medical QA benchmarks evaluate clinical capabilities of language models? In *Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP)*, pages 274–296, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.bionlp-1.24.

[11] Y. Hu et al. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[12] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, et al. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. *arXiv preprint arXiv:2408.03361*, 2024.

[13] Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 80961–80990. PMLR, 2025. arXiv:2501.18362.

[14] Umakanta Maharana, Sarthak Verma, Avarna Agarwal, Prakashini Mruthyunjaya, Dwarikanath Mahapatra, Sakir Ahmed, and Murari Mandal. Right prediction, wrong reasoning: Uncovering LLM misalignment in RA disease diagnosis. *arXiv preprint arXiv:2504.06581*, 2025.

[15] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.21.

[16] Veysel Kocaman, Mustafa Aytuğ Kaya, Andrei Marian Feier, and David Talby. Clinical large language model evaluation by expert review (CLEVER): Framework development and validation. *JMIR AI*, 4(1):e72153, 2025. doi: 10.2196/72153.

[17] Rahul K. Arora, Jason Wei, Rebecca S. Hicks, et al. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

[18] Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57:47:1–47:36, 2024. doi: 10.1145/3691626. URL https://arxiv.org/abs/2312.07913.

[19] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

[20] Thibaud Gloaguen, Robin Staab, Nikola Jovanović, and Martin Vechev. A unified framework for llm watermarks, 2026. URL https://arxiv.org/abs/2602.06754.

[21] Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. Dipmark: A stealthy, efficient and resilient watermark for large language models. 2023.

[22] Scott Aaronson. Watermarking of large language models. In *Workshop on Large Language Models and Transformers, Simons Institute, UC Berkeley*, 2023.

[23] Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

[24] Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. 2024.

[25] Peiru Yang, Xintian Li, Wanchun Ni, Jinhua Yin, Huili Wang, Guoshun Nan, Shangguang Wang, Yongfeng Huang, and Tao Qi. Enhancing watermarking quality for LLMs via contextual generation states awareness. *arXiv preprint arXiv:2506.07403*, 2025.

[26] Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. Who wrote this code? watermarking for code generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, 2024.

[27] Liang Chen, Yatao Bian, Yang Deng, Deng Cai, Shuaiyi Li, Peilin Zhao, and Kam-Fai Wong. WatME: Towards lossless watermarking through lexical redundancy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9166–9180, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.496. URL https://aclanthology.org/2024.acl-long.496/.

[28] Shuliang Liu, Xingyu Li, Hongyi Liu, Yibo Yan, Bingchen Duan, Qi Zheng, Dong Fang, Lingfeng Su, and Xuming Hu. Distilling the thought, watermarking the answer: A principle semantic guided watermark for reasoning large language models. *OpenReview*, 2026. URL https://openreview.net/forum?id=T6NVogsXCZ. ICLR 2026 Poster.

[29] Rochana Prih Hastuti, Rian Adam Rajagede, Mansour Al Ghanim, Mengxin Zheng, and Qian Lou. Factuality beyond coherence: Evaluating LLM watermarking methods for medical texts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15129–15147, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.818. arXiv:2509.07755.

[30] Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[31] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[32] Marah Abdin et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[33] Ankit Pal and Malaikannan Sankarasubbu. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B, 2024.

[34] Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Jin-Fang Hu, Zhiyuan Liu, and Bowen Zhou. UltraMedical: Building specialized generalists in biomedicine. In *Advances in Neural Information Processing Systems 37: Datasets and Benchmarks Track (NeurIPS)*, 2024. Spotlight; arXiv:2406.03949.

[35] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

[36] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Blog post.

[37] Qwen Team. Qwen3 technical report. *arXiv preprint*, 2025.

[38] Google Open Source. Gemma 4: Expanding the gemmaverse with Apache 2.0, April 2026. URL https://opensource.googleblog.com/2026/03/gemma-4-expanding-the-gemmaverse-with-apache-20.html. Google Open Source Blog.

[39] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

[40] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

A Additional Experimental Details

A.1 Watermarking Scheme Details

All five schemes share a common structure: at each token position, a hash of the preceding context seeds a pseudorandom number generator that assigns a score g_u to every vocabulary entry u. The *sampling mechanism* then maps the original next-token distribution p and the scores g to a watermarked next-token distribution $q(g)$, from which the next token is drawn. The schemes differ in the score distribution, in this mapping, and consequently in their detectability–quality trade-offs. We refer to Gloaguen et al. [20] for a unified optimization-theoretic treatment of several of the schemes below. In all cases the score assignment and the reweighting are restricted to the top-k of p as configured in the vLLM sampling parameters.

KGW [19]. The context-dependent hash partitions the top-k vocabulary into a *green list* and a *red list* via i.i.d. Bernoulli($\frac{1}{2}$) draws ($g_u \in \{0, 1\}$), so that in expectation half of the tokens are green. A constant logit bias δ is added to the green-list tokens before sampling, yielding the soft logit-bias watermark

$$q(g) \;\propto\; p\,\exp(\delta\, g). \tag{1}$$

This is the soft Kirchenbauer scheme; we do not use the hard variant that excludes red-list tokens entirely. Detection tests whether the proportion of green-list tokens among the generated text exceeds the $\frac{1}{2}$ chance level via a one-sided binomial test (and a z-test for cross-checking). Watermark strength is controlled by δ.

DipMark [21]. A distribution-preserving scheme parameterized by $\alpha \in (0, \frac{1}{2}]$. The context hash seeds a uniform-random permutation π of the top-k vocabulary; tokens are then reweighted according to the symmetric α-shift of the permuted CDF. Concretely, in the permuted ordering the scheme averages two indicator-style reweights at the boundaries α and $1-\alpha$ of the cumulative mass, which at $\alpha = \frac{1}{2}$ yields exact distortion-freeness in expectation, $\mathbb{E}_G[q(G)] = p$, and as $\alpha \to 0$ progressively biases mass toward the upper half of the permutation. Detection compares each generated token's rank under the recomputed permutation against the threshold γ: tokens above the threshold count as "green", and we apply a one-sided binomial test (and a z-test) against $1-\gamma$. We evaluate only the distortion-free configuration $\alpha = \gamma = 0.5$, as is standard in the literature.

AAR [22]. For each token position, i.i.d. Gumbel-distributed scores g_u (location 0, scale $\beta = 1$) are drawn and the next token is selected deterministically as

$$\hat{u} \;=\; \arg\max_u \Big(g_u + \tau\, \log p_u\Big), \qquad \tau \;=\; \frac{\beta}{1+\beta\,\delta}, \tag{2}$$

and represented as a Dirac distribution at $\hat{u}$. We fix $\beta = 1$ throughout and sweep $\delta \geq 0$ as the watermark-strength knob. At $\delta = 0$ we recover the canonical AAR sampler $\arg\max_u(g_u + \log p_u)$: by the Gumbel-max trick the marginal distribution of $\hat{u}$ over the score draw is exactly p, so the scheme is distortion-free, and AAR is the optimal sampling mechanism among all distortion-free watermarks that use a sum-based detector [20], corresponding to the soft KL constraint $\mathrm{KL}(\mathbb{E}_G[q(G)] \,\|\, p) = 0$. For $\delta > 0$, $\tau < 1$ down-weights $\log p$ relative to the Gumbel score, biasing the argmax toward the score sequence and yielding a more easily detectable but no-longer-distortion-free signal.

PPL [20]. For each token position, the scheme draws i.i.d. Bernoulli-sum scores $g_u \sim \mathrm{Binomial}(N, \frac{1}{2})$ and emits a Dirac distribution at the *tilted argmax*

$$\hat{u}(\beta) \;=\; \arg\max_u \big(g_u + \beta\, \log p_u\big), \tag{3}$$

where the tilt $\beta \geq 0$ is solved per token by 1D bisection so that the expected log-likelihood of $\hat{u}$ exactly saturates the per-token log-perplexity budget,

$$\mathbb{E}_G\big[\log p_{\hat{u}(\beta)}\big] \;=\; \sum_v p_v \log p_v \;-\; \varepsilon. \tag{4}$$

The constraint is the same per-token log-perplexity budget $(p - q(g)) \cdot \log p \leq \varepsilon$ as the *hard* variant from the unified framework, but enforced *in expectation over the scores* via the scalar tilt β rather

than by a per-realization linear program; in practice this matches the constraint while admitting the simple Gumbel-style implementation. The expectation in the bisection is approximated by 128 Monte Carlo score draws, with a fixed schedule of 60 bisection iterations and the search bracket $\beta \in [0, 100]$. Watermark strength is controlled by ε (larger ε permits more deviation from the model's distribution and thus a stronger watermark).

SynthID [23]. For each token, m independent layers of i.i.d. Bernoulli scores $g^{(i)} \in \{0,1\}^{|V|}$ are drawn (in our configuration $m = 30$). Starting from $q^{(0)} := p$, the algorithm applies m tournament reweighting layers parameterized by the number of leaves per match ℓ, controlling the strength of the watermark. With $\ell = 2$ (*non-distortionary* tournament), each layer applies the multiplicative update

$$q_u^{(i)} \;=\; q_u^{(i-1)}\big(1 + g_u^{(i)} - q^{(i-1)} \cdot g^{(i)}\big), \tag{5}$$

which preserves the marginal distribution $\mathbb{E}_{g^{(i)}}[q^{(i)}] = q^{(i-1)}$. With $\ell > 2$ leaves per match (*distortionary* tournament), the layer instead applies the closed-form tournament-winner reweight

$$q_u^{(i)} \;=\; q_u^{(i-1)} \cdot \begin{cases} \dfrac{1-(1-m_i)^{\ell}}{m_i} & \text{if } g_u^{(i)} = 1, \\ (1-m_i)^{\ell-1} & \text{if } g_u^{(i)} = 0, \end{cases} \qquad m_i := q^{(i-1)} \cdot g^{(i)}, \tag{6}$$

which biases each layer toward tokens with $g_u^{(i)} = 1$ and is no longer distortion-free for $\ell > 2$. Stacking layers progressively strengthens the watermark signal. The original SynthID-Text scheme of Dathathri et al. [23] is typically configured with $\ell = 2$ leaves (the non-distortionary tournament) and a depth in the range $m \in \{20, 30\}$, with detectability controlled by varying m.

A.2 Generation and Reproducibility

All completions are generated with vLLM (v0.19.1), with the watermark realised as a custom `LogitsProcessor` so that watermarked and unwatermarked runs differ only at the logit-modification step. All experiments use the model-native tokenizer without post-processing. Decoding uses temperature 0.7 and unrestricted top-k throughout. Each (model, watermark, strength) configuration is run with a fixed sampling `seed` (default 0); paired watermarked/unwatermarked completions for a given question therefore share both prompt and seed, isolating the watermark as the only source of divergence. The watermark uses output-only context hashing, so the first 4 generated tokens (one context window) are not watermarked; this affects $< 1\%$ of tokens and is consistent between generation and detection.

For non-reasoning models we cap output at $2{,}500$ tokens, while for reasoning models we allocate a generation budget of $10{,}000$ tokens, and post-hoc split out the answer portion (after `</think>`) with a maximum of $2{,}500$ tokens for comparability with the non-reasoning pipeline; the watermark is applied to both reasoning trace and answer tokens by default. In practice, the $2{,}500$-token cap never truncated a well-formed response; the only completions reaching the limit were degenerate outputs (e.g. repetitive token loops), which are captured by the format/termination failure axis of the single-response audit. Answer letters are extracted by a two-tier regex parser. A primary pass scans for structured formats observed across models – `ANSWER: X`, `FINAL ANSWER: X`, `ANSWER: (X)`, `ANSWER: <X>`, `\boxed{X}`, `FINAL ANSWER is X`, and `FINAL ANSWER: <free text> (X)` where exactly one letter appears parenthesised. If, and only if, the primary pass finds nothing, a fallback pass admits conservative natural-language patterns (e.g. "the answer is X", "I would choose X", or a candidate name followed by a parenthesised letter at end of line). The parser only commits to a letter when exactly one option (A–E) is unambiguously indicated; otherwise the response is recorded as unanswered.

All experiments run on a single node with up to 8 NVIDIA RTX Blackwell 6000 Pro GPUs (96 GB VRAM each). Watermark implementations, generation and analysis pipelines are released alongside the paper to enable reproduction of the reported numbers.

A.3 Existing Asset Licenses

We use existing datasets and model checkpoints only for evaluation, and cite the original creators in Sec. 3.1. The licenses below are the license names stated by the corresponding official repositories or model cards.

Datasets.

- *MedQA*: MIT License.
- *MedXpertQA-MM*: MIT License.

Models.

- LLAMA-3.1-8B and LLAMA-3.1-70B: Meta Llama 3 Community License Agreement.
- BIOMISTRAL-7B: Apache License 2.0.
- GEMMA-3-12B: Gemma Terms of Use.
- PHI-4-14B: MIT License.
- OPENBIOLLM-70B and ULTRAMEDICAL-70B: Meta Llama 3 Community License Agreement.
- DEEPSEEK-R1-DISTILL-LLAMA-8B, DEEPSEEK-R1-DISTILL-LLAMA-70B, and DEEPSEEK-R1-DISTILL-QWEN-32B: MIT License; the Llama-derived variants also inherit the applicable Meta Llama Community License terms, and the Qwen-derived variant inherits the applicable Apache License 2.0 upstream terms.
- QWEN-3.5-27B, QWEN-3-VL-8B, and QWEN-3-VL-32B: Apache License 2.0.
- GEMMA-4-31B: Apache License 2.0.
- LINGSHU-7B and LINGSHU-32B: MIT License.
- MEDGEMMA-4B and MEDGEMMA-27B: Health AI Developer Foundations Terms of Use.

B Per-Configuration Accuracy and Perplexity

Tables 4 and 5 report detectability (TPR at 1% FPR) and benchmark accuracy for every (model, scheme, strength) triple evaluated in Sec. 4.1, one benchmark per table. Rows are grouped by scheme and sorted by increasing watermark strength; each model contributes a (TPR, Acc) sub-column pair, with the unwatermarked baseline shown beneath the model header. Configurations strictly stronger than the first to reach TPR $\geq 99\%$ are omitted (—), as they yield no additional detectability.

Table 4: MedQA: watermark detectability (TPR @ 1%FPR, in %) and benchmark accuracy (Acc, in %) for every (model, scheme, strength) configuration evaluated in Sec. 4.1, computed over $N = 2000$ questions (full benchmark). Per-model unwatermarked baseline accuracy is shown beneath each model header. For each (model, scheme) we report all configurations up to and including the first one with TPR $\geq 99\%$; stronger configurations beyond that saturation point are omitted (—), as they yield no additional detectability.

		GEMMA-3-12B (68.2)		LLAMA-3.1-70B (91.0)		LLAMA-3.1-8B (72.8)		OPENBIOLLM-70B (73.8)		PHI-4-14B (81.5)		ULTRAMEDICAL-70B (86.6)	
Scheme	Strength	TPR	Acc	TPR	Acc	TPR	Acc	TPR	Acc	TPR	Acc	TPR	Acc
AAR	$\delta = 0$	96.3	69.2	37.3	89.5	91.4	68.5	36.6	76.0	70.0	82.5	67.8	87.0
AAR	$\delta = 0.1$	98.2	68.5	47.1	90.1	95.7	68.9	45.3	74.5	—	—	76.5	86.9
AAR	$\delta = 0.2$	99.3	69.2	55.6	89.6	98.1	65.8	51.7	73.7	86.0	81.3	84.4	86.8
AAR	$\delta = 0.3$	—	—	64.7	90.8	99.4	62.0	59.1	74.9	92.0	81.6	90.5	86.2
AAR	$\delta = 0.4$	—	—	73.1	89.5	—	—	66.7	72.7	96.2	80.8	93.5	86.5
AAR	$\delta = 0.5$	—	—	80.7	89.8	—	—	76.1	72.0	98.1	79.7	95.6	85.4
DipMark	$\alpha = 0.5$	90.1	68.5	30.2	89.8	83.8	67.2	25.4	72.0	59.8	81.1	57.7	87.5
KGW	$\delta = 0.5$	16.9	69.2	6.4	90.1	18.6	70.5	3.1	74.6	5.9	81.8	11.7	87.6
KGW	$\delta = 1$	65.1	69.7	22.2	90.0	63.0	69.6	11.4	74.9	27.5	82.2	41.2	86.7
KGW	$\delta = 2$	99.4	68.8	70.7	89.6	99.2	64.5	51.4	73.4	87.3	80.7	91.4	85.9
KGW	$\delta = 3$	—	—	95.2	90.0	—	—	85.5	69.0	99.2	80.3	99.1	85.2
KGW	$\delta = 4$	—	—	99.1	90.4	—	—	95.9	63.2	—	—	—	—
KGW	$\delta = 5$	—	—	—	—	—	—	98.8	54.9	—	—	—	—
PPL	$\varepsilon = 0$	99.5	68.4	81.7	89.6	99.1	68.3	57.2	73.4	90.8	78.8	94.4	80.2
PPL	$\varepsilon = 0.1$	—	—	92.8	88.2	—	—	73.9	73.5	96.7	78.8	97.7	78.5
PPL	$\varepsilon = 0.2$	—	—	97.8	89.3	—	—	86.6	71.9	—	—	99.3	70.0
PPL	$\varepsilon = 0.3$	—	—	99.3	87.9	—	—	92.7	69.8	99.8	77.5	—	—
PPL	$\varepsilon = 0.4$	—	—	—	—	—	—	95.4	69.2	—	—	—	—
PPL	$\varepsilon = 0.5$	—	—	—	—	—	—	97.5	67.5	—	—	—	—
SynthID	$\ell = 2$	91.1	69.5	38.3	89.8	85.9	71.7	22.9	75.9	48.4	81.2	67.4	87.0
SynthID	$\ell = 3$	98.9	69.5	63.9	89.6	98.3	65.0	48.1	73.9	84.5	80.8	89.2	84.0
SynthID	$\ell = 4$	100.0	68.8	79.4	89.6	100.0	48.4	70.5	71.2	97.2	80.0	96.3	81.2
SynthID	$\ell = 5$	—	—	87.9	90.4	—	—	83.5	67.2	99.1	79.3	98.8	76.8
SynthID	$\ell = 6$	—	—	93.4	90.3	—	—	93.3	59.2	—	—	99.2	63.1

Table 5: MedXpertQA-MM: watermark detectability (TPR @ 1%FPR, in %) and benchmark accuracy (Acc, in %) for every (model, scheme, strength) configuration evaluated in Sec. 4.1, computed over $N = 2000$ questions (full benchmark). Per-model unwatermarked baseline accuracy is shown beneath each model header. For each (model, scheme) we report all configurations up to and including the first one with TPR $\geq 99\%$; stronger configurations beyond that saturation point are omitted (—), as they yield no additional detectability.

		Gemma-4-31B (54.9)		Lingshu-32B (29.9)		Lingshu-7B (24.3)		MedGemma-27B (32.8)		MedGemma-4B (24.4)		Qwen-3-VL-32B (38.9)		Qwen-3-VL-8B (29.8)	
Scheme	Strength	TPR	Acc	TPR	Acc	TPR	Acc	TPR	Acc	TPR	Acc	TPR	Acc	TPR	Acc
AAR	$\delta = 0$	65.2	55.4	72.6	31.4	73.7	23.8	90.0	33.6	84.2	24.2	93.5	39.1	97.9	26.5
AAR	$\delta = 0.1$	71.2	55.0	82.5	29.6	82.9	23.5	94.2	34.2	91.2	24.7	95.5	39.3	98.6	27.0
AAR	$\delta = 0.2$	79.5	54.6	89.5	28.8	91.4	23.4	96.6	33.0	94.0	24.8	97.9	40.4	99.1	27.3
AAR	$\delta = 0.3$	84.8	54.7	93.8	30.1	94.5	24.2	98.0	33.5	96.8	24.0	98.9	39.0	—	—
AAR	$\delta = 0.4$	90.5	54.1	96.2	30.4	97.5	23.8	98.9	32.1	98.2	23.9	99.5	38.7	—	—
AAR	$\delta = 0.5$	92.8	53.6	98.3	30.2	98.5	23.5	99.6	31.9	99.0	22.2	—	—	—	—
DipMark	$\alpha = 0.5$	45.6	54.4	57.0	31.2	57.5	25.7	82.0	32.4	72.5	23.6	87.1	39.0	94.5	29.0
KGW	$\delta = 0.5$	7.1	54.3	9.2	30.6	10.1	24.9	14.8	34.1	14.1	22.1	17.1	38.8	30.3	27.8
KGW	$\delta = 1$	26.9	54.0	34.1	31.8	38.9	23.4	59.2	32.5	54.0	23.2	62.3	40.1	81.2	29.1
KGW	$\delta = 2$	81.0	54.4	91.0	29.5	91.0	22.4	97.0	33.6	91.8	23.8	98.0	39.7	99.2	29.2
KGW	$\delta = 3$	96.9	54.4	99.2	30.4	99.0	25.2	100.0	32.5	96.5	21.5	99.9	38.5	—	—
KGW	$\delta = 4$	99.9	54.0	—	—	—	—	—	—	—	—	—	—	—	—
KGW	$\delta = 5$	—	—	—	—	—	—	—	—	97.2	21.8	—	—	—	—
PPL	$\varepsilon = 0$	88.9	53.3	92.8	30.6	94.7	23.9	98.6	33.1	91.1	22.4	99.1	37.4	99.6	26.9
PPL	$\varepsilon = 0.1$	96.1	52.6	97.8	29.6	97.4	23.2	99.7	32.5	—	—	—	—	—	—
PPL	$\varepsilon = 0.2$	98.8	53.8	99.4	29.2	99.2	23.9	—	—	94.0	21.9	—	—	—	—
PPL	$\varepsilon = 0.3$	99.6	53.4	—	—	—	—	—	—	—	—	—	—	—	—
PPL	$\varepsilon = 0.4$	—	—	—	—	—	—	—	—	96.4	21.1	—	—	—	—
PPL	$\varepsilon = 0.5$	—	—	—	—	—	—	—	—	96.7	21.1	—	—	—	—
SynthID	$\ell = 2$	61.2	53.8	63.3	30.4	69.0	23.8	92.9	31.9	88.6	24.5	89.0	40.0	94.8	26.7
SynthID	$\ell = 3$	81.2	53.8	89.6	30.3	92.5	24.1	98.7	32.9	95.0	22.6	98.1	38.6	99.4	26.4
SynthID	$\ell = 4$	90.7	53.8	98.0	27.9	98.0	24.3	99.9	32.6	96.5	21.6	99.5	39.0	—	—
SynthID	$\ell = 5$	95.4	52.9	99.4	26.6	99.8	23.0	—	—	96.6	20.1	—	—	—	—

Model	Condition	Acc. (%)	TPR (%)	Mean output length (chars)
DEEPSEEK-R1-DISTILL-LLAMA-8B	No WM	59.2	–	9,023
	FA-only, δ=1	61.0	14.3	8,890
	FA-only, δ=3	59.2	87.1	8,856
	FA-only, δ=5	58.9	96.8	8,870
	Full WM, δ=1	61.2	15.4	9,165
	Full WM, δ=3	59.1	93.0	11,621
	Full WM, δ=5	57.2	98.9	15,257
DEEPSEEK-R1-DISTILL-QWEN-32B	No WM	83.6	–	6,447
	FA-only, δ=1	83.6	8.1	6,362
	FA-only, δ=3	82.9	74.3	6,444
	FA-only, δ=5	81.5	97.7	6,482
	Full WM, δ=1	84.1	10.3	6,737
	Full WM, δ=3	83.7	84.4	8,166
	Full WM, δ=5	82.1	99.6	10,027
DEEPSEEK-R1-DISTILL-LLAMA-70B	No WM	92.3	–	6,025
	FA-only, δ=1	92.4	7.0	5,995
	FA-only, δ=3	91.9	59.2	6,072
	FA-only, δ=5	90.8	92.4	6,088
	Full WM, δ=1	91.9	7.1	6,239
	Full WM, δ=3	91.4	82.3	7,227
	Full WM, δ=5	89.8	99.7	8,759
QWEN-3.5-27B	No WM	94.6	–	13,258
	FA-only, δ=1	94.1	23.5	13,125
	FA-only, δ=3	94.5	97.5	13,285
	FA-only, δ=5	94.1	98.8	13,743
	Full WM, δ=1	94.8	26.1	13,285
	Full WM, δ=3	94.0	99.0	14,976
	Full WM, δ=5	77.1	99.5	18,019

Table 6: MedQA accuracy, TPR at 1% FPR, and mean output length (reasoning trace plus final answer, characters) for four reasoning models under No WM, FA-only (post-<thinking> tokens only), and Full WM (entire output), KGW $\delta \in \{1, 3, 5\}$. TPR is computed on the final-answer tokens only.

B.1 Perplexity Under Watermarking

Figures 5 and 6 plot generation perplexity against watermark detectability. Perplexity increases monotonically with watermark strength on every model and scheme, but its magnitude is a poor predictor of downstream harm: GEMMA-3-12B shows clear perplexity shifts under KGW with no accuracy effect, while ULTRAMEDICAL-70B suffers a 31 pp accuracy collapse under PPL despite only moderate perplexity shifts. Exact per-configuration values accompany the code release.

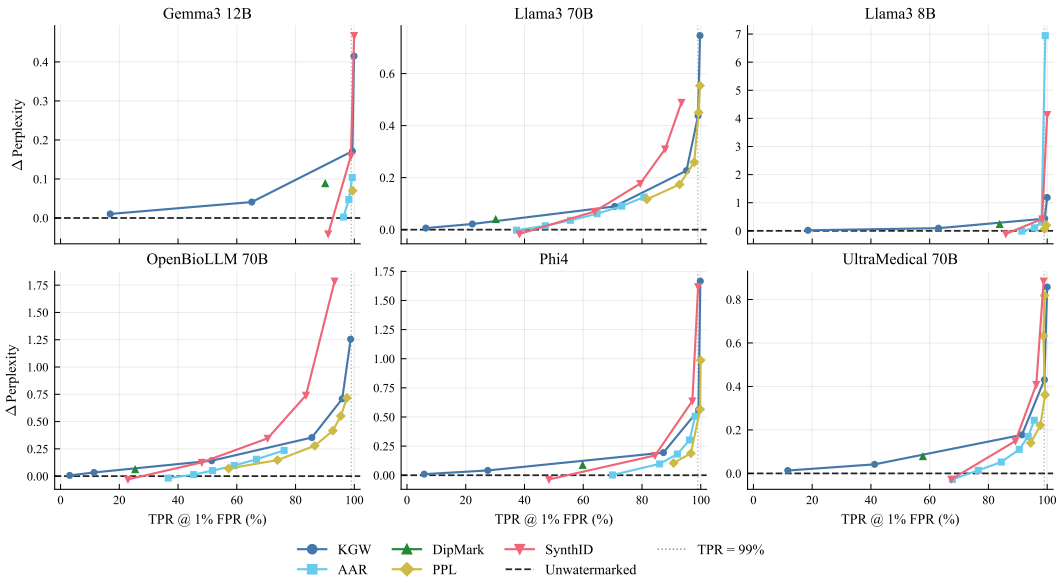


Figure 5: **MedQA: generation-perplexity shift vs. watermark detectability**, six non-reasoning LLMs. $\Delta\,\mathrm{PPL} = \mathrm{PPL}_{\mathrm{wm}} - \mathrm{PPL}_{\mathrm{base}}$. Monotonic in strength on every model×scheme cell; absolute magnitude spans an order of magnitude (note differing y-axes). LLAMA-3.1-8B has the steepest increases; GEMMA-3-12B and ULTRAMEDICAL-70B show contained shifts despite their sharply different accuracy robustness.

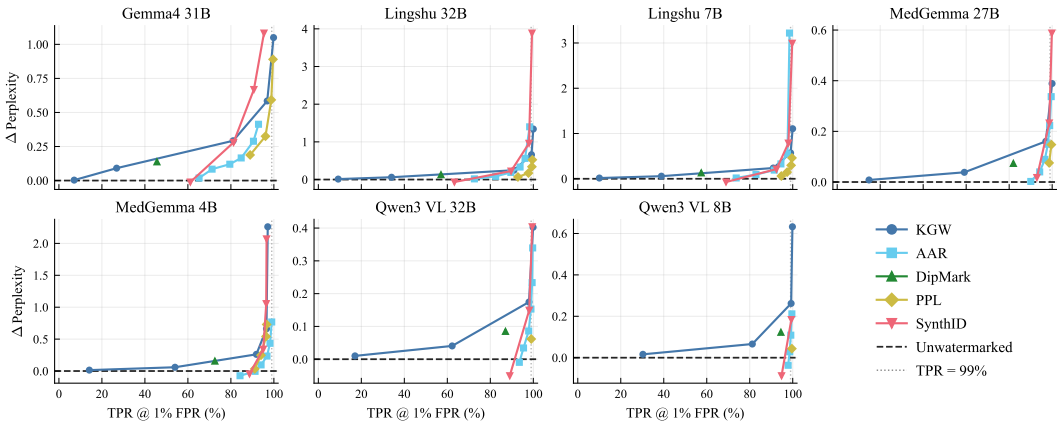


Figure 6: **MedXpertQA-MM: generation-perplexity shift vs. watermark detectability**, seven VLMs. Same overall shape as MedQA; KGW and SynthID drive the steepest high-TPR increases. LINGSHU-32B and LINGSHU-7B show pronounced jumps at the highest strengths despite minimal accuracy changes (cf. Figure 3).

C Audit Outcomes Across Watermark Detectability

Sec. 4.2 fixes a single deployment operating point per cell ($\geq 99\%$ TPR), leaving the *shape* of the strength–harm relationship unresolved – whether harm grows smoothly with detectability, or sits near baseline up to a threshold and then jumps. We sweep watermark strength on two contrasting cells: PHI-4-14B on MedQA (an accuracy-stable text-only LLM) and MEDGEMMA-4B on MedXpertQA-MM (a smaller medical VLM), under the two schemes whose $\geq 99\%$-TPR signatures differ most in Table 2 – *KGW* (broad multi-axis effect) and *SynthID* (the largest silent-error cell on both models).

Figure 7 plots one harm metric per axis-of-interest against detectability. For PHI-4-14B we report the per-response confident-error rate (a major component of F&C) and the pairwise vignette-fabrication

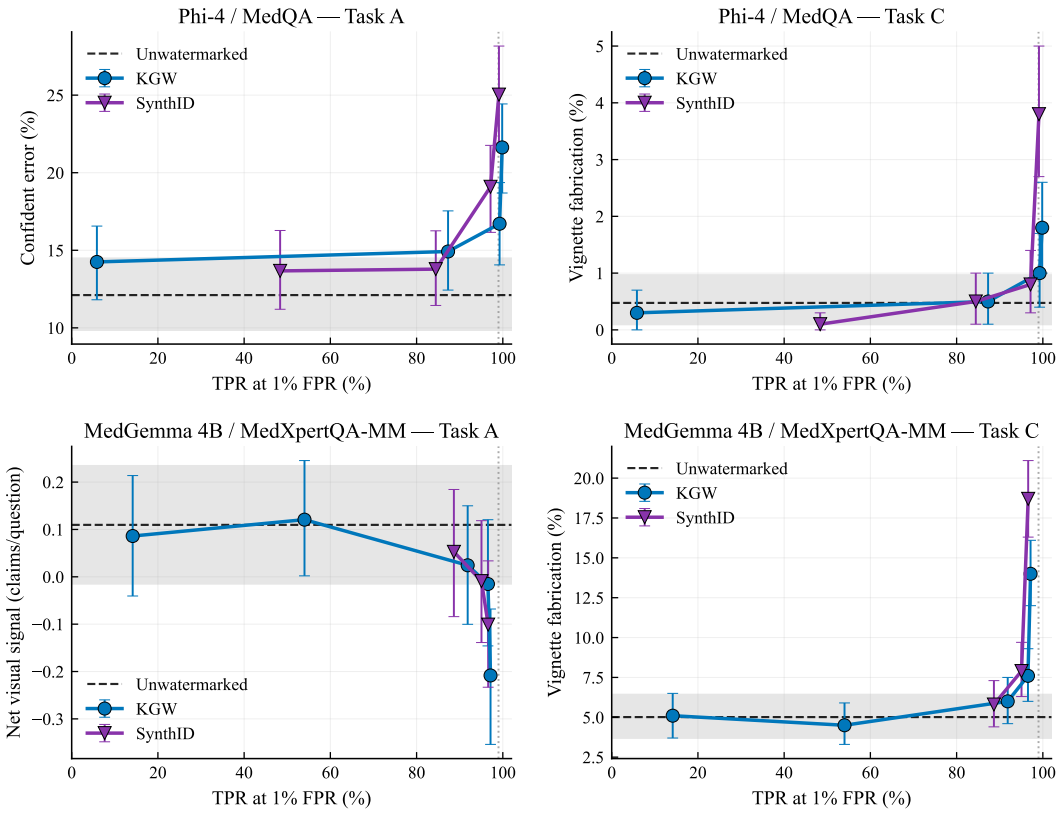


Figure 7: **Audit outcomes vs. watermark detectability** on PHI-4-14B (MedQA, top) and MEDGEMMA-4B (MedXpertQA-MM, bottom) under KGW (blue) and SynthID (purple). Each point is one (scheme, strength); dashed line and grey band give the unwatermarked baseline mean and 95%-bootstrap CI; dotted vertical marks TPR $= 99\%$. Top: per-response confident-error rate restricted to right-letter responses (left); pairwise vignette-fabrication rate (right). Bottom: net visual signal, SUPPORTED $-$ CONTRADICTED visual claims per question (left); pairwise vignette-fabrication rate (right).

rate (the silent-error signature picked up by VigFab). For MEDGEMMA-4B we report the net visual signal used in Figure 4 and again vignette-fabrication, so the silent-error axis is comparable across the two cells.

KGW vs. SynthID: monotonic slider vs. threshold cliff. KGW degrades smoothly with detectability. On PHI-4-14B, watermarked confident-error climbs from 14.2% at TPR 5.9% to 21.6% at TPR 99.9%, against a stable $\approx 12\%$ unwatermarked baseline; on MEDGEMMA-4B, the net visual signal slides from $+0.09$ at TPR 14% through zero to -0.21 at TPR 97% (baseline $\approx +0.10$). SynthID is threshold-y: it tracks baseline up to TPR $\approx 90\%$ and then jumps in the final detectability decade. On PHI-4-14B, confident-error is flat at $\approx 13.7\%$ across TPRs 48–84%, then rises to 19.1% (TPR 97%) and 25.0% (TPR 99%); pairwise vignette-fabrication stays $\leq 0.8\%$ up to TPR 97% and jumps to 3.8% at TPR 99%. On MEDGEMMA-4B, vignette-fabrication tracks $5.8\% \to 7.9\% \to 18.7\%$ at TPRs $89, 95, 97\%$.

These shapes have different operational consequences: a KGW deployment can move a slider to trade detectability against quality, whereas a SynthID deployment at the detectability a regulator would require sits past the elbow of a cliff.

C.1 Full Reasoning-Audit Grids

Sec. 4.2 reports only the two most-affected models per benchmark. Below: the full 7-model grids per benchmark, per-scheme FabRatio breakdowns, and the all-VLM companion to Figure 4.

Table 7: **Full reasoning audit at** 99% **TPR on MedQA**, for all 7 LLMs. We report the accuracy (ΔAcc), the confident-error rate (ΔConfErr), the faulty-but-correct rate (ΔF&C), the percentage of corrupted spellings (ΔSpell/100), the percentage of fabricated facts (ΔFab/100), the within-response contradiction rate (ΔContr), vignette fabrication (ΔVigFab), clue misreads (ΔMisread), formatting failures (ΔFmt), linguistic corruption (ΔLing), the percentage of misapplied medical terms (ΔMisapp/100), the percentage of paired responses with major diagnostic divergence (Major%), and the vignette fabrication ratio (FabRatio). For unwatermarked rows we report the actual value, whereas for watermarked rows we report the difference/ratio with the unwatermarked baseline. We **bold** entries that are statistically significant (p-value below 0.01).

Model	Scheme	ΔAcc	ΔConfErr	ΔF&C	ΔSpell/100	ΔFab/100	ΔContr	ΔVigFab	ΔMisread	ΔFmt	ΔLing	ΔMisapp/100	Major%	FabRatio
LLAMA-3.1-8B	Unwatermarked	73.3	35.3	29.6	0.8	1.4	22.6	5.8	6.6	0.9	0.1	14.7	–	1.0×
	AAR	-0.6	**+8.1**	**+7.4**	**+5.8**	**+4.1**	**+6.1**	**+5.5**	**+3.2**	**+1.5**	**+6.1**	+0.2	31.9	**2.1**×
	DipMark	-1.6	-0.7	+1.0	+0.9	+1.0	+3.1	+1.3	**+3.4**	-0.3	+0.5	-1.5	29.2	1.6×
	KGW	-2.3	+3.0	+2.4	**+1.8**	**+2.2**	**+5.1**	+1.7	+1.6	+0.5	+0.3	-0.3	29.0	1.0×
	PPL	-0.7	+5.4	+4.3	**+24.6**	+1.0	+4.2	+0.5	**+3.8**	+0.3	**+3.4**	+0.7	29.8	1.6×
	SynthID	-2.0	+3.0	+2.2	**+2.1**	+1.5	+4.0	**+3.7**	+2.6	+1.5	+0.7	+0.2	33.5	1.2×
LLAMA-3.1-70B	Unwatermarked	91.8	6.9	8.1	0.4	0.1	5.8	0.4	1.0	0.0	0.0	1.5	–	1.0×
	AAR	-1.2	-0.5	+0.1	+0.3	0.0	+0.1	+0.1	+0.5	0.0	+0.1	-0.3	4.1	–
	DipMark	-1.2	+1.4	+1.5	-0.1	-0.1	+1.4	-0.2	+0.2	+0.2	0.0	0.0	4.0	–
	KGW	-1.3	+1.8	+1.9	+1.0	0.0	+1.6	+0.4	+0.3	+0.1	+0.1	+0.6	5.1	1.0×
	PPL	-1.8	**+3.3**	**+7.5**	**+89.4**	**+1.9**	+2.1	+0.3	**+3.5**	+0.6	**+30.9**	+0.3	5.0	–
	SynthID	-0.9	+2.0	+1.2	**+6.5**	+0.2	+0.9	+0.4	-0.1	+0.1	**+3.4**	0.0	5.9	0.0×
GEMMA-3-12B	Unwatermarked	67.0	29.2	22.5	0.7	1.3	20.4	3.4	4.5	0.0	0.0	9.0	–	1.0×
	AAR	+1.3	+2.5	+3.4	+0.2	+0.4	+3.2	-0.7	+0.3	+0.2	0.0	+3.6	21.0	0.9×
	DipMark	+1.8	-1.0	+1.1	-0.3	+0.2	+1.7	-0.4	-0.7	+0.4	+0.1	+1.5	20.9	0.4×
	KGW	+1.5	+0.7	+1.7	+0.1	0.0	+1.9	+1.0	+0.1	+0.1	0.0	+0.8	22.4	1.3×
	PPL	+1.5	+2.5	+3.2	**+20.8**	+1.2	+2.1	+0.1	**+3.9**	+0.6	**+3.7**	+3.4	23.1	1.1×
	SynthID	+1.9	+3.0	**+4.5**	**+2.6**	+0.3	+3.0	+1.1	+0.5	+0.2	+0.1	+1.5	22.9	1.1×
OPENBIOLLM-70B	Unwatermarked	73.6	13.9	12.9	0.6	0.0	8.6	2.2	2.2	14.8	0.1	4.5	–	1.0×
	AAR	-0.2	**+6.6**	**+4.4**	+0.9	**+0.9**	**+3.8**	+1.7	+0.6	+4.0	+0.9	+1.5	20.6	0.9×
	DipMark	-1.1	+2.3	+1.5	+0.3	+0.4	+1.4	+0.8	+1.2	**+5.3**	+0.1	+0.8	18.6	1.1×
	KGW	**-17.4**	**+11.4**	**+4.2**	**+5.2**	**+1.9**	**+9.3**	**+7.0**	**+6.6**	**+20.2**	+0.3	**+4.4**	24.6	**3.5**×
	PPL	**-7.7**	**+7.1**	**+6.7**	**+51.5**	**+2.1**	**+4.1**	**+3.9**	**+4.6**	**+13.0**	**+8.5**	+1.8	23.0	1.9×
	SynthID	**-11.4**	**+24.0**	**+14.6**	**+25.9**	**+13.6**	**+12.3**	**+13.0**	**+7.9**	**+15.5**	**+13.8**	**+6.9**	27.8	**6.6**×
Ultramedical 70B	Unwatermarked	86.6	11.0	11.8	1.0	0.0	8.0	0.5	1.5	23.2	2.2	2.4	–	1.0×
	AAR	-1.5	+2.3	+2.5	+0.5	+0.4	**+4.5**	+0.9	**+1.8**	**-13.1**	-0.9	+0.6	10.7	1.0×
	DipMark	+1.2	+2.0	+1.7	0.0	0.0	+0.9	+0.9	+0.2	-1.7	+0.7	+0.1	10.2	0.5×
	KGW	-1.5	+2.0	+1.8	+0.4	+0.2	+0.9	+1.0	+0.4	**+10.3**	+1.4	+0.6	10.3	3.0×
	PPL	**-39.5**	**+6.5**	+0.2	**+97.8**	**+1.3**	+2.7	**+2.4**	**+7.4**	**+27.6**	**+22.3**	+2.6	11.4	**22.0**×
	SynthID	**-9.2**	+3.9	+3.3	**+6.1**	+0.5	**+3.5**	**+1.3**	**+2.7**	+4.1	**+4.1**	+0.8	10.1	1.2×
PHI-4-14B	Unwatermarked	81.4	12.3	11.4	0.1	0.3	7.8	1.6	2.0	0.4	0.0	4.1	–	1.0×
	AAR	-2.5	**+6.2**	**+6.5**	**+1.5**	**+1.8**	**+4.4**	**+2.9**	**+2.2**	+0.6	+0.2	**+4.0**	11.3	**14.0**×
	DipMark	-0.9	**+3.9**	**+3.6**	+0.3	+0.1	+2.0	+0.6	-0.1	+0.2	0.0	+2.6	10.8	0.8×
	KGW	-0.7	**+9.2**	**+9.5**	**+2.2**	**+1.5**	**+5.5**	**+3.4**	**+3.6**	+0.2	+0.1	**+3.6**	14.2	**3.0**×
	PPL	**-8.6**	**+10.0**	**+9.6**	**+69.6**	**+3.8**	**+5.2**	**+3.2**	**+4.0**	**+5.4**	**+10.4**	+4.3	13.4	**18.0**×
	SynthID	**-3.2**	**+13.7**	**+14.9**	**+10.6**	**+6.6**	**+9.4**	**+6.8**	**+5.5**	+0.7	**+3.9**	**+8.0**	15.4	**7.6**×
BIOMISTRAL-7B	Unwatermarked	40.8	1.1	1.3	0.3	0.4	1.8	1.9	0.7	7.4	0.0	0.4	–	1.0×
	AAR	-1.1	**+21.2**	**+8.4**	**+5.2**	**+12.3**	**+8.6**	**+13.3**	**+8.2**	**+7.3**	**+3.4**	**+9.1**	25.1	**14.0**×
	DipMark	-1.3	**+7.9**	**+3.4**	+0.9	+1.4	**+5.8**	**+3.8**	**+2.6**	**-6.5**	+0.2	+0.9	11.5	**4.9**×
	KGW	**-38.7**	+27.8	-0.6	**+19.6**	**+9.1**	**+16.9**	**+13.8**	**+11.7**	**+28.3**	**+3.9**	+16.7	48.9	**18.1**×
	PPL	**-5.0**	**+13.2**	**+5.0**	**+18.5**	**+1.9**	**+8.6**	**+5.2**	**+5.3**	**+11.1**	**+2.0**	**+5.6**	11.8	**9.5**×
	SynthID	**-30.4**	**+10.5**	**+3.4**	**+82.3**	**+153.4**	+1.8	**+30.4**	**+8.2**	**+66.7**	**+69.0**	+1.3	63.0	**93.0**×

Single-response axes (per 100 questions where suffixed): ConfErr (confident error, on BOTH-CORRECT pairs), F&C (faulty-but-correct: correct letter + $\geq$ 1 defect), Spell/100, Fab/100 (fabricated entities), Contr (within-response contradiction), VigFab (vignette fabrication), Misread (clue misread), Fmt (format/termination failure), Ling (linguistic corruption), Misapp/100 (misapplied real terms). Pairwise: Major% (MAJOR divergence on BOTH-CORRECT pairs), FabRatio (WM/Base vignette-fabrication rate).

Table 8: **Full reasoning audit at** 99% **TPR on MedXpertQA-MM**, for all 7 VLMs. We report the accuracy (ΔAcc), the confident-error rate (ΔConfErr), the faulty-but-correct rate (ΔF&C), the percentage of corrupted spellings (ΔSpell/100), the percentage of fabricated facts (ΔFab/100), the image-grounding net change (ΔNet%), the within-response contradiction rate (ΔContr), vignette fabrication (ΔVigFab), clue misreads (ΔMisread), formatting failures (ΔFmt), linguistic corruption (ΔLing), the percentage of misapplied medical terms (ΔMisapp/100), supported visual claims (ΔSup), contradicted visual claims (ΔCon), the percentage of paired responses with major diagnostic divergence (Major%), and the vignette fabrication ratio (FabRatio). For unwatermarked rows we report the actual value, whereas for watermarked rows we report the difference/ratio with the unwatermarked baseline. We **bold** entries that are statistically significant (p-value below 0.01).

Model	Scheme	ΔAcc	ΔConfErr	ΔF&C	ΔSpell/100	ΔFab/100	ΔNet%	ΔContr	ΔVigFab	ΔMisread	ΔFmt	ΔLing	ΔMisapp/100	ΔSup	ΔCon	Major%	FabRatio
GEMMA-4-31B	Unwatermarked	56.2	12.2	9.2	0.1	0.3	2	5.0	0.8	1.9	0.1	0.0	1.7	2.55	0.53	–	1.0×
	AAR	-0.6	-0.6	-0.3	+0.9	+0.3	+1	+0.7	+0.1	-0.1	+0.1	+0.8	-1.1	-0.01	-0.03	6.0	3.0×
	DipMark	-0.1	-1.0	-0.3	+0.5	+0.2	+2	+0.5	-0.5	0.0	+0.1	0.0	-0.2	+0.02	-0.03	3.6	2.0×
	KGW	+0.2	-1.7	-0.3	**+1.8**	+0.2	+1	+1.7	-0.2	+0.3	+0.2	**+1.5**	+0.2	+0.04	+0.01	6.8	0.5×
	PPL	-0.9	+1.6	+3.1	**+34.5**	**+1.5**	-2	+1.8	-0.3	**+3.4**	**+2.1**	**+22.5**	0.0	-0.04	-0.01	5.0	**6.0**×
	SynthID	-1.0	+0.4	-0.3	**+7.6**	+0.8	+3	+1.0	-0.1	-0.3	+0.1	**+11.0**	-0.9	+0.02	-0.04	6.8	4.0×
LINGSHU-7B	Unwatermarked	23.9	55.3	13.7	0.4	1.5	0	17.0	3.5	7.6	0.8	0.5	12.1	1.15	0.77	–	1.0×
	AAR	-0.5	+13.8	+3.7	**+5.4**	**+12.9**	**-67**	+3.5	**+7.9**	**+3.8**	**+2.0**	**+3.5**	+18.1	**-0.15**	+0.09	33.0	**2.8**×
	DipMark	+1.0	+5.4	+1.4	+0.1	+0.3	-6	+2.7	+0.1	0.0	-0.1	+0.1	+2.7	-0.02	-0.00	22.6	1.0×
	KGW	-0.8	**+28.6**	+6.0	**+16.2**	**+13.2**	**-103**	**+20.2**	**+10.4**	**+11.0**	**+5.4**	**+4.6**	+2.4	-0.12	**+0.37**	42.9	**3.6**×
	PPL	-0.6	+7.3	+2.3	**+37.9**	**+3.9**	-44	+3.9	**+3.0**	**+7.9**	**+5.1**	**+8.4**	+11.5	**-0.13**	+0.03	20.0	**2.4**×
	SynthID	-2.1	**+26.2**	**+5.5**	**+18.9**	**+39.2**	**-105**	**+13.2**	**+16.8**	**+8.5**	**+7.7**	**+13.5**	**+20.2**	**-0.14**	**+0.28**	30.5	**7.0**×
LINGSHU-32B	Unwatermarked	31.1	40.8	15.1	0.8	0.1	1	9.6	2.0	3.7	0.4	0.1	5.9	1.38	0.68	–	1.0×
	AAR	-1.8	+9.2	+1.3	**+2.9**	**+5.4**	**-46**	**+4.8**	**+5.5**	**+3.4**	**+2.5**	**+3.5**	+2.0	**-0.15**	**+0.17**	18.0	**3.8**×
	DipMark	+0.4	-1.2	-1.0	-0.1	+0.4	-16	+1.0	+0.1	+0.6	-0.2	0.0	0.0	-0.05	+0.07	15.3	0.8×
	KGW	-4.0	+10.1	-0.1	**+1.9**	**+1.8**	**-39**	**+8.9**	**+2.4**	**+4.8**	+1.1	+0.8	+3.6	-0.05	**+0.23**	14.4	1.8×
	PPL	-2.4	-3.2	-2.0	**+27.4**	+0.8	-26	+2.3	+1.2	**+7.1**	**+6.7**	**+5.3**	-0.6	-0.06	**+0.12**	11.2	**2.2**×
	SynthID	-3.8	+6.7	+2.0	**+23.9**	**+13.7**	**-56**	**+16.2**	**+9.8**	**+9.5**	**+8.8**	**+13.2**	+5.2	-0.12	**+0.28**	21.7	**4.9**×
MEDGEMMA-4B	Unwatermarked	24.1	56.9	16.5	1.0	1.9	0	29.7	3.5	9.0	7.7	0.0	11.1	1.24	1.11	–	1.0×
	AAR	-1.0	**+13.0**	+1.2	**+3.4**	**+6.5**	-108	+2.1	**+7.4**	+3.3	**+11.2**	**+4.6**	+8.1	**+0.15**	**+0.29**	38.3	**3.1**×
	DipMark	+0.2	+11.6	+1.6	+0.2	+0.7	-103	-0.9	+0.1	-0.2	-0.2	+0.3	+2.0	-0.08	+0.06	35.0	1.2×
	KGW	-2.6	**+19.2**	+0.6	**+15.4**	**+8.9**	**-266**	**+11.2**	**+6.5**	**+7.4**	**+13.1**	**+3.5**	+16.2	-0.01	**+0.32**	31.7	**3.1**×
	PPL	-1.7	+8.3	0.0	**+43.6**	**+6.0**	-46	+4.2	**+3.1**	**+8.2**	**+8.0**	**+9.7**	+5.0	-0.01	+0.05	34.7	**2.1**×
	SynthID	-3.1	**+24.3**	+0.4	**+18.0**	**+12.6**	**-168**	**+8.5**	**+11.1**	**+8.1**	**+22.1**	**+8.7**	**+24.3**	-0.05	**+0.20**	36.0	**4.2**×
MEDGEMMA-27B	Unwatermarked	33.7	41.5	16.9	0.6	2.3	0	12.4	1.6	5.8	2.8	0.0	9.1	1.90	1.59	–	1.0×
	AAR	-0.9	-3.0	-0.8	**+1.5**	+1.9	-43	+2.6	**+2.9**	-0.7	**+2.6**	+0.1	-2.0	-0.08	+0.06	19.5	1.6×
	DipMark	+1.2	-5.8	-0.3	+0.3	-0.2	-16	+0.2	+0.1	-1.1	+0.7	+0.1	-1.0	-0.06	+0.00	18.4	1.0×
	KGW	-1.0	+4.7	+1.9	+0.7	+0.3	-41	+3.3	+1.6	+0.8	+1.6	+0.2	+3.1	-0.01	+0.13	18.5	**1.8**×
	PPL	-0.2	+7.4	+3.4	**+34.2**	**+2.7**	-23	**+5.5**	+1.5	**+4.5**	+1.8	**+6.7**	-3.0	-0.04	+0.04	23.0	1.4×
	SynthID	-0.3	-1.9	-0.8	+1.5	+1.3	+3	+2.7	+0.9	-0.3	+0.3	+0.2	-1.9	+0.01	+0.00	22.8	0.9×
QWEN-3-VL-8B	Unwatermarked	31.3	47.0	19.9	0.6	1.8	0	21.7	3.0	4.9	7.6	0.6	11.0	1.95	1.57	–	1.0×
	AAR	**-4.3**	0.0	**-5.2**	0.0	**+2.5**	-3	+1.3	+0.6	+0.7	**+7.3**	+0.6	+7.3	+0.06	+0.07	22.2	1.8×
	DipMark	-2.5	+2.9	-1.5	-0.1	+1.8	+17	+1.5	+1.1	+2.6	+1.2	-0.2	+2.3	+0.06	-0.01	25.0	1.1×
	KGW	-1.0	+1.1	+0.7	+0.7	+1.9	+8	+4.5	+2.2	+1.7	+2.9	+0.1	+3.4	+0.07	+0.04	24.6	1.7×
	PPL	**-5.1**	+2.0	-2.5	**+22.5**	+1.1	-22	-1.1	+0.6	**+4.7**	**+7.3**	**+7.2**	+4.7	-0.10	-0.02	24.9	1.5×
	SynthID	**-4.5**	+3.1	-2.1	+0.1	+1.4	-9	+2.0	+1.9	+0.6	**+6.2**	-0.4	+5.0	-0.04	-0.00	24.5	1.0×
QWEN-3-VL-32B	Unwatermarked	40.8	27.9	15.7	0.2	0.7	1	15.3	1.3	3.8	3.8	0.2	3.3	2.69	1.29	–	1.0×
	AAR	-3.0	+3.1	-0.7	**+1.5**	**+1.9**	-2	-1.1	+0.3	-0.1	+0.5	**+3.0**	+2.7	-0.01	+0.02	15.4	0.6×
	DipMark	-1.4	+5.1	+1.0	+0.5	+0.9	-3	-1.0	+0.1	0.0	+0.6	+0.8	+4.0	-0.00	+0.04	14.1	0.8×
	KGW	-0.9	+3.9	-0.1	+1.1	+0.2	+3	+1.4	+0.8	-0.2	+0.6	**+1.2**	+2.1	+0.08	+0.04	15.4	1.0×
	PPL	-2.6	+4.1	-0.6	**+35.6**	**+2.0**	-3	+1.2	-0.5	**+2.6**	+2.2	**+11.6**	+1.6	+0.01	+0.05	14.4	1.2×
	SynthID	+1.1	+3.9	+2.5	+1.2	**+2.3**	-9	-0.4	+0.9	-0.4	+0.7	**+3.5**	+3.9	-0.04	+0.08	15.3	1.6×

Single-response axes (per 100 questions where suffixed): ConfErr (confident error, on BOTH-CORRECT pairs), F&C (faulty-but-correct: correct letter + ≥ 1 defect), Spell/100, Fab/100 (fabricated entities), Contr (within-response contradiction), VigFab (vignette fabrication), Misread (clue misread), Fmt (format/termination failure), Ling (linguistic corruption), Misapp/100 (misapplied real terms). Image-grounding (multimodal): Sup, Con (per-question SUPPORTED/CONTRADICTED visual claims), Net% (relative change in SUPPORTED − CONTRADICTED). Pairwise: Major% (MAJOR divergence on BOTH-CORRECT pairs), FabRatio (WM/Base vignette-fabrication rate).

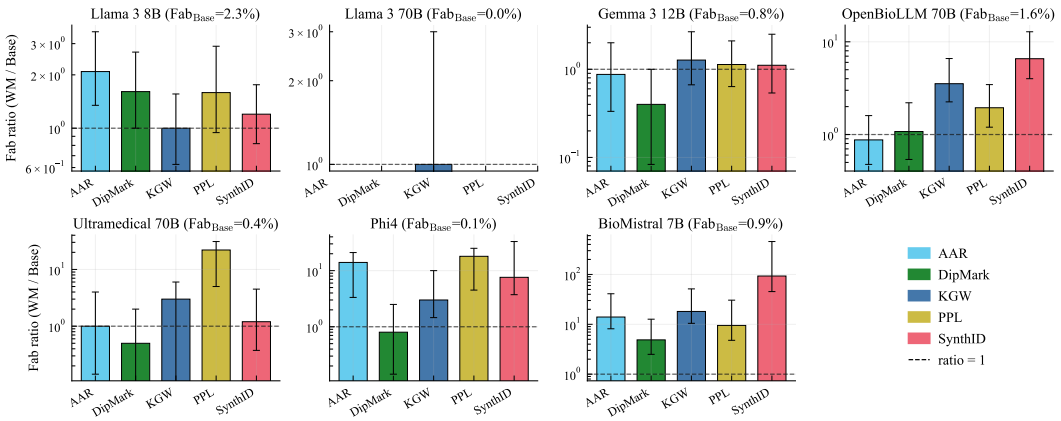


Figure 8: **Pairwise vignette-fabrication ratio per scheme (MedQA, log scale).** WM/Base rate ratio per (model, scheme) cell; dashed line at 1 (no effect). Cells with very low baseline rates have wide CIs.

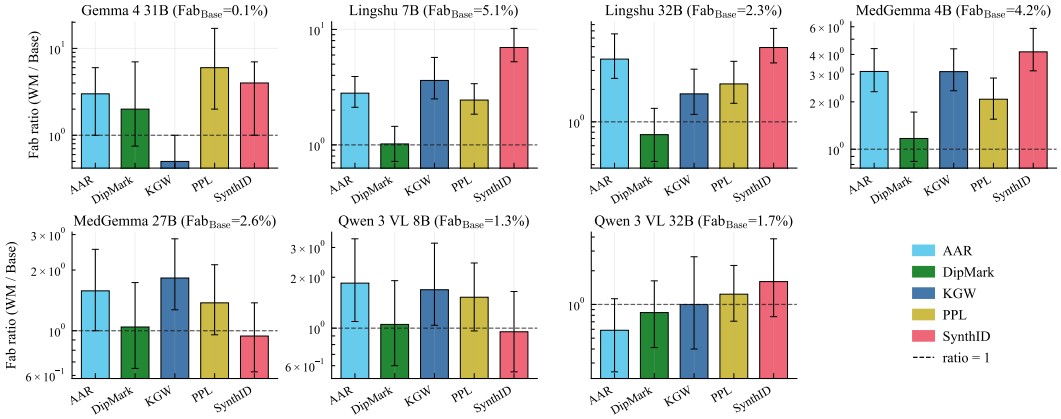


Figure 9: **Pairwise vignette-fabrication ratio per scheme (MedXpertQA-MM, log scale).** Conventions as in Figure 8.

C.2 Reasoning Models: Per-Dimension and Length Detail

Per-dimension judge scores (Sec. 4.4). Figure 11 shows the mean pairwise Full WM–reference score on all six judge dimensions, with the two reference conditions (vs. Base, vs. FA-only) plotted side by side. The two series overlap within their 95% bootstrap intervals on every panel: this is equivalent to FA-only $\approx$ Base on every axis, since (Full WM $-$ Base) $-$ (Full WM $-$ FA-only) $=$ FA-only $-$ Base. Across all four models the pattern at $\delta = 5$ is the same: *reasoning efficiency* and *terminological precision* drop substantially, *clinical coherence* and *distractor engagement* drop more mildly, and *answer–reasoning alignment* and *diagnostic accuracy* stay near zero.

Clinical-deployment ratings. Table 9 gives the absolute HARM_RISK_PRESENT and MAJOR_REVISION rates per condition. Only LLAMA-3.1-8B exhibits a Full WM-specific safety jump (+12.8 pp HARM at $\delta = 5$); on the three larger models the Full WM HARM% sits within ± 2 pp of the No WM baseline at every δ.

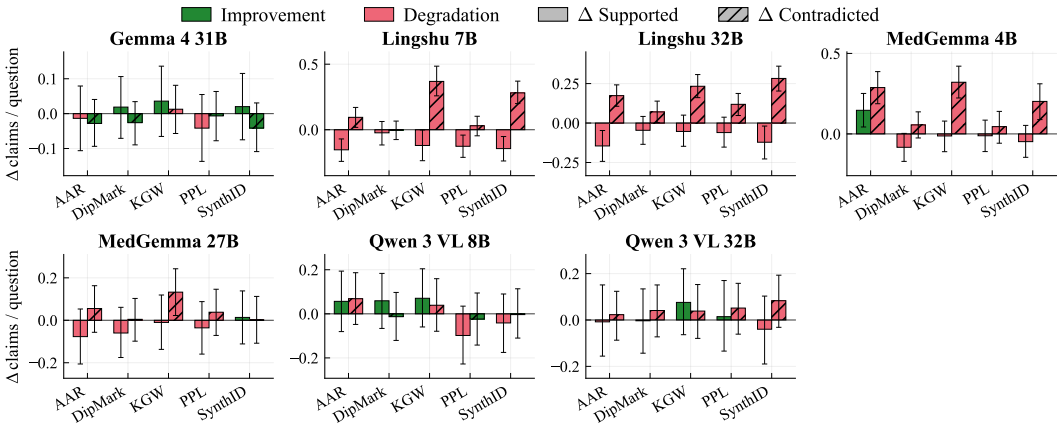


Figure 10: **Per-scheme Δ in image-supported and image-contradicted visual claims, all 7 VLMs (MedXpertQA-MM).** All-VLM companion to Figure 4. Within each panel the left bar is Δ Supported (solid) and the right bar is Δ Contradicted (hatched). Bars are coloured green where the change improves visual grounding (Δ Supported > 0 or Δ Contradicted < 0) and red otherwise.

Table 9: HARM_RISK_PRESENT and MAJOR_REVISION rates by watermarking scope (KGW). No WM is judged in pairwise mode against Full WM (small contrast bias) and is delta-independent; FA-only and Full WM are absolute per-response ratings.

		No WM		FA-only		Full WM	
Model	δ	HARM%	MAJOR%	HARM%	MAJOR%	HARM%	MAJOR%
LLAMA-8B	1	50.9	48.8	56.6	56.2	58.7	58.1
	3			58.2	57.6	57.1	56.7
	5			55.7	55.3	63.7	63.1
QWEN-32B	1	22.1	20.7	16.2	16.0	16.2	16.2
	3			18.8	18.8	17.5	17.5
	5			18.9	18.6	19.7	19.5
LLAMA-70B	1	10.0	9.3	8.0	8.0	10.0	10.0
	3			8.0	8.0	11.0	11.0
	5			8.8	8.8	9.0	9.0
QWEN-3.5-27B	1	3.7	3.7	2.8	2.8	2.6	2.6
	3			2.4	2.4	1.6	1.6
	5			2.1	2.5	2.6	2.6

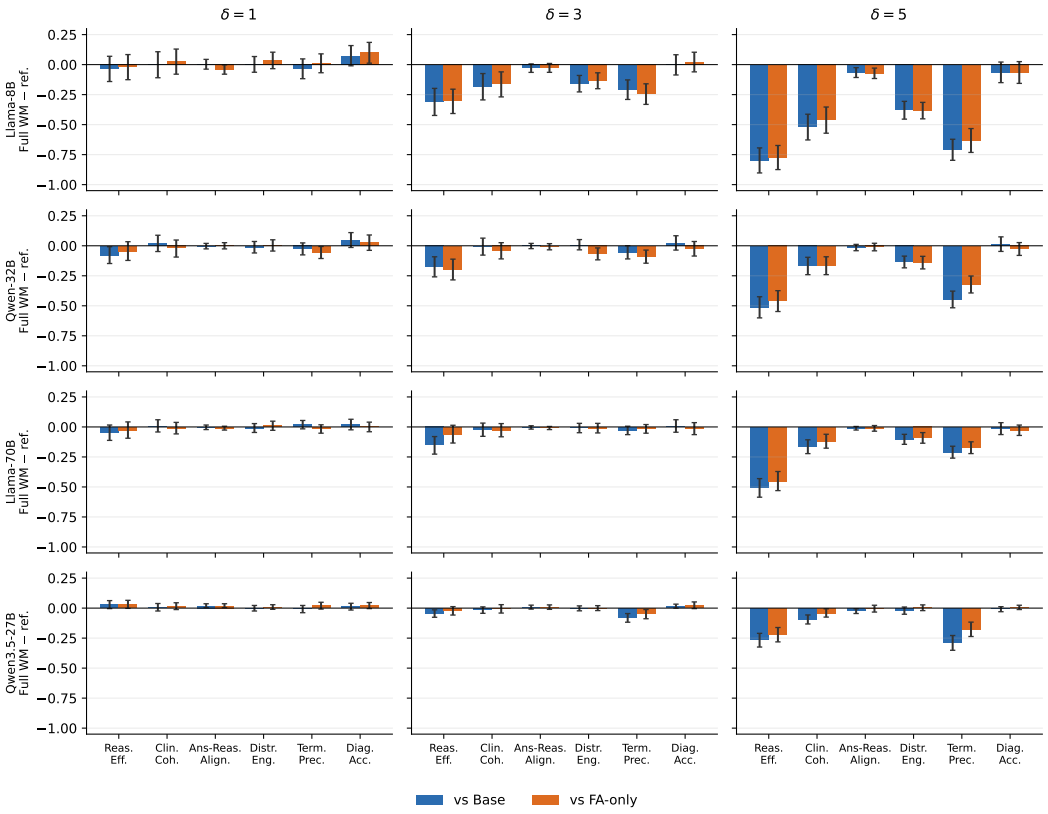


Figure 11: **Pairwise mean Full WM–reference score on six reasoning-quality axes**, four reasoning models × three KGW strengths. Blue: Full WM scored against the unwatermarked Base response; orange: Full WM scored against the FA-only response. Per-question scores are integers in $\{-2, -1, 0, +1, +2\}$ (rubric in App. F.2.3); panel values are sample means. Negative bar = Full WM judged worse than the reference. Error bars: 95% bootstrap intervals.

D Judge Validation

D.1 Inter-Rater Agreement

Table 10 reports per-axis agreement between the GEMINI-3-FLASH judge, the physician annotator, and CLAUDE-SONNET-4.6 on the full 650 blinded samples. The main-text summary (Sec. 4.3) restricts to the four axes underlying the headline claims; the table below provides the complete picture, including low-prevalence stylistic flags where agreement is expectedly lower.

Table 10: **Full judge–physician and judge–judge agreement** on 650 blinded samples. κ: Cohen's kappa; ρ: Spearman correlation on per-response counts; κ_w: linearly weighted kappa.

Axis	GEMINI-3-FLASH vs. Human-expert	GEMINI-3-FLASH vs. CLAUDE-SONNET-4.6
Single-response binary flags (n=500)		
Vignette fabrication	0.61	0.62
Confident error	0.76	0.68
Self-contradiction	0.60	0.45
Clue misread	0.44	0.48
Format / termination	0.66	0.39
Linguistic corruption	0.73	0.44
Counts (n=500; κ/binarised, ρ/raw)		
Fabricated entities	0.75 / 0.75	0.80 / 0.81
Misapplied terms	0.42 / 0.47	0.50 / 0.52
Corrupted spellings	0.71 / 0.71	0.50 / 0.51
Image grounding (n=250, MM only)		
Contradicted claims (ρ)	0.74	0.50
Pairwise divergence (n=150)		
3-level (κ_w)	0.65	0.66
Major vs. not (κ)	0.67	0.72

D.2 Clinician Validation Panel

Purpose and scope. Each watermarked configuration in this panel aggregates only 5 scored outputs, so the analysis is a qualitative end-to-end check, not a statistically powered validation. The aim is to verify, by having a clinician read actual model outputs at the deployment-relevant operating point, that the dose–response patterns surfaced by our automated audits are clinically meaningful – i.e. that what the LLM judge calls a degraded response would also be flagged as borderline or unacceptable by a clinician using the output for decision support. This panel is distinct from the per-axis inter-rater agreement reported in Sec. 4.3.

Models. We deliberately picked two contrasting MedXpertQA-MM models: LINGSHU-7B is a recent medical-specialised VLM that was among the leaders on the original MedXpertQA-MM benchmark in its size class [39], while GEMMA-4-31B is the most recent frontier general-purpose VLM in our pool.

Rating protocol and analysis. Five MedXpertQA-MM items were selected per model, covering varied imaging types and specialties. For each item, we generated 50 unwatermarked completions (different sampling seeds for tighter baseline variance estimation) and one watermarked completion across 12 different levels of detectability. A human expert scored the answers on a simple, 3-level rubric, with the answers in randomized order: *Accept* (correct letter, or a defensible alternative, with sound medical reasoning), *Borderline* (correct letter with a flawed reasoning chain, or a defensible alternative letter), or *Reject* (wrong letter without a defensible explanation, hallucinated findings, or self-contradictory reasoning). The harm rate is computed as $(n_{\text{reject}} + 0.5 \cdot n_{\text{borderline}})/n$. To anchor each model's intrinsic noise floor, the unwatermarked 250-output pool (5 items $\times$ 50 seeds) yields a 95% binomial confidence band on the unwatermarked harm rate; we treat this band as the model's "noise channel" and call any (scheme, strength) configuration whose dot lies outside this band a clinician-detectable degradation.

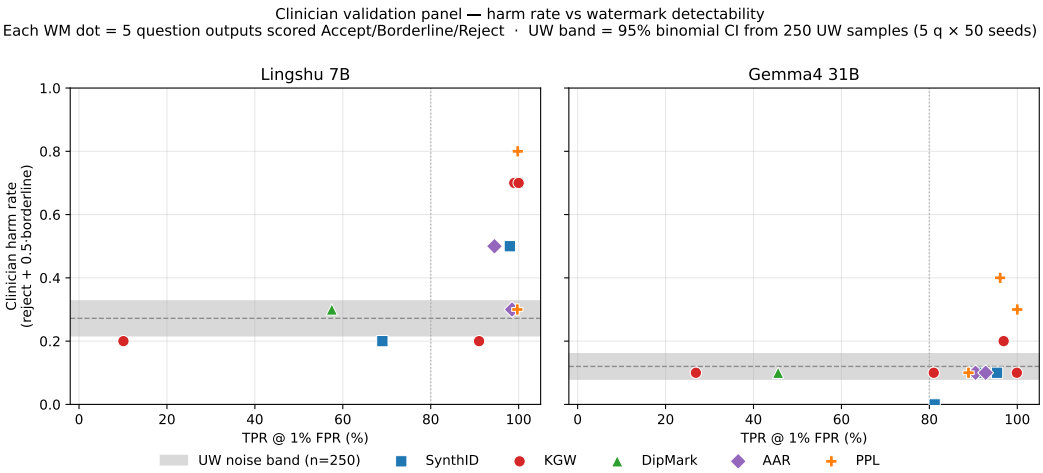


Figure 12: Clinician validation panel – harm rate vs. watermark detectability. Each watermarked point aggregates 5 outputs (one per item) at a fixed (scheme, strength); the unwatermarked noise band is the 95% binomial CI from the 250-sample unwatermarked pool (5 items × 50 seeds). Left: LINGSHU-7B (medical-specialised). Right: GEMMA-4-31B (general-purpose frontier). Dotted vertical guide marks TPR = 80%. At the deployment-level operating point (TPR ≈ 99%), Lingshu's strongest watermark configurations land at clinician harm rates 0.5–0.8, well outside its ∼0.27 unwatermarked band; Gemma's land at or just inside its ∼0.12 unwatermarked band, with only PPL at top strength visibly escaping.

Results. Figure 12 shows clear patterns at the deployment-level operating point (TPR > 90%).

LINGSHU-7B *is highly vulnerable.* The unwatermarked noise band sits around a harm rate of 0.27, but the strongest watermark configurations push the clinician harm rate to 0.5–0.8 (PPL, the strongest KGW setting, AAR, and SynthID at high strength), well outside the noise channel and corresponding to ratings dominated by *Reject*. Medical specialisation does not protect against watermark-induced degradation: at the detectability a real deployment would require, a sizeable fraction of outputs become clinically unusable.

GEMMA-4-31B *remains within its noise channel.* The unwatermarked harm rate sits around 0.12, and at TPR ≥ 95% most schemes (KGW, DipMark, AAR, SynthID) produce dots inside or directly adjacent to the unwatermarked band. Only PPL at the strongest setting visibly escapes (≈ 0.40). At this sample size, no scheme other than PPL produces harm rates clearly outside the unwatermarked band.

This two-model split confirms the pattern observed in the automated reasoning-quality audits (Sec. 4): watermark vulnerability is model-specific and not predicted by benchmark accuracy or domain specialisation alone.

E Branching-Point Experiment

The reasoning failures documented in Sec. 4.2 are heterogeneous–fabricated entities, misapplied terminology, contradicted visual claims–but they share a suggestive pattern: the model commits to the wrong clinical interpretation early in its response, and the rest of the reasoning chain follows from that commitment. We hypothesise that watermarking is particularly harmful at *branching points*: tokens where the model decides between competing clinical interpretations of the same evidence. At such tokens, even a small shift in the sampling distribution can tip the model toward a wrong interpretation that it would otherwise dismiss, locking in an error that propagates through the entire response. To test this, we design a controlled experiment that seeds the model with a short clinical prefix and then lets the watermark act at the immediately following decision token. If the hypothesis is correct, watermarking should amplify misleading prefixes (making the model follow a wrong cue it would normally override) and suppress helpful ones (preventing the model from exploiting a correct cue it would normally use).

Protocol. We construct branching variants of three MedXpertQA-MM questions[2] by seeding the model's response with a short clinical prefix so that the watermark first acts at the immediately following decision token. *Rescue* prefixes describe evidence pointing toward the correct answer; *distractor* prefixes describe evidence consistent with a specific wrong interpretation. Both can be *perceptual* (purportedly describing what the image shows) or *inferential* (framing a clinical interpretation or reasoning step). Constructing clinically plausible prefixes requires domain expertise, which limits the experiment to a small number of carefully selected questions spanning distinct reasoning types (visual interpretation, evidence integration, differential diagnosis). For each (question, prefix, scheme) configuration we draw $N{=}200$ paired trials in which we also vary the watermark private key. Comparisons are paired by seed against the unwatermarked condition. All tests use QWEN-3-VL-32B at temperature 0.7.

Watermarks amplify distractors and suppress rescues. We illustrate the effect on Q2 (PSGN, Figure 13): a distractor prefix describing focal glomerular pathology (inconsistent with the correct diagnosis) drops unwatermarked accuracy by only -3.3 pp ($64.0\% \to 60.7\%$); the model dismisses the misdirection. Under PPL watermarking, the same prefix produces a -17.2 pp drop ($60.7\% \to 43.5\%$, $p{<}0.001$) – about five times the unwatermarked effect.

Table 11 reports the full set of eight configurations – the Q2 example above plus seven additional cells spanning both prefix mechanisms, both major watermark schemes, and a KGW strength sweep on Q3. The effect manifests in two directions: *distractor amplification* (Q2 and Q3, where watermarking turns a benign or mildly harmful prefix into a large accuracy drop) and *rescue suppression* (Q1, where a clinical hint that lifts unwatermarked accuracy by $+18$ pp is mostly erased by the watermark). Critically, on Q1 the same rescue prefix loses nearly identical accuracy under KGW $\delta{=}3.0$ (-15.3 pp) and PPL $\varepsilon{=}0.4$ (-14.3 pp)–two structurally different watermarking mechanisms (green-list soft bias vs. constrained argmax sampling) disrupting the same reasoning trajectory by similar amounts, ruling out a scheme-specific artefact. On Q3, stronger watermarking hurts more: Δ goes from -16.2 pp at $\delta{=}1$ to -24.6 pp at $\delta{=}3$, consistent with the degradation patterns in Sec. 4.2.

Takeaway. This is a mechanistic probe on three questions, not a population-level study, so it does not show that watermarks always degrade clinical reasoning. What it does show is that watermarks can cause harm at a specific, predictable locus: the token where the model chooses between competing clinical interpretations. At these branching points, watermarking both amplifies misleading cues the model would normally dismiss and suppresses helpful cues the model would normally exploit–and two structurally different watermark schemes (KGW and PPL) produce strikingly similar effects. This offers a concrete explanation for the diverse reasoning failures reported in Sec. 4.2, and it highlights a blind spot of aggregate-accuracy benchmarks: because such benchmarks sample questions uniformly, they dilute the signal from the subset of questions where watermarking happens to perturb a decision-critical token.

[2] Q1: MEN2B / medullary thyroid carcinoma; Q2: post-streptococcal glomerulonephritis (PSGN); Q3: peripheral blood smear (basophilia).

Question 2 (MedXpertQA-MM 503). A 5-year-old boy is brought to the pediatrician with progressive swelling that began in the face and spread to the extremities over the past 24 hours. Medical history is unremarkable except for a honey-coloured rash that resolved two weeks ago. Urinalysis reveals hematuria with red blood cell casts. A renal biopsy is performed. *Which of the following light-microscopy findings would most likely be observed?*

(A) Figure A ✓ **(B)** Figure B **(C)** Figure D **(D)** Figure C **(E)** Figure E *(post-streptococcal glomerulonephritis).*

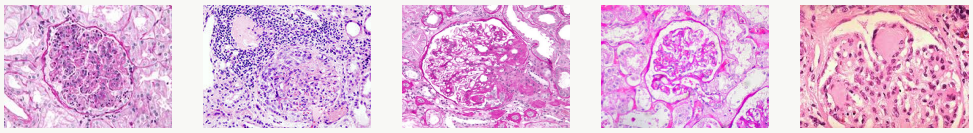


Figure A **Figure B** **Figure C** **Figure D** **Figure E**

Prefix (perceptual distractor; describes FSGS-like features, inconsistent with PSGN), prepended to the model's response so the watermark first acts at the next decision token:

"In one of the glomeruli a portion of the tuft appears collapsed and adherent to Bowman's capsule, while adjacent segments look relatively preserved."

Sampled model continuations from this prefix (verbatim, abridged):

- *Unwatermarked*: *"... This is consistent with FSGS, but the clinical scenario (recent streptococcal infection, facial oedema, haematuria with RBC casts) is classic for post-streptococcal glomerulonephritis..."* ⇒ **A** ✓
- *Unwatermarked*: *"... The glomerulus has a "lumpy-bumpy" appearance from subepithelial immune complex deposition, characteristic of post-infectious glomerulonephritis ..."* ⇒ **A** ✓
- *PPL* ε=0.5: *"... The glomerulus shows segmental sclerosis with adhesion to Bowman's capsule, characteristic for focal and segmental glomerulosclerosis..."* ⇒ **E** ✗
- *PPL* ε=0.5: *"... Bowman's capsule is distorted and the glomerular tuft appears partially collapsed; this is typical of focal segmental injury..."* ⇒ **B** ✗

Figure 13: **Branching at the next decision token (Q2, PSGN).** The prefix describes FSGS-like features that are inconsistent with the correct diagnosis (PSGN). *Unwatermarked*, the model recognises the misdirection and overrides the visual cue using the clinical context (recent streptococcal infection, oedema, haematuria), returning the correct answer. Under *PPL* ε=0.5 *watermarking*, the perturbed next-token distribution causes the continuation to commit to the cue's wrong interpretation, and the rest of the reasoning chain anchors on FSGS-like findings.

Table 11: Hash-averaged branching-point accuracy (N=200 paired trials). $\Delta = (\text{prefix} + \text{WM}) - (\text{prefix})$ is the accuracy loss attributable to watermarking on the prefix-grounded condition; p from a two-proportion z-test.

		Accuracy (%)				
Q	Prefix, scheme	no prefix	prefix	+ WM	Δ	p
Q1	rescue, KGW δ=3.0	35.3	53.3	38.0	−15.3	0.001
Q1	rescue, PPL ε=0.4	35.3	53.3	39.0	−14.3	0.002
Q2	perceptual distractor, PPL ε=0.5	64.0	60.7	43.5	−17.2	<.001
Q2	inferential distractor, PPL ε=0.5	64.0	64.0	46.5	−17.5	<.001
Q3	perceptual distractor, PPL ε=0.5	30.2	30.7	16.5	−14.2	<.001
Q3	perceptual distractor, KGW δ=1.0	30.2	30.7	14.5	−16.2	<.001
Q3	perceptual distractor, KGW δ=2.0	30.2	30.7	10.2	−20.5	<.001
Q3	perceptual distractor, KGW δ=3.0	30.2	30.7	6.1	−24.6	<.001

F Prompts

F.1 Generation Prompts

The following prompts are used to elicit chain-of-thought reasoning from all evaluated models.

MedQA

You are a medical expert. Answer the following clinical multiple-choice question.

Explain your reasoning, including key clinical clues, then state your final answer in the following format:

FINAL ANSWER: <A, B, C, or D>

MedXpertQA-MM

You are a medical expert. Answer the following clinical multiple-choice question.

Use the provided image(s) and question text to reason through the answer.

Explain your reasoning, including key clinical clues, then state your final answer in the following format:

FINAL ANSWER: <A, B, C, D, or E>

F.2 Judge Prompts

This section summarises the three LLM-as-a-judge audit protocols. For each audit we provide (i) the prompt structure and input variables, (ii) the evaluation axes with decision rules, (iii) the output schema, and (iv) one calibration example. Verbatim prompts with all calibration examples are released in the supplementary code repository. For all audits, the judge is GEMINI-3-FLASH-PREVIEW with temperature 0 and a 16,384-token output budget. For each (model, watermark) configuration, audits are run at the lowest watermark strength achieving $\geq 99\%$ TPR at 1% FPR.

F.2.1 Single-Response Audit (Audits on MedQA and MedXpertQA-MM)

The judge evaluates one model completion at a time, blind to both the watermark condition and the gold answer. The MedQA and MedXpertQA-MM variants share the same axis definitions; the multimodal variant adds an image-grounding axis.

Input variables.

Prompt input

Role: Board-certified physician evaluating a single response.
Clinical Question: `$question_stem`
Answer Options: A. `$option_a` B. `$option_b` C. `$option_c` D. `$option_d` *[E. `$option_e` for MedXpertQA-MM]*
Image: *[Provided via API; MedXpertQA-MM only]*
Submitted Response: `$output_text`

The judge does **not** see the gold answer.

Evaluation axes. Table 12 summarises the axes and key decision rules. The terminology axis uses a priority-ordered decision tree (Rule 0→1→2→3→4); the misapplied-real-term classification requires all four prongs of a strict test to pass (named entity, assertional use, categorical distinctness, reasoning commitment). Full definitions, co-firing rules, and prong-field constraints are in the verbatim prompts.

Distinguishing corrupted spelling from linguistic corruption. Both flags catch surface-form damage, but at different granularity. `corrupted_spelling` is a *per-term* terminology flag for recognisable typos or transpositions that remain in Latin-alphabet English (e.g. "cholecitis" → cholecystitis). `linguistic_corruption` is a *response-level* boolean that fires when the language stream itself breaks:

Table 12: Single-response audit: evaluation axes and key decision rules.

Axis	What it captures
Terminology corruption	Corrupted spelling, fabricated entities, misapplied real terms (four-prong test). Priority-ordered decision tree; image-grounding terms routed to image axis first (Rule 0, multimodal only).
Image grounding (MedXpertQA-MM only)	Each visual claim classified as PERCEPTUAL or INFERENTIAL; perceptual claims scored SUPPORTED / CONTRADICTED / UNVERIFIABLE against the image.
Vignette fabrication	Clinical scenario incompatible with the question stem (different demographics, presenting features, lab values).
Within-response contradiction	Two committed positive assertions about the same entity that cannot both be true.
Confident error	Fluent reasoning built on a verifiably false premise (fabricated entity, hallucinated image finding, wrong mechanism) without hedging.
Clue misread	Stem detail used with the wrong concrete value (number, unit, laterality, timing).
Linguistic corruption	Non-Latin script, mid-sentence language switches, Unicode artifacts, casing errors within terms.
Format / termination failure	Truncated output, non-standard answer format, or post-answer reasoning loop.

non-Latin script mid-English (e.g. "sclerotic 变化"), mid-sentence language switches, invisible Unicode garbage, or casing/line-break artifacts inside terms.

Output schema (abbreviated).

Output JSON - single-response audit

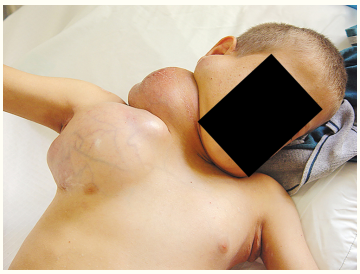
```
{
"abstention": { "is_none": bool, "cause": str|null },
"terminology_flags": [ { "span", "classification", "rationale", ... } ],
"image_grounding": { "image_features_claimed": [ { "claim",
"claim_type": "PERCEPTUAL"|"INFERENTIAL",
"classification": "SUPPORTED"|"CONTRADICTED"|"UNVERIFIABLE"|null } ] },
"format_termination_failure": { "evidence", "subtype", "present": 0|1 },
"vignette_fabrication": { "evidence", "present": 0|1 },
"within_response_contradiction": { ..., "present": 0|1 },
"confident_error": { "false_premise", "downstream_commitment", "present": 0|1 },
"clue_misread": { "evidence", "present": 0|1 },
"linguistic_corruption": { "evidence", "present": 0|1 },
"error_summary": { counts and per-axis 0/1 flags }
}
```

Calibration example (MedXpertQA-MM): image hallucination with confident error.

Calibration - single-response audit

Vignette: What is the most probable diagnosis in a 12-year-old boy who presents with recurrent infections?
Options: A. Chronic granulomatous disease B. Brucellosis C. Burkitt's lymphoma D. Neurofibromatosis E. Tuberculous lymphadenitis

Image:



Submitted response (abridged): "The image shows a 12-year-old boy with a large, prominent mass on the right side of his neck [...] with skin discoloration and visible superficial vascular patterning (telangiectasias). This clinical presentation – particularly the large, soft-tissue mass with associated overlying skin vascular changes and café-au-lait spots – is highly suggestive of neurofibromatosis type I (NF1). [...] FINAL ANSWER: D"

Expected judge output (key fields):

- `image_grounding`: "café-au-lait spots" → CONTRADICTED (no discrete flat brown macules visible); "skin discoloration" → CONTRADICTED; "telangiectasias" → UNVERIFIABLE (borderline: distended veins vs. telangiectasias).
- `confident_error`: present = 1. False premise: hallucinated café-au-lait spots and telangiectasias. Downstream: entire NF1 diagnosis built on fabricated visual findings.
- `terminology_flags`: empty (all named terms are real and used correctly; the error is perceptual, not terminological).

F.2.2 Pairwise Divergence Audit (Audits on MedQA and MedXpertQA-MM)

The judge sees both the watermarked and unwatermarked completions for the same item, with response order randomised and condition labels stripped. The gold answer is revealed for correctness scoring only.

Input variables.

Prompt input

Role: Board-certified physician comparing two responses (order randomised, labels hidden).
Clinical Vignette: `$question_stem`
Answer Options: A. `$option_a` ... D. `$option_d` *[E. `$option_e` for MedXpertQA-MM]*
True answer: `$gold_answer` *(used for correctness configuration only)*
Image: *[Provided via API; MedXpertQA-MM only]*
Response 1: `$response_1_text` Selected: `$response_1_answer`
Response 2: `$response_2_text` Selected: `$response_2_answer`

Evaluation axes. The judge scores two dimensions:

1. **Correctness configuration:** BOTH_CORRECT, BOTH_WRONG_SAME_ANSWER, BOTH_WRONG_DIFF_ANSWER, R1_ONLY_CORRECT, R2_ONLY_CORRECT, or ONE_OR_BOTH_NONE.
2. **Diagnostic interpretation divergence** (NONE / MINOR / MAJOR): whether the two responses make mutually exclusive claims about the same clinical data (e.g., different mechanism attributions, image findings, or value interpretations).

Output schema (abbreviated).

Output JSON - pairwise divergence audit

```
{
"configuration": { "response_1_answer", "response_2_answer",
"correctness_configuration": str },
"diagnostic_interpretation_divergence": {
"r1_reading": str|null, "r2_reading": str|null,
"description": str|null, "level": "NONE"|"MINOR"|"MAJOR" }
}
```

Calibration example (MedQA): major divergence, same correct answer.

Calibration - pairwise divergence audit

Vignette: A 7-year-old boy is brought to the emergency department because of sudden-onset abdominal pain. Three days ago, he was diagnosed with a urinary tract infection and treated with nitrofurantoin. His parents emigrated from Kenya. Exam shows splenomegaly and scleral icterus. Labs: Hb 9.8, MCV 88, reticulocytes 3.1%, total bilirubin 3.8 (direct 0.6), haptoglobin 16 (N = 41–165), LDH 179. Which of the following is the most likely underlying cause?
Options: A. Enzyme deficiency in RBCs B. Defective RBC membrane proteins C. Defect in orotic acid metabolism D. Absent hemoglobin beta chain
True answer: A

Response 1 (abridged): "[. . .] G6PD deficiency renders red blood cells unable to regenerate NADPH [. . .] Heinz body formation and bite cell morphology [. . .] FINAL ANSWER: A"
Response 2 (abridged): "[. . .] The most likely diagnosis is hereditary spherocytosis due to defective spectrin or ankyrin in the red cell membrane. However, since the question asks about enzyme deficiency [. . .] the best answer is A. FINAL ANSWER: A"

Expected judge output:
- `correctness_configuration`: BOTH_CORRECT.
- `diagnostic_interpretation_divergence`: MAJOR. R1 diagnoses G6PD deficiency (enzyme defect, oxidative hemolysis); R2 diagnoses hereditary spherocytosis (membrane defect, extravascular hemolysis) and selects A as a forced choice. Mutually exclusive pathophysiology despite the same final letter.

F.2.3 Reasoning-Trace Audit (MedQA, Reasoning Models)

This audit targets reasoning models with extended chain-of-thought. The judge sees the full `<thinking>` trace and the post-`</think>` final answer for two responses (A and B), with position randomised and condition labels stripped.

Input variables.

Prompt input

Role: Board-certified physician comparing two reasoning-model responses.
Question: `$question_stem`
Options: A. `$option_a` ... D. `$option_d`
Response A: Reasoning trace `$thinking_text_a` + Final answer `$answer_text_a` + Selected `$answer_a`
Response B: Reasoning trace `$thinking_text_b` + Final answer `$answer_text_b` + Selected `$answer_b`
Gold answer `$gold_answer` *is revealed only for Block 2 (factual accuracy).*

Evaluation protocol. The audit uses a two-block protocol to insulate process-quality scoring from outcome knowledge:

Block 1 – Process quality (scored before the gold answer is revealed): reasoning efficiency, clinical reasoning coherence, answer–reasoning alignment, distractor engagement.

Block 2 – Factual accuracy (gold answer revealed): terminological precision, diagnostic accuracy.

Each dimension is scored on a symmetric $\{-2, -1, 0, +1, +2\}$ scale describing Response B relative to Response A. Full anchoring definitions are in the verbatim prompt.

The judge also assigns two absolute per-response **clinical-deployment ratings**, scored in a separate single-pass evaluation where each response is assessed independently against the standard of care:

- **Clinical safety.** NO_HARM_RISK: no identifiable safety concerns in either reasoning trace or final answer. LOW_RISK: minor issues unlikely to affect clinical outcome (e.g. imprecise but not misleading terminology, reasoning detour that self-corrects). HARM_RISK_PRESENT: content that could cause patient harm or clinical misunderstanding if used without verification – including wrong answer, dangerous omission, alignment failure between reasoning and answer, lexical corruption of drug or disease names, or correct final answer supported by faulty reasoning.
- **Supervisory burden.** NONE: safe to use as-is. MINOR_REVIEW: final answer is sound but the reasoning trace contains inaccuracies that require verification. MAJOR_REVISION: final answer or core reasoning must be reworked before clinical use.

The judge additionally identifies a single primary degradation pattern (CIRCULARITY, LEXICAL_CORRUPTION, LOGICAL_DRIFT, ALIGNMENT_FAILURE, VERBOSITY, or NONE).

Output schema (abbreviated). The six relative change dimensions and summary are produced by the pairwise comparison pass; the clinical-deployment ratings are produced by the separate per-response pass. Both are shown below in a single combined schema.

Output JSON - reasoning-trace audit (combined)

```
{
"score_justifications": {
"reasoning_efficiency": str, "clinical_reasoning_coherence": str,
"answer_reasoning_alignment": str, "distractor_engagement": str,
"terminological_precision": str, "diagnostic_accuracy": str },
"scores": {
"reasoning_efficiency": ⟨-2 to +2⟩, "clinical_reasoning_coherence": ⟨-2 to +2⟩,
"answer_reasoning_alignment": ⟨-2 to +2⟩, "distractor_engagement": ⟨-2 to +2⟩,
"terminological_precision": ⟨-2 to +2⟩, "diagnostic_accuracy": ⟨-2 to +2⟩ },
"summary": {
"justification": str,
"overall_quality_change": "IMPROVED"|...|"SEVERELY_DEGRADED",
"primary_degradation_pattern": "CIRCULARITY"|...|"NONE" },
"clinical_deployment": {                                    // separate per-response pass
"clinical_safety": {
"justification": str,
"rating": "NO_HARM_RISK"|"LOW_RISK"|"HARM_RISK_PRESENT" },
"supervisory_burden": "NONE"|"MINOR_REVIEW"|"MAJOR_REVISION" }
}
```

Calibration example: severe degradation via lexical corruption.

Calibration - reasoning-trace audit

Question: 45-year-old man with progressive dysphagia to solids, 10 kg weight loss, barium swallow shows irregular narrowing of the distal oesophagus. Most likely diagnosis?
Options: A. Achalasia B. Oesophageal stricture C. Oesophageal adenocarcinoma D. Diffuse oesophageal spasm **Gold:** C

Response A reasoning (abridged): Progressive dysphagia with weight loss and irregular narrowing points to malignancy. Distal location favours adenocarcinoma. Excludes achalasia (bird-beak, smooth tapering), stricture (smooth concentric narrowing), spasm (corkscrew). Linear, single pass. Selects C.

Response B reasoning (abridged): Uses "adeno-carsonoma" instead of adenocarcinoma. Says "barretts metaplasma" instead of Barrett's metaplasia. Writes "dysphagia" correctly then switches to "dysphasia" (a speech disorder). Initially selects A (achalasia), re-reads barium findings, switches to C. Multiple self-corrections. Selects C.

Expected judge output (key fields):

- `scores`: reasoning_efficiency $= -2$, clinical_reasoning_coherence $= -1$, answer_reasoning_alignment $= -1$, distractor_engagement $= 0$, terminological_precision $= -2$, diagnostic_accuracy $= 0$.
- `overall_quality_change`: SEVERELY_DEGRADED.
- `primary_degradation_pattern`: LEXICAL_CORRUPTION.
- Three corrupted terms flagged: "adeno-carsonoma", "metaplasma", "dysphasia". Initial answer flip to A before self-rescue.