TOP-k PROBING FOR BLACK-BOX LLM WATERMARK DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Auditing an LLM watermark is hard when the model keeps its watermark key and token scores private. Existing black-box detectors usually examine responses under a fixed decoding setting; some also need carefully designed prompts that restrict the answer. We use a different clue: what happens when the model's list of allowed tokens changes. Our top-k probe sends the same prompt with a short and a long token list, then checks whether the model changes how it chooses between tokens on both lists. To the best of our knowledge, this is the first black-box LLM watermark test to use a controlled change in top-k on the same prompt as its probe. We build two detectors around this idea. Water-Zero uses an exact conditional test on ordinary text prefixes. It needs no answer template, trained detector, watermark key, or clean-model calibration. Water-Sonar learns patterns in the responses from each token list and covers more watermark families; it can use natural or crafted prompts. For natural prompts, we test held-out keys on C4 and OpenGen with three Qwen and Llama models. A separate crafted-prompt study covers nine models and nine watermark families. Comparisons include plain-model audits, pooled exact tests, and tests that keep top-k fixed. Water-Zero detects watermarks that respond to a change in the allowed tokens, and its natural-prompt results transfer across the two datasets. Water-Sonar detects more watermark families, while its natural-prompt version remains effective for several major families. The fixed-top-k tests show that changing the token list creates the direct signal used by the exact detector.

1 INTRODUCTION

An LLM may watermark its responses while keeping the key, token scores, and sampling rule private. A user can query the model but sees only returned text. Existing black-box tests therefore ask many questions and look for watermark patterns in the answers. Some use crafted prompts to restrict the possible answers Liu et al. (2025); others test properties of ordinary text that depend on the watermark family Gloaguen et al. (2025).

We use a different clue: the requested top-k value. The auditor sends the same text with a short and a long list of allowed tokens, recording which list produced each answer. With ordinary softmax sampling, adding tokens can change each token's probability, but the odds between two tokens on both lists stay the same. Some watermarks act after top-k has limited the choices. For them, adding tokens can change those odds. This makes top-k an input to the audit rather than a setting kept fixed throughout it.

To the best of our knowledge, this is the first work to change top-k on the same prompt and test the resulting token choices for black-box LLM watermark detection. We base this comparison on the primary accounts of watermark methods and black-box detectors reviewed in Section 2. Figure 1 shows the two detectors built around it.

Water-Zero is the simpler detector. It accepts unedited text prefixes, uses an initial set of queries to choose informative token groups, and tests their change across menus with an exact conditional p-value. It needs no trained detector, clean-model calibration set, or designed answer template. The text can look like an ordinary prefix because the change is in the sampling setting. On held-out C4 prefixes and SynthID keys, Water-Zero detects 84.4–91.4% of marked audits at 2,000 responses across three base models; plain-model rejection is 3.0–4.5% at the 5% level. We then keep the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

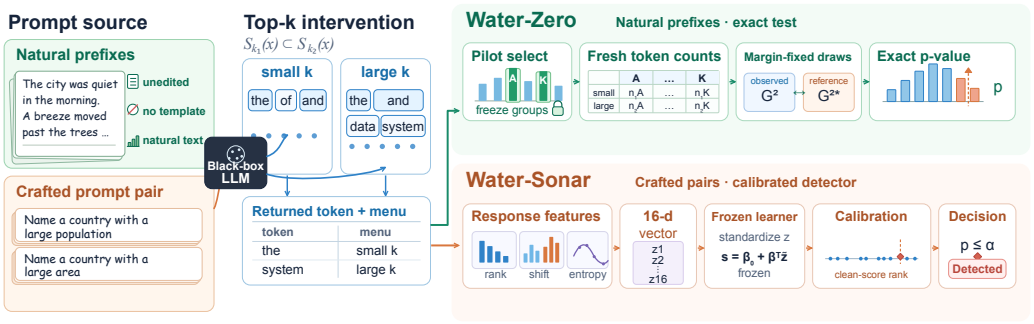

Figure 1: **One top-k intervention, two detectors.** For a fixed prompt x, the smaller allowed-token set $S_{k_1}(x)$ lies inside $S_{k_2}(x)$. Water-Zero uses unedited prefixes at top-8 and top-32. Its pilot fixes prompts and up to eight shared-token categories per prompt; fresh counts form $2 \times K$ tables for an exact conditional test. Water-Sonar uses crafted prompt pairs at top-32 and top-128. It standardizes 16 response features, applies a frozen trained score, and ranks that score against separate clean audits.

C4 settings fixed and run the same 18-word test on OpenGen. It reaches 86.1–88.2% TPR, with 5.0–8.5% observed rejection on plain models across the three checkpoints.

Water-Sonar is the learned detector. It uses crafted prompt pairs and responses collected under two top-k settings. From the token counts it computes 16 features, then compares its trained score with scores from clean audits. Its separate nine-model study covers nine watermark families. Mean TPR is 79.5% at 2,000 responses and 87.6% at 20,000; observed plain FPR is 4.7% and 5.2%. Those numbers describe the complete crafted-prompt system, including clean calibration. We also fit the same 16-feature calculation on unedited C4 and OpenGen prefixes and compare both detectors on the same nine families and three models. At 2,000 responses, this natural-prompt Sonar variant detects SynthID, Aar, DiPMark, and γ-Reweight strongly on both corpora. Matched controls show how its learned score draws on both the prompt responses and their top-k labels (Section G).

Our contributions have three parts. First, we use controlled top-k changes to probe a black-box watermark and show why the odds between shared tokens provide a signal. Second, Water-Zero tests that signal with only unedited prefixes and responses from the target model. A sweep over ten menu pairs shows that changing the allowed tokens drives its detection. Third, Water-Sonar learns from the same records. We test its crafted-prompt system across nine families, fit a separate score on natural prompts, and compare both detectors on matched C4 and OpenGen panels. Fixed-menu controls show which results depend on changing top-k and which a learned score can obtain without that change.

Together, these findings show that changing a standard decoding control can reveal watermark behavior from returned tokens alone, with an exact test for ordinary text and a learned detector for broader family coverage.

2 Background and Related Work

Watermark sampling. Red–green watermarks favor a secret set of tokens by changing their token scores Kirchenbauer et al. (2023). Aar uses keyed exponential or Gumbel sampling Aaronson (2022). DiPMark and γ-Reweight preserve the overall token distribution while changing how tokens are sampled Wu et al. (2024); Hu et al. (2024). EXP-Edit and ITS-Edit use fixed pseudorandom key lists and edit-distance detection Kuditipudi et al. (2024). KGW-Min and KGW-Skip are hashing variants distributed in MarkLLM Pan et al. (2024), while SynthID-Text uses keyed tournament sampling Dathathri et al. (2024). These methods leave different clues for an outside auditor. A fixed change to token scores preserves the odds between two shared tokens; a tournament can depend on the other tokens allowed into the choice set. Related work tests watermark reliability under text changes and studies constructions with stronger guarantees Kirchenbauer et al. (2024); Christ et al. (2024); Zhao et al. (2024). MorphMark adapts its token-score watermark during generation Wang

et al. (2025). Theoretical and empirical studies of SynthID-Text examine why its keyed sampling rule is detectable under different conditions Omidi et al. (2026). More broadly, impossibility results limit what any strong text watermark can guarantee after arbitrary text changes Zhang et al. (2024). Systematic evaluations also show that watermark performance changes with the generation and detection setting Piet et al. (2023).

Black-box watermark tests. Gloaguen et al. test several watermark families from returned text, with a separate statistical rule for each mechanism Gloaguen et al. (2025). Water-Probe designs pairs of prompts with related answer spaces and compares their response distributions Liu et al. (2025). These studies show that repeated answers can expose a watermark without the key. In the papers we reviewed, top-k sets up generation, tests robustness, or enters the watermark's sampling rule. We did not find a black-box test that changes top-k for the same text and checks the resulting token counts. Here the requested top-k value is the controlled input. Water-Zero can therefore use ordinary text rather than a crafted answer template; Water-Sonar combines crafted prompt pairs with a score that retains the top-k label of each response. Work on stealing a watermark or inferring hidden model information from returned text further motivates the limited-query setting Jovanović et al. (2024); Carlini et al. (2024); Finlayson et al. (2024). Language-model inversion can also recover a hidden prompt from next-token probabilities, which are a stronger interface than returned tokens alone Morris et al. (2024).

Top-k and conditional tests. Top-k is a standard sampling control Fan et al. (2018); top-p sampling is another common way to limit token choices Holtzman et al. (2020). Hu et al. measure how a mismatch in top-k changes a watermark detector's score Hu et al. (2024). Here the auditor asks the same prompt under more than one top-k setting and compares the answers. The rule that shared-token odds stay fixed is a known property of multinomial logit Luce (1959); McFadden (1974). Water-Zero uses this property across two token menus; choice-set effects have also been studied outside watermarking Benson et al. (2016). It then conditions on the row and column totals of token-count tables, which removes the unknown token probabilities from its exact test Agresti (2013).

3 The Top-k Audit

Take a text prefix x and request two values, $k_1 < k_2$. The service returns one sampled next token per request. The auditor saves the prompt, returned token, and requested k, but cannot see the model's token scores or watermark key. Let $S_k(x)$ be the tokens actually available at setting k, and let $q_x^k(i)$ be the chance of returning token i. In our experiment, top-k limits the tokens before the watermark is applied, so $S_{k_1}(x) \subset S_{k_2}(x)$.

Both detectors change top-k, but they ask different questions. Water-Zero asks whether each token can keep the same score $v_x(i)$ across all the requested token lists at prompt x:

$$q_x^k(i) = \frac{\exp(v_x(i))}{\sum_{h \in S_k(x)} \exp(v_x(h))}, \qquad i \in S_k(x). \tag{1}$$

The scores may change with the prompt. This rule covers ordinary softmax sampling and fixed changes to token scores at that prompt. Water-Zero gives a conditional p-value for the rule: a small value means that one set of fixed token scores has difficulty explaining the counts from both menus.

Water-Sonar instead asks whether an audit looks more like marked examples than clean ones. It learns a score from development data and sets its threshold with separate clean audits. It can use patterns that Water-Zero's fixed-score rule allows. Its decision is calibrated for the clean audits used in the study; it is not an exact test of Equation (1).

Water-Zero needs real control over the tokens available before watermark sampling, with the same first-token context in both arms. Water-Sonar also requests two menus, and the matched control in Section G tests how much its learned score needs that difference. Repeated requests at a fixed prompt and menu are independent draws from the same token distribution; requests across prompts and menus are independent too. The natural-prefix and crafted-prompt experiments have different prompts, keys, and menu pairs, so we report their rates separately.

4 WHY CHANGING TOP-k GIVES EVIDENCE

Consider two tokens i and j that remain available under both top-k settings. The key question is whether adding other tokens changes the choice between i and j.

Theorem 4.1 (Stable odds under fixed scores)**.** *If Equation* (1) *holds at a prompt x, then for any two queried menus S, T containing i, j,*

$$\frac{q_x^S(i)}{q_x^S(j)} = e^{v_x(i)-v_x(j)} = \frac{q_x^T(i)}{q_x^T(j)}. \tag{2}$$

Thus, if the response is either i or j, the chance that it is i is the same under both menus.

The proof divides two probabilities that share the same normalizing term; Section A gives every step. For example, suppose "cat" accounts for 70% of responses that are either "cat" or "dog" under top-8. With fixed token scores, it accounts for the same 70% under top-32. A fixed red–green boost at the same first-token context simply changes $v_x(i)$ and $v_x(j)$; it preserves the equality in Equation (2).

SynthID-style tournament sampling can behave differently. It starts with 2^m independent draws from the ordinary next-token distribution limited to menu S. At each of m layers, it compares disjoint pairs and keeps the token whose secret bit is preferred; a fair coin breaks equal-bit ties. Let $p^{(0),S}$ be the starting distribution and $g_{\ell i} \in \{0, 1\}$ be token i's secret bit at layer ℓ. The key and prompt are held fixed across menus.

Theorem 4.2 (Tournament response law)**.** *For a tournament of m layers, define $s_\ell(S) = \sum_{h\in S} p_h^{(\ell-1),S} g_{\ell h}$. The distribution after layer ℓ is*

$$p_i^{(\ell),S} = p_i^{(\ell-1),S}\big(1 + g_{\ell i} - s_\ell(S)\big), \qquad \ell = 1, \ldots, m. \tag{3}$$

Thus the returned-token distribution is $q_i^S = p_i^{(0),S} \prod_{\ell=1}^m (1 + g_{\ell i} - s_\ell(S))$.

Here $s_\ell(S)$ is the total probability of tokens with the preferred bit. Adding tokens can change that total. To see what this does to two tokens shared by both menus, divide the displayed response probability for i by the one for j under each menu. The base ratio cancels by Equation (2). At a layer where $g_{\ell i} = 1$ and $g_{\ell j} = 0$, the remaining ratio is $(2 - s_\ell(S))/(1 - s_\ell(S))$; when their bits are reversed it is the reciprocal, and equal bits contribute one. Thus changing the menu can change their final odds. Section A derives the update and the ratio step by step. Water-Zero looks for the change in returned-token counts.

5 WATER-ZERO AND WATER-SONAR

Both detectors change top-k and collect one next token per request. Each record contains the prompt, requested top-k, and returned token. Water-Zero uses these records in an exact conditional test; Water-Sonar uses them to compute a learned score. Only Water-Sonar needs training and clean calibration data.

5.1 WATER-ZERO: AN EXACT TEST WITH NATURAL PROMPTS

Water-Zero sends unedited text prefixes. It queries each under top-8 and top-32, spending half its budget on an initial set of 32 uniformly sampled prompts. At each prompt, it considers tokens seen under both menus. It ranks them by their smaller count across the two arms, then by total count and token ID, and keeps up to eight. It scores the resulting $2 \times K$ table with Equation (4). It then ranks prompts by G^2, category count, and prompt ID, selecting the first two. These tie rules fix the selection before fresh confirmation responses are collected. As specified in Section 3, the new responses are independent of the initial set and of each other, and every request counts toward the budget.

For selected group h, let C_{hri} count confirmation responses in token category i under menu $r \in \{1, 2\}$. Only the scoring step discards responses outside the categories chosen in the initial set. Let C_{hr+}, C_{h+i}, and C_{h++} be the row, column, and total counts. The group's deviance is

$$G_h^2 = 2\sum_{r=1}^{2}\sum_{i} C_{hri} \log\left(\frac{C_{hri}C_{h++}}{C_{hr+}C_{h+i}}\right), \qquad 0 \log 0 := 0. \tag{4}$$

For H retained groups, we consider the fixed set $\mathcal{Q} = \{1, 2, 4, \ldots\} \cap [H] \cup \{H\}$. For each q in this set, the test adds the q largest group deviances and ranks the sum against conditional table draws. It then finds the strongest rank and ranks that choice again using the same draws. In this way, the reference distribution includes the choice of q. If no group is usable, Water-Zero returns $p = 1$.

Proposition 5.1 (Exact level of selected Water-Zero)**.** *Suppose the target follows Equation* (1) *at every queried prompt, and the pilot and confirmation responses are independent. Conditional on the pilot, menus, selected token groups, and confirmation row and column totals, the Water-Zero confirmation tables follow a product of multivariate-hypergeometric laws. Its weak-rank p-value satisfies* $\Pr(p \le \alpha) \le \alpha$ *for every* $\alpha \in [0, 1]$.

Section A derives the table probabilities, accounts for discarded tokens, and proves that choosing q from the data preserves the test level.

We also compare with the earlier pooled Water-Zero rule. It first sends top-2 queries to every prompt and keeps a prompt when both available tokens appear. It spends the remaining budget on top-8 queries at those prompts, then compares the counts of the two shared tokens across menus. Its exact reference distribution includes both the prompt-selection step and the redistribution of requests.

Proposition 5.2 (Exact level of pooled Water-Zero)**.** *Under Equation* (1) *and independent requests as specified in Section 3, the pooled rule's conditional p-value satisfies* $\Pr(p \le \alpha) \le \alpha$ *for every budget, including random admission and reallocation.*

The full conditioning proof appears in Section A. Both Water-Zero rules use only target responses and need no clean-model calibration set.

5.2 Water-Sonar: a learned decision with paired prompts

Water-Sonar uses two crafted prompt families with related answer spaces. It asks each prompt under top-32 and top-128, assigning one quarter of responses to the shorter token list. For each of 50 fixed prefixes z, the two templates have the same instruction wrapper but ask either "Name a country with a large population" or "Name a country with a large area." Section C summarizes the grid; the packaged protocol gives the literal prompt text and prefix order.

For each prompt, Water-Sonar counts the returned tokens separately under the two top-k settings. Its features measure how the token counts, diversity, and entropy change between them. They also include a rank-similarity score computed after pooling responses. The resulting 16 numbers go into a ridge-logistic model trained on development audits, yielding one score per target. Section D gives the exact feature order and preprocessing; the supplement includes the fitted weights.

Water-Sonar compares a target score s with n scores $c_1, \ldots, c_n$ from separate clean audits. Its calibrated upper-tail rank is

$$p_{\mathrm{cal}}(s) = \frac{1 + \sum_{j=1}^{n} \mathbf{1}[c_j \ge s]}{n + 1}. \tag{5}$$

If the target and clean calibration audits come from the same clean distribution, the target's rank among their $n + 1$ scores is equally likely to take any position. Counting ties against rejection keeps the test conservative. This gives a level-α decision for that clean audit distribution. The final family study uses 1,000 separate clean calibration audits per model and budget. If one calibration set serves M targets, the average response cost is $B(1 + 1000/M)$ per target, where B counts target responses.

The same 16-feature calculation can use two groups of unedited natural prefixes. In the matched study, Sonar-N groups the even- and odd-indexed C4 and OpenGen prefixes, without adding instructions or words. It trains a new score on natural-prompt development audits and sets the threshold with separate clean audits. This lets us study a different type of prompt while keeping the feature calculation fixed.

6 Experimental Setup

We test the detectors in their separate held-out studies and in a matched natural-prompt study. The experiments draw independent one-token responses from next-token distributions computed with real, frozen model checkpoints. The Qwen2.5, Llama, and Mistral checkpoints follow their released

model families Yang et al. (2024); Llama Team (2024); Jiang et al. (2023). The detector sees only the returned token and requested top-k value. Only the experiment generator sees the token scores and watermark keys. Every response, including one used by Water-Zero for initial selection, counts toward the stated budget.

Natural-prefix Water-Zero study. We choose the first 96 distinct 18-word prefixes from the cached C4 `realnewslike` validation stream before looking at token scores or watermark results. Prefixes 0–47 and SynthID keys 0–7 are used to choose the method on Qwen2.5-1.5B. The final study uses prefixes 48–95 and keys 24–35 on Qwen2.5-1.5B, Qwen2.5-3B, and Llama-3.2-3B base models. We carry the selected top-(8,32) setting to the other two models unchanged. At each budget, $B = 2{,}000$ or $5{,}000$, every model–rule cell has 1,200 marked audits, 200 plain audits, and 100 fixed-bias KGW controls. The two Water-Zero rules use independent response streams on the same final prompts and keys. Each final exact test uses 999 conditional draws and the decision rule $p \leq 0.05$.

We next apply this fixed test to the released OpenGen prefixes of Krishna et al. Krishna et al. (2023). In release order, we take the first 96 distinct prefixes that meet the C4 length rule, without a model-based prompt screen. The main test cuts each prefix to 18 words, matching the C4 prompt length. We evaluate positions 48–95 and do not use the earlier positions to adjust the detector. A second test uses the authors' two-sentence generation prompts; a third uses each complete source context to check longer prefixes. All three keep the C4-selected rule, pooled comparison, three checkpoints, 12 final keys, two budgets, audit counts, and $p \leq 0.05$ decision. Their response streams are independent. Together, these tests separate the effect of text source from that of prompt length.

Matched natural-prompt family study. For the natural-prompt curves in Figure 2, we test both detectors on the same 18-word C4 and OpenGen panels, three checkpoints, nine watermark families, held-out keys 24–35, and budgets of 2,000 and 5,000 responses. Water-Zero keeps the C4-selected exact $8/32$ test. Sonar-N uses the same 16 features with a $32/128$ menu pair. The even- and odd-indexed prefixes form two groups of 24, with their wording unchanged. For each budget, we fit one ridge-logistic score using the separate first 48 prefixes of both corpora, keys 0–7, all three models, and all nine families; the ridge value is fixed at 10. Every final family cell has 20 marked audits per key. For each corpus, model, and budget, Sonar-N uses 1,000 separate plain audits for calibration and 1,000 for evaluation; Water-Zero uses 200 plain evaluations. The detectors draw independent responses. Sonar-N has its own fitted score and never uses the one trained on crafted prompts.

Crafted-prompt Water-Sonar study. We use two prompt families with 50 prefixes each, nine base-model checkpoints, and nine watermark families: SynthID Dathathri et al. (2024), KGW Kirchenbauer et al. (2023), Aar Aaronson (2022), KGW-Min and KGW-Skip Pan et al. (2024), DiPMark Wu et al. (2024), γ-Reweight Hu et al. (2024), and EXP-Edit and ITS-Edit Kuditipudi et al. (2024). The support pair is (32,128). Water-Sonar is trained on three development models and keys 48–59, checked on separate models and keys 60–71, and evaluated on nine models with untouched keys 72–83. All final watermark keys and response streams are held out. Some model checkpoints also appear during training or validation: Qwen2.5-1.5B, Llama-3.2-3B, and Mistral-7B appear in development; OPT-2.7B, Qwen2.5-3B, and Llama2-7B appear in validation; Qwen2.5-7B, Llama-3.1-8B, and Llama2-13B first appear in final evaluation. Thus every final result uses new keys, while only the last three models are also new checkpoints. Each model–family–budget cell contains 240 marked audits; each model–budget has 1,000 separate plain evaluations and 1,000 clean calibration audits. Budgets are $B \in \{2{,}000, 5{,}000, 10{,}000, 20{,}000\}$ and the decision level is 0.05. The pooled exact Water-Zero rule is included as a same-grid reference. We report these crafted-prompt results separately from the natural-prefix Water-Zero study.

Measures. True-positive rate (TPR) is the share of marked audits detected at the fixed threshold. False-positive rate (FPR) is the share of independent plain audits incorrectly detected. The natural study also reports area under the ROC curve (AUC), which measures score ranking across thresholds. Because the marked and plain groups have different sizes, Table 2 gives them equal weight in F1, yielding $2\,\mathrm{TPR}/(1 + \mathrm{TPR} + \mathrm{FPR})$. We also show the weakest held-out key and fixed-bias control rejection. In the crafted study, each model–family cell has equal weight in the average TPR; the tables keep these different averages separate.

Table 1: **Water-Zero on natural C4 and OpenGen text.** Both datasets use 18-word prefixes and the test chosen on C4. Each row has 12 held-out SynthID keys, 1,200 marked audits, 200 plain audits, and 100 fixed-bias KGW controls at $p \leq 0.05$. Selected and pooled Water-Zero draw independent responses. Lowest-key TPR and both control rates use the selected rule. TPR intervals resample keys; plain-FPR intervals use exact binomial limits. All rates are fractions.

		SynthID marked audits				Key spread	Controls	
Model	B	Selected TPR		Pooled TPR		Lowest-key	Plain FPR	KGW
		Rate	95% CI	Rate	95% CI	TPR	[95% CI]	Rejection
C4 realnewslike: matched 18-word natural prefixes								
Qwen 2.5-1.5B	2k	**0.914**	[0.858, 0.956]	0.818	[0.667, 0.936]	0.660	0.030 [0.011, 0.064]	0.040
	5k	**0.991**	[0.979, 0.999]	0.937	[0.833, 0.997]	0.940	0.065 [0.035, 0.109]	0.050
Qwen 2.5-3B	2k	**0.883**	[0.809, 0.949]	0.636	[0.402, 0.856]	0.590	0.030 [0.011, 0.064]	0.060
	5k	**0.984**	[0.974, 0.993]	0.765	[0.558, 0.941]	0.960	0.030 [0.011, 0.064]	0.050
Llama 3.2-3B	2k	**0.844**	[0.756, 0.921]	0.633	[0.399, 0.848]	0.550	0.045 [0.021, 0.084]	0.040
	5k	**0.953**	[0.910, 0.987]	0.781	[0.605, 0.938]	0.760	0.030 [0.011, 0.064]	0.050
OpenGen: matched 18-word natural prefixes								
Qwen 2.5-1.5B	2k	**0.875**	[0.814, 0.930]	**0.875**	[0.780, 0.959]	0.710	0.055 [0.028, 0.096]	0.060
	5k	**0.993**	[0.989, 0.998]	**0.993**	[0.985, 0.998]	0.980	0.045 [0.021, 0.084]	0.020
Qwen 2.5-3B	2k	**0.882**	[0.834, 0.922]	0.648	[0.432, 0.844]	0.720	0.050 [0.024, 0.090]	0.060
	5k	**0.983**	[0.971, 0.993]	0.822	[0.657, 0.956]	0.930	0.045 [0.021, 0.084]	0.060
Llama 3.2-3B	2k	**0.861**	[0.813, 0.904]	0.761	[0.589, 0.909]	0.700	0.085 [0.050, 0.133]	0.060
	5k	**0.981**	[0.957, 0.995]	0.951	[0.899, 1.000]	0.860	0.040 [0.017, 0.077]	0.060

7 RESULTS

7.1 WATER-ZERO: STRONG SYNTHID DETECTION FROM UNEDITED PREFIXES

Table 1 shows how often Water-Zero detects SynthID from unedited prefixes. On C4, it detects 84.4–91.4% of marked audits across the three models with 2,000 responses, rising to 95.3–99.1% with 5,000. At 2,000 responses, its observed false-positive rate on plain models is 3.0–4.5%. It also rejects 4.0–6.0% of the fixed-bias KGW controls, close to the intended 5% rate in these 100-audit cells.

We chose the rule on C4, then applied it to OpenGen text of the same length without changing it. Its TPR is 0.861–0.882 at 2,000 responses and 0.981–0.993 at 5,000, compared with 0.844–0.914 and 0.953–0.991 on C4. OpenGen plain FPR is 0.050–0.085 at 2,000 responses and 0.040–0.045 at 5,000. Water-Zero therefore works on unedited prefixes without clean calibration, and the same test settings remain effective across the two corpora.

7.2 WATER-SONAR: NATURAL-PROMPT COVERAGE ACROSS FAMILIES

Table 2 compares the nine families on matched C4 and OpenGen prefixes. At 2,000 responses, Sonar-N detects a mean 0.529 and 0.550 of marked audits on the two corpora. Their AUCs are 0.762 and 0.777, and their balanced F1 scores are 0.667 and 0.684. All three measures rise at 5,000 responses, while observed plain FPR remains between 0.043 and 0.059. SynthID and Aar reach perfect mean detection in every setting; DiPMark and γ-Reweight are also strong. KGW, its two variants, and the fixed-key-list families are harder. This matched natural-prompt study uses 2,000 and 5,000 responses.

Figure 2 lets readers compare the two natural-prompt detectors by family and model, with the crafted-prompt results shown for context. The crafted curves repeat across rows because that study used neither corpus. At 2,000 responses, Sonar-N's mean TPR over nine families and three models is 0.529 on C4 and 0.550 on OpenGen, compared with 0.171 and 0.174 for Water-Zero. At 5,000 responses, the corresponding rates are 0.543 and 0.574 for Sonar-N, and 0.251 and 0.264 for Water-Zero. Sonar-N detects SynthID and Aar at 1.000 in every corpus–budget average, with DiPMark

Table 2: **Sonar-N on natural C4 and OpenGen text.** Each row is a watermark family. TPR and AUC average three models. FPR averages three separate plain-model tests and is the same across families within each dataset and budget. F1 gives marked and plain audits equal weight: $2\,\mathrm{TPR}/(1+\mathrm{TPR}+\mathrm{FPR})$. Each family–model cell has 12 held-out keys and 240 marked audits. Each model uses 1,000 separate plain audits for calibration and 1,000 for evaluation. The decision level is $\alpha = 0.05$.

	C4								OpenGen							
	2,000 responses				5,000 responses				2,000 responses				5,000 responses			
Watermark	TPR	FPR	AUC	F1	TPR	FPR	AUC	F1	TPR	FPR	AUC	F1	TPR	FPR	AUC	F1
Support-dependent watermark																
SynthID	1.000	0.057	1.000	0.972	1.000	0.043	1.000	0.979	1.000	0.059	1.000	0.972	1.000	0.046	1.000	0.977
N-gram or context-keyed watermark																
KGW	0.263	0.057	0.694	0.398	0.249	0.043	0.685	0.385	0.271	0.059	0.707	0.407	0.328	0.046	0.740	0.477
Aar	1.000	0.057	1.000	0.972	1.000	0.043	1.000	0.979	1.000	0.059	1.000	0.972	1.000	0.046	1.000	0.977
DiPMark	0.951	0.057	0.987	0.947	0.982	0.043	0.996	0.970	0.994	0.059	0.999	0.969	1.000	0.046	1.000	0.977
γ-Reweight	0.963	0.057	0.991	0.953	0.993	0.043	0.998	0.975	0.996	0.059	0.999	0.969	1.000	0.046	1.000	0.977
KGW-Min	0.243	0.057	0.662	0.374	0.215	0.043	0.667	0.342	0.333	0.059	0.718	0.479	0.368	0.046	0.753	0.520
KGW-Skip	0.229	0.057	0.627	0.356	0.261	0.043	0.634	0.400	0.243	0.059	0.687	0.373	0.264	0.046	0.722	0.403
Fixed-key-list watermark																
EXP-Edit	0.011	0.057	0.327	0.021	0.028	0.043	0.373	0.052	0.012	0.059	0.309	0.023	0.019	0.046	0.365	0.036
ITS-Edit	0.103	0.057	0.567	0.177	0.156	0.043	0.664	0.259	0.097	0.059	0.575	0.168	0.188	0.046	0.683	0.304
Overall summary																
Mean	0.529	0.057	0.762	0.667	0.543	0.043	0.780	0.684	0.550	0.059	0.777	0.684	0.574	0.046	0.807	0.709

and γ-Reweight at 0.951–1.000. KGW reaches 0.249–0.328; EXP-Edit and ITS-Edit stay below 0.2. Sonar-N's mean plain FPR is 0.043–0.059 across the four corpus–budget groups. The separate crafted-prompt system covers more families, especially KGW, but its trained score and tested audits differ from Sonar-N. The plotted gaps therefore cannot isolate the effect of prompt wording.

8 Discussion

What changing top-k adds. We change the tokens the model may choose while keeping the text fixed, so the two sets of answers can be compared directly. Water-Zero tests whether the odds between shared tokens change. On natural prefixes, it detects the tested SynthID tournament without hand-written answer templates. Water-Sonar keeps the top-k labels in a learned, crafted-prompt system that covers more families in its separate study. With unedited prefixes, Sonar-N detects SynthID, Aar, DiPMark, and γ-Reweight strongly, but its coverage varies across families. Water-Zero directly tests the effect of changing top-k, while Water-Sonar combines the response patterns with a learned score.

When to use each detector. With top-k control, Water-Zero collects one-token responses, chooses groups from initial queries, and computes a conditional p-value. It needs neither training nor clean calibration. Water-Sonar adds a trained score and clean calibration when broader family coverage matters. The two decisions answer different questions: Water-Zero tests whether token choices depend on the menu, while Water-Sonar tests whether responses resemble marked examples. The natural-prompt results show how these two approaches complement each other: Water-Zero gives a direct test for SynthID, while Sonar-N extends strong detection to Aar, DiPMark, and γ-Reweight.

What the experiments establish. The studies use next-token distributions from real models, frozen before audit responses are drawn. Water-Zero and Sonar-N use natural C4 and OpenGen prefixes; the original Water-Sonar study uses crafted prompt pairs. Water-Zero sends 30 initial requests per prompt at 2,000 responses and 78 at 5,000. Its results show that a crafted answer template is unnecessary. Whether ordinary prompts also pass a particular provider's prompt screen requires a live-endpoint test with enforced top-k settings and a stated screening rule. The matched Sonar-N study shows how its learned family coverage changes on unedited prefixes, with a score fitted separately from the crafted-prompt Water-Sonar model.

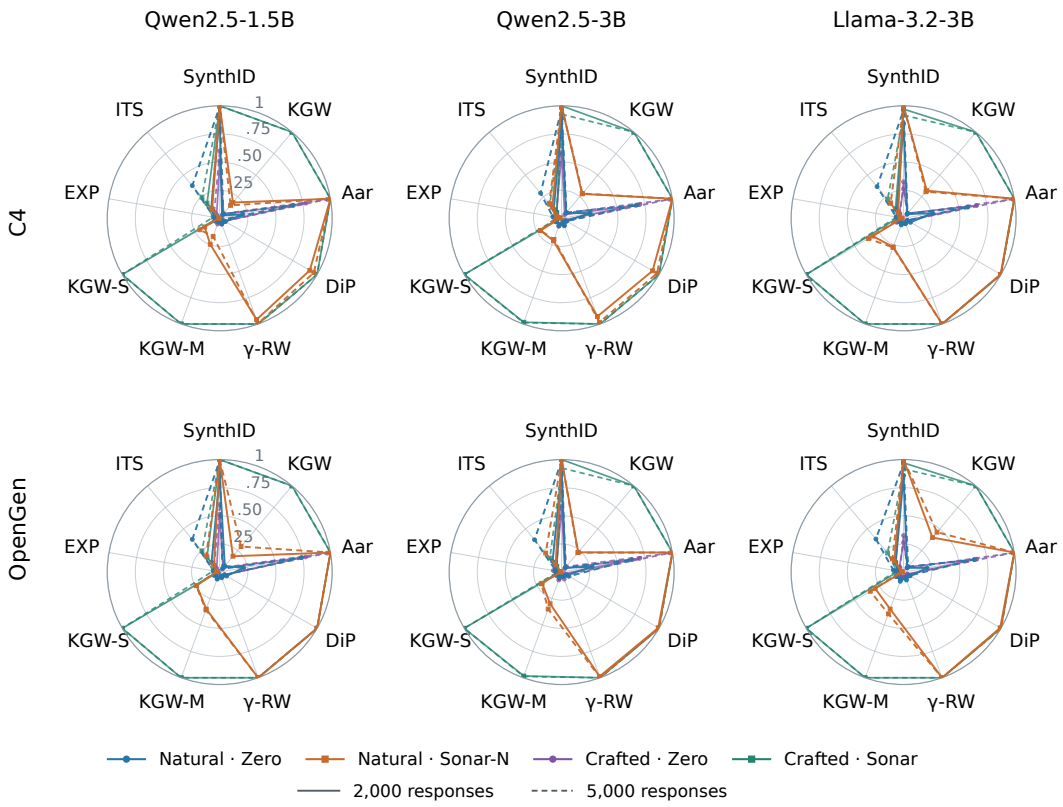

Figure 2: **Detection across watermark families.** Rows use natural 18-word prefixes from C4 and OpenGen; columns show the three base models. Each spoke is a watermark family, with TPR at the fixed 0.05 decision shown from 0 to 1. Blue and orange are natural Water-Zero and Sonar-N; purple and green are Water-Zero and Water-Sonar from the separate crafted-prompt study. Solid lines use 2,000 responses and dashed lines use 5,000. The crafted study has no C4 or OpenGen split, so its measured curves appear in both rows as a reference. Each family cell includes 12 held-out keys and 240 marked audits. The studies differ in prompts, keys, trained Sonar scores, calibration audits, and Zero menu settings; the gaps between curves do not by themselves measure the effect of prompt wording. DiP, γ-RW, KGW-M/S, EXP, and ITS mean DiPMark, γ-Reweight, KGW-Min/Skip, EXP-Edit, and ITS-Edit.

9 Conclusion

This work turns top-k from a generation setting into a tool for black-box watermark detection. By keeping the text fixed and changing the tokens the model may choose, we expose patterns that help reveal a watermark without its key. Water-Zero uses this idea to test ordinary text without a trained detector or clean calibration, while Water-Sonar learns from the returned tokens and their top-k labels to cover more watermark families. Experiments on C4 and OpenGen show that unedited text can support effective detection across models, making natural prompts a practical starting point for watermark audits.

The top-k finding gives future detectors a simple way to gather evidence through the model's sampling controls. It also connects watermark design to a concrete question: how do token choices change when the allowed list changes? This opens useful directions for choosing top-k settings adaptively, reducing the number of responses, and studying other sampling controls. Together, our results broaden the tools available for black-box audits and establish top-k probing as a useful approach to understanding watermarked generation.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX formatting, figure and table design, and code generation for manuscript checks. Reported measurements come from archived experiment outputs. AI-assisted code and displayed values were checked with validation scripts and by inspecting the compiled paper. The authors take responsibility for the final text, claims, code, and artifacts.

REPRODUCIBILITY STATEMENT

Sections 6 and 7 describe the response budgets, data splits, and decisions used in each study. The appendices give full proofs, results for each model, and the detector definitions. The supplement contains the frozen code and protocols, fitted Water-Sonar weights, result records, hashes, and the source of each table and figure. It also contains all 60 matched natural-family data shards and the new Sonar-N weights. A packaged check recomputes the 108 family cells, 12 plain controls, and all 646,800 paired-prefix curve rows.

REFERENCES

Scott Aaronson. My AI safety lecture for UT effective altruism. Shtetl-Optimized blog post, 2022. URL `https://scottaaronson.blog/?p=6823`. Describes the Gumbel/exponential watermarking construction.

Alan Agresti. *Categorical Data Analysis*. Wiley, 3rd edition, 2013.

Austin R. Benson, Ravi Kumar, and Andrew Tomkins. On the relevance of irrelevant alternatives. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pp. 963–973, 2016. doi: 10.1145/2872427.2883025.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5680–5705. PMLR, 2024.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *Proceedings of the 37th Annual Conference on Learning Theory (COLT)*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1125–1139. PMLR, 2024.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, A. Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. Scalable watermarking for identifying large language model outputs. *Nature*, 634 (8035):818–823, 2024. doi: 10.1038/s41586-024-08025-4.

Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 889–898, 2018.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. Logits of API-protected LLMs leak proprietary information, 2024.

Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin T. Vechev. Black-box detection of language model watermarks. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=uWVC5FVidc.

Albert Q. Jiang et al. Mistral 7B, 2023.

Nikola Jovanović, Robin Staab, and Martin T. Vechev. Watermark stealing in large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22570–22593. PMLR, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/575c450013d0e99e4b0ecf82bd1afaa4-Paper-Conference.pdf.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=FhZi7r4nzA.

Aiwei Liu, Sheng Guan, Yiming Liu, Leyi Pan, Yifei Zhang, Liancheng Fang, Lijie Wen, Philip S. Yu, and Xuming Hu. Can watermarked LLMs be identified by users via crafted prompts? In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Llama Team. The Llama 3 herd of models, 2024.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, New York, 1974.

John X. Morris, Wenting Zhao, Justin T. Chiu, Vitaly Shmatikov, and Alexander M. Rush. Language model inversion. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Romina Omidi, Yun Dong, and Binghui Wang. On Google's SynthID-Text LLM watermarking system: Theoretical analysis and empirical validation, 2026.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. MarkLLM: An open-source toolkit for LLM watermarking. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 61–71. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-demo.7/.

Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David A. Wagner. Mark my words: Analyzing and evaluating language model watermarks, 2023.

Zongqi Wang, Tianle Gu, Baoyuan Wu, and Yujiu Yang. MorphMark: Flexible adaptive watermarking for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4842–4860. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.240/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53443–53470. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24h.html.

An Yang et al. Qwen2.5 technical report, 2024.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58851–58880. PMLR, 2024.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Appendix contents.

A Complete proofs
B Algorithms and example
C Implementation and result records
D Sonar features and score
E Detailed result views
F Held-out result matrices
G Top-k ablations

A Complete Derivations and Proofs

A.1 Stable odds under fixed scores

We prove Theorem 4.1. Fix the prompt and let $Z_S = \sum_{h \in S} e^{v_x(h)}$ and $Z_T = \sum_{h \in T} e^{v_x(h)}$. Since i and j occur in both menus,

$$\frac{q_x^S(i)}{q_x^S(j)} = \frac{e^{v_x(i)}/Z_S}{e^{v_x(j)}/Z_S} = e^{v_x(i) - v_x(j)}.$$

Replacing S by T yields the same value. Conditional on the returned token being in $\{i, j\}$, its probability of being i is

$$\frac{q_x^S(i)}{q_x^S(i) + q_x^S(j)} = \frac{e^{v_x(i)}}{e^{v_x(i)} + e^{v_x(j)}},$$

This expression has no menu term. A red–green bias $b_x(i)$ fixed at the same prompt replaces $v_x(i)$ by $v_x(i) + b_x(i)$ under every queried menu, so the same terms cancel. The auditor needs the two requested settings, not a list of every token in the actual menus.

A.2 Tournament response law and cross-menu odds

We prove Theorem 4.2 by induction, one layer at a time. At depth zero, the winner is one draw from $p^{(0),S}$. Assume that each tournament branch of depth $\ell - 1$ returns an independent draw from $p^{(\ell-1),S}$. The two branches at depth ℓ start from disjoint draws, so their winners A and B are independent and have that same distribution. Put $p_i = p_i^{(\ell-1),S}$ and $s = \sum_h p_h g_{\ell h}$.

If $g_{\ell i} = 1$, a token i in the first slot wins against a non-preferred opponent with probability one and against a preferred opponent with probability $1/2$. Its conditional chance of winning is $(1 - s) + s/2 = 1 - s/2$. The token can win from either slot. These two ways of winning do not overlap, even when both candidates are i, because the tie coin chooses one slot. Its total winner probability is therefore $2p_i(1 - s/2) = p_i(2 - s)$. If $g_{\ell i} = 0$, it wins only against a non-preferred opponent and then only with probability $1/2$. The corresponding probability is $2p_i(1 - s)/2 = p_i(1 - s)$. Both cases equal $p_i(1 + g_{\ell i} - s)$, which proves Equation (3) at layer ℓ. The update sums to one because $\sum_i p_i(1 + g_{\ell i} - s) = 1 + s - s = 1$. Induction from layer one to m gives the stated product law.

We now compare the same token pair across menus, as used in the main text. For $i, j \in S \cap T$, divide the product law for i by the law for j:

$$\frac{q_i^S}{q_j^S} = \frac{p_i^{(0),S}}{p_j^{(0),S}} \prod_{\ell=1}^{m} \frac{1 + g_{\ell i} - s_\ell(S)}{1 + g_{\ell j} - s_\ell(S)}.$$

The first ratio is $e^{v_x(i) - v_x(j)}$ for the base softmax and is the same under T by Theorem 4.1. If $g_{\ell i} = g_{\ell j}$, the layer factor is one. If $(g_{\ell i}, g_{\ell j}) = (1, 0)$, it is $(2 - s_\ell(S))/(1 - s_\ell(S))$; if the bits are $(0, 1)$, it is the reciprocal. For a pair with positive final probabilities, every denominator used for an unequal-bit layer is positive: a token with $g = 0$ retains positive mass only when $1 - s_\ell(S) > 0$. Hence each ratio is well defined. Dividing the expression for S by the one for T removes the base ratio and leaves only factors that can depend on the menu. This is why we compare shared-token counts across menus.

A.3 Exact level of selected Water-Zero

We prove Proposition 5.1. First fix everything observed in the initial queries. This fixes the prompts, both menus at each selected prompt, the category set A_h, the number of groups, and the number

of fresh queries in each arm. The initial responses may determine all those choices because the confirmation responses are independent.

Fix one selected group h and let $r = 1, 2$ denote its menus. Under Equation (1), a category token $i \in A_h$ has probability $e^{v_h(i)}/Z_{hr}$ under arm r, where Z_{hr} sums over that arm's full menu. The probability that a response falls inside A_h is $\rho_{hr} = \sum_{i \in A_h} e^{v_h(i)}/Z_{hr}$. Thus, conditional on a hit inside A_h, the category probability is

$$\theta_{hi} = \frac{e^{v_h(i)}/Z_{hr}}{\rho_{hr}} = \frac{e^{v_h(i)}}{\sum_{j \in A_h} e^{v_h(j)}}.$$

It does not depend on r. Conditional on the numbers n_{h1} and n_{h2} of category hits, the two retained count rows are independent multinomials with sizes n_{h1} and n_{h2} and common probability vector θ_h. Conditioning also removes the arm-specific chances ρ_{hr} of hitting A_h. We therefore need no probability model for responses outside A_h.

Let $m_{hi} = C_{h1i} + C_{h2i}$, let $N_h = n_{h1} + n_{h2} = \sum_i m_{hi}$, and write $c_i = C_{h1i}$. The joint multinomial probability of the two count rows, given their sizes, is

$$\frac{n_{h1}! n_{h2}!}{\prod_i c_i!(m_{hi} - c_i)!} \prod_i \theta_{hi}^{m_{hi}}.$$

The column totals m_h themselves have a multinomial law with size N_h and probability θ_h, of mass $N_h! \prod_i \theta_{hi}^{m_{hi}} / \prod_i m_{hi}!$. Dividing these expressions shows that, conditional on both row totals and all column totals,

$$\Pr(C_{h1} = c \mid n_{h1}, n_{h2}, m_h) = \frac{\prod_i \binom{m_{hi}}{c_i}}{\binom{N_h}{n_{h1}}}, \qquad \sum_i c_i = n_{h1}, \quad 0 \le c_i \le m_{hi}. \tag{6}$$

This is the multivariate-hypergeometric law. Independent confirmation query streams give the product of these laws across selected groups after conditioning on each group's margins.

We also show why choosing the strongest statistic from the data keeps the same error level. Take M table configurations: one observed and $M - 1$ independent configurations drawn from Equation (6); final evaluation uses $M = 1000$. Conditional on the pilot and margins, all M configurations are identically distributed and exchangeable. For every configuration a, compute all group deviances, all fixed-grid top-q sums T_{aq}, and their weak upper ranks $R_{aq} = M^{-1} \sum_b \mathbf{1}[T_{bq} \ge T_{aq}]$. Set $Z_a = \max_q[-\log R_{aq}]$ and return $p = M^{-1} \sum_b \mathbf{1}[Z_b \ge Z_0]$ for the observed index zero. Reordering the M configurations only reorders their deviances, ranks, and Z values. Thus the Z values remain exchangeable even after choosing q. Conditional on their unordered multiset, the observed index is uniform over its M positions. At most $\lfloor \alpha M \rfloor$ positions have weak upper rank at most α; ties can only reduce this count. Therefore the conditional rejection probability is at most $\lfloor \alpha M \rfloor / M \le \alpha$. Average this bound over the margins and pilot. If the pilot selected no usable group, the rule returns $p = 1$ and does not reject.

A.4 Exact level of pooled Water-Zero

We prove Proposition 5.2. The first stage gives each prompt c top-2 requests. At a prompt whose two available token identities are ordered canonically as i, j, let X be the number of i responses. The prompt is admitted exactly when $1 \le X \le c - 1$. Condition on the complete admission vector and pair identities over all prompts. This fixes how many prompts are admitted and the number of top-8 requests assigned to each. It does not fix X for an admitted prompt.

For one admitted prompt, the stable-odds rule gives a common conditional probability θ of token i among $\{i, j\}$ in both arms. At top-2, every response is in that pair, so X has binomial mass $\binom{c}{x}\theta^x(1-\theta)^{c-x}$ before conditioning on admission. At top-8, condition on n_8, the number of responses that hit the pair. Fresh top-8 responses give $Y \sim \mathrm{Bin}(n_8, \theta)$ independently of the top-2 counts conditional on the admission decision. Let $A = X + Y$. The joint mass for a feasible $X = x$

Algorithm 1 Selected Water-Zero on natural prefixes

Require: Frozen natural-prefix panel, budget B, menus $(8, 32)$, level α

Uniformly choose 32 prompts from the available panel using the audit seed.
Send $n_p = \lfloor \lfloor B/2 \rfloor / (32 \cdot 2) \rfloor$ independent requests to each prompt–menu cell.
for each pilot prompt **do**
    Keep up to eight tokens returned in both arms, ordered by minimum arm count, total count, then token ID.
    If the category table is usable, score it by the deviance in Equation (4).
end for
Order usable prompts by deviance, category count, then prompt ID; lock the first $H \leq 2$.
if $H = 0$ **then**
    return $p = 1$
end if
Send $n_c = \lfloor (B - 64n_p)/(2H) \rfloor$ *fresh* requests to each selected prompt–menu cell.
Form each $2 \times K$ confirmation table on its pilot-locked token categories.
Draw 999 independent tables per group from its conditional multivariate-hypergeometric law in Equation (6); include the observed configuration as configuration zero.
On every configuration, compute all group deviances and the fixed-grid top-q sums for $q \in \{1, 2, 4, \ldots\} \cap [H] \cup \{H\}$.
Compute a weak upper rank for each q, maximize its negative log over q, and rank that maximum against the 999 drawn maxima.
return the final weak-rank p and decision $\mathbf{1}[p \leq \alpha]$.

is proportional to

$$\binom{c}{x}\binom{n_8}{A-x}\theta^A(1-\theta)^{c+n_8-A}.$$

Conditioning on A cancels the factor involving θ. Conditioning on admission restricts x to $1 \leq x \leq c-1$. Thus the exact fiber is

$$\Pr(X = x \mid c, n_8, A, \text{admitted}) = \frac{\binom{c}{x}\binom{n_8}{A-x}}{\sum_{z \in J}\binom{c}{z}\binom{n_8}{A-z}}, \quad x \in J, \tag{7}$$

where $J = \{z : 1 \leq z \leq c-1,\ 0 \leq A - z \leq n_8\}$. The denominator adds the weights of every feasible value of x that passes the selection rule. The resulting distribution sums to one and has no unknown parameter.

Different prompts use independent response streams. The joint conditional law over admitted prompts is therefore the product of their fibers in Equation (7). The pooled rule draws synchronized configurations from that product, recomputes every prompt deviance and every top-q rank for each configuration, and computes a final weak upper rank for the adaptive maximum. The same argument from Section A.3 applies: reordering the conditional draws only reorders their final scores. Their ranks therefore give a conditional rejection probability of at most α. Averaging over the admission vector, pair identities, hit totals, and token totals gives the unconditional result. With no admitted prompt, the returned value is $p = 1$.

B Audit Algorithms and a Worked Example

Both detectors send a fixed prompt with a chosen top-k value and record one returned token. The steps below show which choices are fixed before the final responses are seen, how requests count toward the budget, and where Water-Sonar uses clean calibration audits.

Every initial request counts toward B, even if its prompt yields no usable token group. With $B = 2{,}000$ and two selected groups, we use 15 initial responses per prompt and menu (960 total), then 260 fresh responses per selected prompt and menu (1,040 total). With $B = 5{,}000$, the corresponding counts are 39 (2,496 total) and 626 (2,504 total). If fewer than two groups are usable, the formula spreads the remaining requests over those selected. The actual request count never exceeds B.

Algorithm 2 Water-Sonar on the paired crafted-prompt grid

Require: Frozen 100-cell prompt grid, budget B, menus $(32, 128)$, frozen B-specific feature scaler and score, 1,000 independent clean calibration scores

1: Allocate one quarter of target requests to the narrow-menu arm and the rest to the wide-menu arm across the 100 prompt cells.
2: Collect independent one-token responses; retain returned token identities, cell IDs, and arm labels.
3: Build a $2 \times K$ count table at each cell and compute its menu-sensitive statistics.
4: Pool within-cell counts to compute the cross-family rank-similarity statistic.
5: Aggregate the ordered 16 features listed in Section D; standardize with the frozen development mean and scale.
6: Apply the budget-specific frozen ridge-logistic score to get s.
7: **return** $p_{\mathrm{cal}}(s)$ from Equation (5) and decision $\mathbf{1}[p_{\mathrm{cal}}(s) \le \alpha]$.

B.1 A THREE-TOKEN EXAMPLE OF THE MENU EFFECT

This is a calculation to illustrate the method, not a measured result. Let the narrow menu be $S = \{a, b\}$ with ordinary probabilities $(1/2, 1/2)$, and the wide menu be $T = \{a, b, c\}$ with probabilities $(2/5, 2/5, 1/5)$. The ordinary odds of a against b are one in both menus. Let a one-layer tournament prefer a and c, but not b. In S, preferred mass is $s_1(S) = 1/2$; Equation (3) gives returned probabilities $(3/4, 1/4)$ and odds three. In T, preferred mass is $s_1(T) = 3/5$, so the returned probabilities are $(14/25, 4/25, 7/25)$ and the odds of a against b are $14/4 = 3.5$. Adding c therefore changes the odds between a and b, although both tokens are available in both menus. Fixed token scores cannot produce that change.

B.2 WHAT THE CONDITIONAL TEST HOLDS FIXED

Suppose a selected prompt has two confirmation rows over three shared token categories, $C_1 = (8, 4, 2)$ and $C_2 = (4, 8, 2)$. The row totals are both 14, and the column totals are $(12, 12, 4)$. Under the fixed-score null, every table with these margins has a known probability. For this particular table, Equation (6) gives

$$\frac{\binom{12}{8}\binom{12}{4}\binom{4}{2}}{\binom{28}{14}}.$$

The denominator counts every way to place the 14 first-arm hits among the 28 hits in these categories; the numerator counts those that give the first row shown above. After fixing the row and column totals, neither the token probabilities nor the chance of a response outside these categories remains in the calculation. Water-Zero draws tables with these same totals for each selected group, recomputes its statistic for each draw, and reports where the observed value ranks.

C IMPLEMENTATION AND RESULT RECORDS

Frozen-model generation and tokenization use the Transformers library Wolf et al. (2020). The supplement packages the detector code, fixed protocols, fitted scores, audit records, and hashes used to check the reported results.

Water-Zero. Algorithm 1 gives the complete selected rule, including the pilot tie breaks and the exact-test draws. On Qwen2.5-1.5B, we chose its settings from 16 development combinations of menu pair, prompt count, pilot budget share, and number of selected groups. We ranked eligible settings by mean marked-audit detection, then weakest-key detection, plain rejection, and request count. The chosen rule uses top-(8,32), 32 pilot prompts, half the budget for the pilot, at most two confirmation groups, and 999 conditional draws. Final evaluation uses new prefixes, keys, and response streams.

The packaged final records contain 18,000 audit rows with split details, p-values, decisions, and summary rates. A separate response-cost study reuses each audit's response stream at 49 ordered

checkpoints from 500 to 5,000 requests, recomputing the exact test from the counts at each checkpoint. Each model has 1,000 plain audits and 100 marked audits per key. The archived audit rows support checks of the reported values; the original confirmation tables were not retained for replaying every earlier conditional draw.

OpenGen transfer. We follow the released source-row order, take the first 96 distinct rows that meet the C4 length rule, and evaluate positions 48–95 with the C4-selected detector unchanged. The archived source IDs and prompt texts identify the 18-word, authors' two-sentence, and full-context versions of each row. Separate response streams are used for these checks.

Water-Sonar. The crafted study uses a fixed grid of 50 prefixes, drawn once with seed 64. Each prefix is placed in two otherwise identical templates asking for a country with either a large population or a large area. The packaged protocol preserves the literal wrapper and prefix order. The requests use temperature one, top-$p = 1$, and top-(32,128), with one quarter of responses assigned to top-32. Section D defines the ordered 16 features, their preprocessing, and the ridge-logistic score. A separate score is fitted for each of the four response budgets with ridge penalty 10.

Development, validation, and final keys are disjoint, and the final study adds three model checkpoints not used in development or validation. Each final decision ranks its score against 1,000 separate clean audits for the same model and budget. The primary comparisons use $\alpha = 0.05$; a separately fixed low-FPR view uses $\alpha = 0.01$. The supplement stores the fitted values and complete result summary, with a hash map connecting displayed tables to their underlying records. The packaged checks verify the displayed results from derived records; the summary also records hashes for the raw token-score streams and family shards.

D Water-Sonar Feature and Score Definitions

This section defines every input to Water-Sonar's 16-feature score so it can be rebuilt from the returned token counts. The first two features combine responses from both top-k settings; the other 14 retain which setting produced each response. The combined rank-similarity feature is related to Water-Probe's prompt-pair statistic Liu et al. (2025). Because this feature can work without a menu change, we test the menu's separate contribution with a matched refit in Section G.

D.1 Statistics for one prompt cell

Let a_{ci} and b_{ci} be the numbers of times token i is returned at cell c under top-32 and top-128. We include every token seen under either setting and enter a zero where a token is missing from one setting. Write $N_{a,c} = \sum_i a_{ci}$ and $N_{b,c} = \sum_i b_{ci}$, and let $p_{ci} = a_{ci}/N_{a,c}$ and $q_{ci} = b_{ci}/N_{b,c}$ when the corresponding row is nonempty. Empty rows map to all-zero normalized vectors. The cell deviance is G_c^2 from Equation (4), applied to this full observed union. The other cell statistics are

$$\mathrm{TV}_c = \tfrac{1}{2}\sum_i |p_{ci} - q_{ci}|, \qquad W_c = \sum_{i:a_{ci}>0} q_{ci}, \tag{8}$$

$$U_{a,c} = |\{i : a_{ci} > 0\}|, \qquad U_{b,c} = |\{i : b_{ci} > 0\}|, \tag{9}$$

$$R_c = \frac{U_{b,c}}{\max(U_{a,c}, 1)}, \qquad D_c = |H(p_c) - H(q_c)|, \tag{10}$$

where $H(p) = -\sum_{i:p_i>0} p_i \log p_i$ and $H(0) = 0$. For an entirely empty cell, the implementation returns $G_c^2 = \mathrm{TV}_c = W_c = U_{a,c} = U_{b,c} = D_c = 0$ and $R_c = 1$. These rules give a finite value to every feature. The 100 prompt cells are ordered as 50 prefixes from the first question family followed by the same 50 prefixes from the second.

D.2 Pooling and rank similarity

For the pooled coordinate, replace the two arm counts at each cell by their sum. For each pair of the 50 prefix IDs, take the up-to-50 most frequent positive-count tokens at each prefix within each family, with token ID breaking equal-count ties. Keep tokens that appear in the resulting unions for both families. Within a family, rank each kept token by its pooled count at each of the

two prefixes, assigning the minimum rank to ties and rank one to zero-count tokens. Replace the difference between its two ranks by its sign in $\{-1, 0, 1\}$. The cosine similarity of the two families' sign vectors is one pair score. A pair with two zero vectors scores one; a pair with exactly one zero vector scores zero. The first feature is the mean over all computable prefix pairs. If no pair is computable, its score is -1; the second feature is then zero, otherwise one. The second feature records whether the pooled score could be computed, keeping a missing value distinct from a real low-similarity result.

D.3 Ordered 16-vector

Let $\langle x \rangle$ denote a mean over all 100 cells, let $\mathrm{top}_r(x)$ denote the mean of the largest $\max(1, \lceil 100r \rceil)$ cell values, and let sd use the population standard deviation. The vector passed to the trained score has this fixed order:

$$
\begin{aligned}
z_1 &= \text{pooled rank similarity}, & z_2 &= \text{pooled-score computability},\\
z_3 &= \sum_c G_c^2, & z_4 &= \langle G_c^2 \rangle,\\
z_5 &= \max_c G_c^2, & z_6 &= \mathrm{top}_{1/4}(G_c^2),\\
z_7 &= \mathrm{top}_{1/2}(G_c^2), & z_8 &= \langle \mathrm{TV}_c \rangle,\\
z_9 &= \max_c \mathrm{TV}_c, & z_{10} &= \langle W_c \rangle,\\
z_{11} &= \mathrm{sd}(W_c), & z_{12} &= \langle U_{a,c} \rangle,\\
z_{13} &= \langle U_{b,c} \rangle, & z_{14} &= \langle R_c \rangle,\\
z_{15} &= \langle D_c \rangle, & z_{16} &= \left| \langle G_c^2 \rangle_{c \le 50} - \langle G_c^2 \rangle_{c > 50} \right|.
\end{aligned}
$$

We compute the 16 features separately for each complete target audit. The calculation uses no watermark key, target token score, or clean-model probability.

D.4 Frozen training and clean rank

For each response budget, the development data set a mean μ_j and standard deviation σ_j for every feature. We replace a standard deviation below 10^{-10} by one and set $\tilde{z}_j = (z_j - \mu_j)/\sigma_j$. The fitted score is the linear value $s = \beta_0 + \sum_{j=1}^{16} \beta_j \tilde{z}_j$; ranking does not need its logistic probability. During development, we minimize balanced logistic loss with ridge penalty $10\|\beta\|_2^2/2$ and leave the intercept unpenalized. Each class gets half the total weight. Newton steps solve the penalized score equations until the maximum step is below 10^{-9} or 100 steps are reached. The supplement includes the frozen means, scales, coefficients, and intercepts for all four budgets.

Once these values are frozen, a target is compared with 1,000 separately collected clean scores from the same model, prompt grid, and budget. The upper-tail rank in Equation (5) includes one for the target and treats equal scores as clean scores at least as large as the target. For n clean calibration audits, the smallest possible p is $1/(n+1)$. Thus a 0.05 decision is possible only when $n \ge 19$, and a 0.01 decision requires $n \ge 99$. The final study uses $n = 1{,}000$ for a finer rank and a stable measured decision threshold.

E Detailed Result Views

The main tables report the Water-Zero natural-prompt results and the Water-Sonar family averages. Here we show how results change with response budget and across individual models, without repeating the main tables.

Figure 3 reuses response streams within each audit and is separate from the independently drawn endpoints in Table 1. It shows a clear change at low budgets. Selected Water-Zero TPR rises from 0.318–0.349 at 500 responses to 0.807–0.853 at 1,500 and 0.950–0.962 at 3,000. It exceeds the pooled rule on all three models from 1,200 responses onward and reaches 0.982–0.989 at 5,000. Across all checkpoints and both rules, observed false-positive rates stay between 0.032 and 0.064 around the 0.05 target.

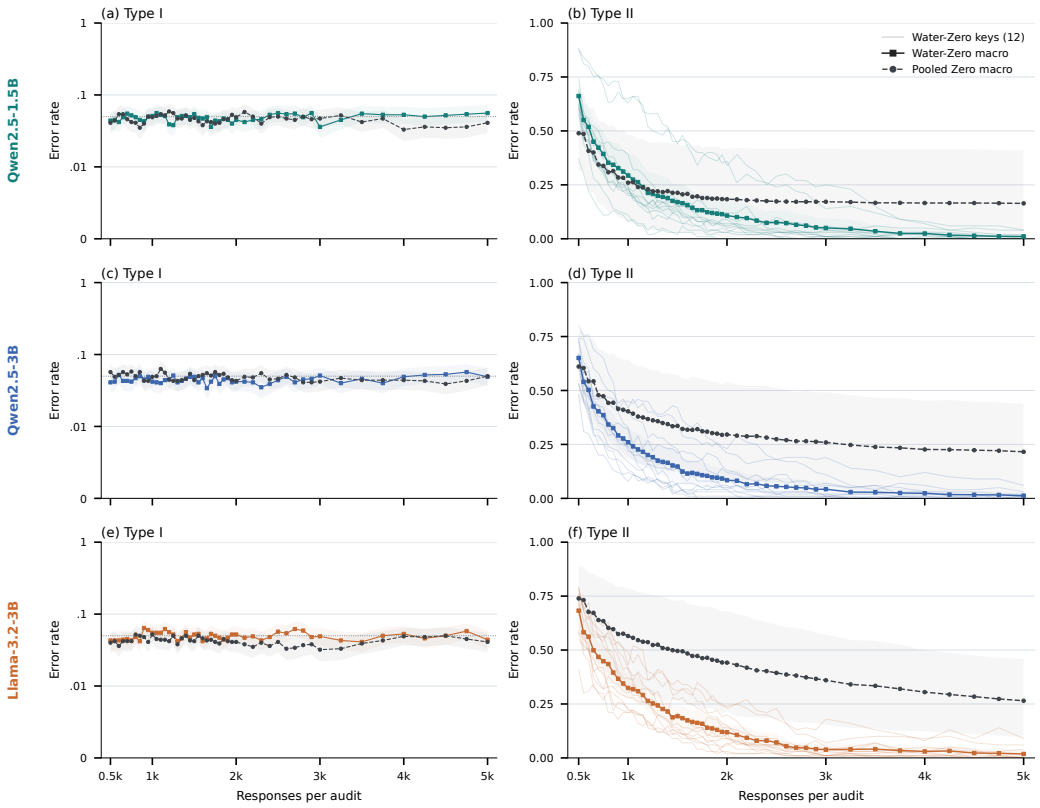

Figure 3: **Water-Zero error by model and response budget.** Each row is a base model. The left column shows Type I error (false positives); the right shows Type II error (missed watermarks). Thin colored lines show all 12 held-out SynthID keys, with the square-marked line giving their mean. Gray dashed circles show the pooled Water-Zero mean. Each curve has 49 measured points from 500 to 5,000 responses. Later points use more responses from the same audit streams. Type I bands use 1,000 plain audits; Type II bands resample keys, with 100 audits per key. On the Type I axis, 0–.01, .01–.1, and .1–1 occupy equal height.

Figure 4 reads progressively longer portions of the original 20,000-response Water-Sonar audits. It is separate from the independently drawn low-budget endpoints in Table 7. Across the three models, mean detection over nine families rises from 0.795 at 2,000 responses to 0.863 at 20,000. The largest gains are on EXP-edit (0.039 to 0.285) and ITS-edit (0.178 to 0.618); most other families are already near full detection at 2,000. Observed false-positive rates range from 0.024 to 0.072 across the 147 model–budget points, around the 0.05 target.

Table 3 places each model's mean TPR across families beside its false-positive rate from separate plain audits at all four response budgets. The same rows let us compare the two detectors without repeating the endpoint FPR in another table.

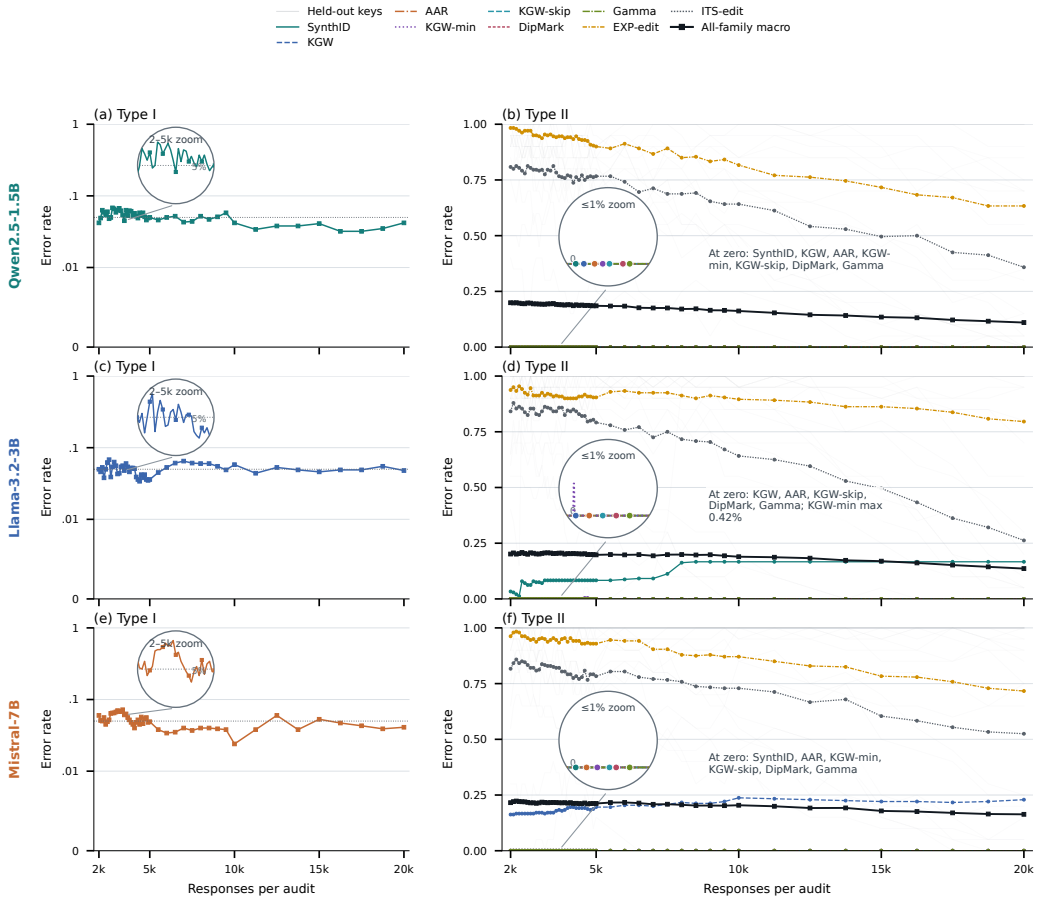

Figure 4: **Water-Sonar error at 49 response budgets.** Rows show Qwen2.5-1.5B, Llama-3.2-3B, and Mistral-7B in the crafted-prompt study. The left column shows false-positive rates from 1,000 held-out plain audits per model and budget. The dotted line marks 0.05, and circles enlarge the region with 2,000–5,000 responses and 2–8% error. The right column shows missed-watermark rates for nine families: colored lines average 12 held-out keys, faint gray lines show individual keys above 0.01, and black squares show the mean over families. Circles enlarge the near-zero region; colored dots and adjacent names distinguish curves that meet at zero. Each key has 20 evaluation audits. The 49 points use progressively more responses from the same 20,000-response top-k streams. At each point, a separate detector is trained on the original development keys and calibrated with 1,000 separate clean audits. The left axis gives equal height to 0–.01, .01–.1, and .1–1; the right runs from 0 to 1.

Table 3: **Crafted-prompt detection and plain false positives by model and response budget.** TPR is the mean over nine watermark families (240 marked audits per family); FPR uses a separate 1,000-audit plain panel. Water-Sonar uses a separate clean calibration panel. This combines the endpoint TPR comparison with the complete four-budget FPR check.

		Water-Zero		Water-Sonar	
Model	Responses	TPR	FPR	TPR	FPR
		Qwen models			
Qwen2.5-1.5B	2,000	0.178	0.052	0.799	0.044
	5,000	0.212	0.045	0.812	0.059
	10,000	0.229	0.052	0.837	0.047
	20,000	0.245	0.046	0.890	0.042
Qwen2.5-3B	2,000	0.181	0.048	0.785	0.063
	5,000	0.239	0.051	0.789	0.049
	10,000	0.245	0.044	0.808	0.049
	20,000	0.259	0.040	0.857	0.051
Qwen2.5-7B	2,000	0.135	0.047	0.794	0.052
	5,000	0.181	0.053	0.783	0.036
	10,000	0.218	0.054	0.809	0.046
	20,000	0.229	0.062	0.872	0.054
		Llama models			
Llama-3.2-3B	2,000	0.140	0.049	0.792	0.050
	5,000	0.185	0.040	0.800	0.030
	10,000	0.200	0.048	0.805	0.037
	20,000	0.215	0.056	0.864	0.048
Llama-3.1-8B	2,000	0.155	0.039	0.794	0.050
	5,000	0.211	0.055	0.812	0.062
	10,000	0.231	0.049	0.817	0.032
	20,000	0.249	0.043	0.851	0.045
Llama2-7B	2,000	0.141	0.047	0.804	0.042
	5,000	0.180	0.038	0.827	0.054
	10,000	0.190	0.057	0.858	0.033
	20,000	0.194	0.040	0.902	0.053
Llama2-13B	2,000	0.117	0.047	0.818	0.046
	5,000	0.172	0.055	0.837	0.038
	10,000	0.179	0.045	0.865	0.041
	20,000	0.185	0.040	0.905	0.066
		Other open models			
OPT-2.7B	2,000	0.155	0.041	0.795	0.042
	5,000	0.194	0.050	0.805	0.045
	10,000	0.209	0.047	0.836	0.062
	20,000	0.231	0.055	0.910	0.071
Mistral-7B	2,000	0.169	0.030	0.772	0.038
	5,000	0.209	0.055	0.779	0.041
	10,000	0.219	0.058	0.799	0.044
	20,000	0.242	0.054	0.837	0.041

F Complete Held-out Result Matrices

The main text reports Water-Zero results on natural prefixes and mean Water-Sonar results by family. The tables here show differences across keys, models, and budgets that those averages can hide. Every number comes from the frozen final records; we made no new model calls for these tables. The natural-prefix and crafted-prompt results remain separate studies with different prompts, keys, and decision rules.

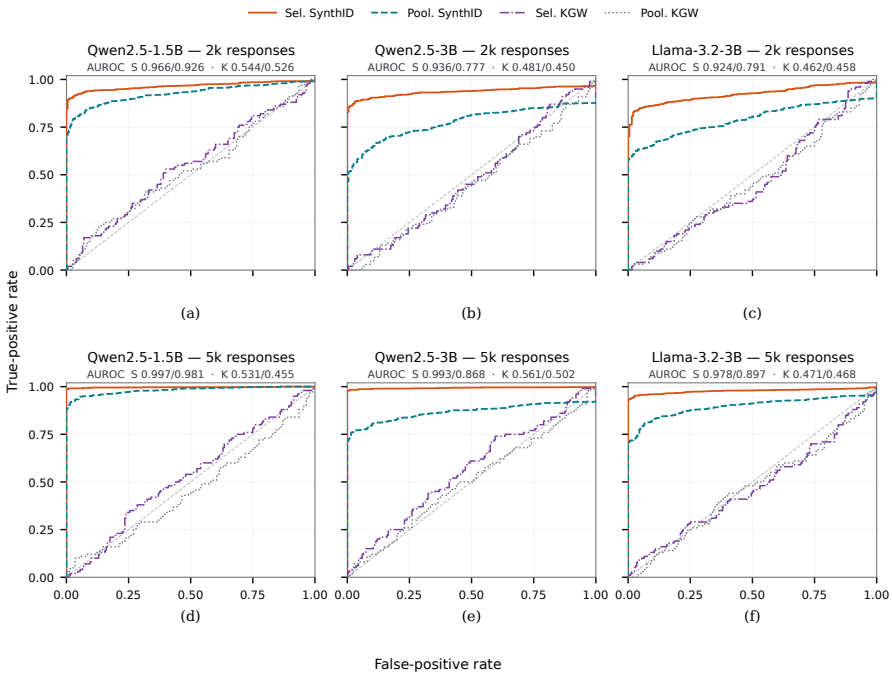

Figure 5: **Water-Zero ROC curves on natural prefixes.** Columns are the three tested models; rows use 2,000 and 5,000 responses. Each SynthID curve compares 1,200 marked audits with 200 plain audits; each KGW control curve compares 100 fixed-bias audits with the corresponding 200 plain audits. Scores are negative exact p-values. Selected and pooled rules use independent response streams. Panel subtitles report selected/pooled AUROC for S (SynthID) and K (KGW). The diagonal is chance ranking; Table 1 retains the fixed 0.05 decision.

F.1 Water-Zero ranking on natural prefixes

Figure 5 shows how the exact test ranks audits as its threshold varies. The selected rule improves SynthID ranking over the pooled rule on the same natural-prefix panel, while the fixed-bias KGW control stays near chance. The primary decision and its control rates are in Table 1.

F.2 Natural-prefix key variation

Table 4: **Key-level spread in the natural-prefix Water-Zero study.** Each key has 100 held-out marked audits per model, budget, and rule. The 12 keys are 24–35. Counts in the last two columns use selected Water-Zero. The model rows summarize the same audits as Table 1; they are not new experiments.

Model	B	Selected	Pooled	Min. key	Median key	Max. key	≥ 0.8	≥ 0.95
Qwen2.5-1.5B	2k	0.914	0.818	0.660	0.950	0.990	11/12	7/12
	5k	0.991	0.937	0.940	1.000	1.000	12/12	11/12
Qwen2.5-3B	2k	0.883	0.636	0.590	0.930	1.000	10/12	6/12
	5k	0.984	0.765	0.960	0.990	1.000	12/12	12/12
Llama-3.2-3B	2k	0.844	0.633	0.550	0.875	1.000	9/12	4/12
	5k	0.953	0.781	0.760	0.985	1.000	11/12	8/12

Table 4 gives the lowest, middle, and highest TPR across held-out keys, plus the number of keys above two TPR thresholds. The supplement gives all 12 keys for each model and budget. We average

across keys, not individual token responses, so a weak key remains visible even when the overall mean is high. The two budgets use independent audit streams, so the two numbers in a row are not a paired trajectory for one audit.

F.3 OpenGen transfer and prompt length

Table 1 places the 18-word OpenGen transfer beside C4 under the same test settings. Table 5 follows the same OpenGen source rows through three text lengths. The two-sentence version is the authors' generation prompt. The complete source context is a separate length check, not the original OpenGen generation prompt. We carried every setting over from C4 without OpenGen tuning. The selected and pooled tests, watermark keys, and audit streams remain separate within every arm.

Table 5: **OpenGen prompt-length sensitivity.** The 18-word matched prompt, authors' two-sentence generation prompt, and complete source context use the same 96 released source rows, positions 48–95, C4-selected test, 12 keys, and independent draws. Each model–budget–arm–rule cell has 1,200 marked, 200 plain, and 100 fixed-bias KGW audits. All rates use $p \leq 0.05$.

		Selected SynthID TPR			Two-sentence prompt: other views			
Model	B	18 words	Two sent.	Full source	Pooled TPR	Lowest key	Plain FPR	KGW reject
Qwen 2.5-1.5B	2k	0.875	0.895	0.834	0.602	0.560	0.040	0.030
Qwen 2.5-1.5B	5k	0.993	0.969	0.969	0.708	0.840	0.015	0.080
Qwen 2.5-3B	2k	0.882	0.888	0.897	0.702	0.520	0.060	0.090
Qwen 2.5-3B	5k	0.983	0.972	0.993	0.768	0.720	0.030	0.070
Llama 3.2-3B	2k	0.861	0.833	0.917	0.743	0.130	0.040	0.000
Llama 3.2-3B	5k	0.981	0.913	0.974	0.830	0.230	0.060	0.080

The 18-word Llama-3.2-3B cell has 17 plain rejections in its primary 200-audit panel at 2,000 responses (0.085). In a separate post-primary diagnostic with a new response stream, selected Water-Zero rejects 54 of 1,000 plain audits (0.054; exact 95% interval [0.041, 0.070]) and 37 of 1,000 fixed-bias KGW audits (0.037). The pooled reference rejects 40 and 68 of 1,000 in the same two conditions. These extra audits investigate the high control rate; the original estimate remains in Table 1. The frozen diagnostic protocol and all 4,000 audit records are archived with the OpenGen transfer study.

F.4 Matched natural-prompt family sweep

Table 6 shows each corpus and model, including the separate plain FPR for both detectors. The new main Table 2 reports the family-level Sonar-N metrics, while the radar shows both detectors. Both detectors use the same final natural prefixes, watermark families, and keys, but draw independent responses. Sonar-N is trained on separate natural-prompt development data, so its results differ from the crafted-prompt Water-Sonar study in Table 7.

Table 6: **Natural detection by corpus and model.** Mean TPR and Sonar-N AUROC average nine families. SynthID TPR is separate; FPR uses independent plain audits. Each marked family cell has 240 audits.

		Mean TPR		SynthID TPR		Plain FPR		Sonar-N
Model	B	Zero	Sonar-N	Zero	Sonar-N	Zero	Sonar-N	AUROC
C4 natural 18-word prefixes								
Qwen2.5-1.5B	2k	0.166	0.509	0.925	1.000	0.025	0.052	0.734
	5k	0.253	0.512	0.992	1.000	0.045	0.044	0.745
Qwen2.5-3B	2k	0.177	0.524	0.887	1.000	0.045	0.072	0.759
	5k	0.256	0.543	0.975	1.000	0.050	0.048	0.783
Llama-3.2-3B	2k	0.169	0.555	0.842	1.000	0.055	0.047	0.792
	5k	0.245	0.573	0.963	1.000	0.045	0.038	0.810
OpenGen natural 18-word prefixes								
Qwen2.5-1.5B	2k	0.170	0.541	0.854	1.000	0.060	0.055	0.765
	5k	0.267	0.568	0.992	1.000	0.060	0.045	0.795
Qwen2.5-3B	2k	0.178	0.533	0.912	1.000	0.055	0.052	0.768
	5k	0.256	0.557	0.992	1.000	0.060	0.053	0.789
Llama-3.2-3B	2k	0.174	0.575	0.858	1.000	0.055	0.069	0.797
	5k	0.269	0.597	0.967	1.000	0.035	0.041	0.838

F.5 SynthID ranking across prompt sources

Figure 6 shows the score rankings behind the SynthID results at 2,000 responses. The two natural-prompt curves change between C4 and OpenGen. The three crafted-prompt curves come from separate audits without a C4/OpenGen split, so each is repeated across the two rows. The circular views enlarge the low-false-positive region: the two Sonar curves coincide on both Qwen models, while the Llama curves separate. Each curve uses 240 marked audits. Plain audit counts are 200 for Water-Zero, 1,000 for Sonar-N, Water-Sonar, and Water-Probe, and 100 for Gloaguen.

F.6 Crafted-prompt family coverage

Table 7 compares same-grid Water-Zero and the complete learned Water-Sonar system at four budgets and two pre-set decision levels. Each pair of rows uses the same scores and audits, making the cost of a stricter cutoff visible. This crafted-prompt study uses different prompts, keys, and decision rules from the C4/OpenGen natural-prompt study.

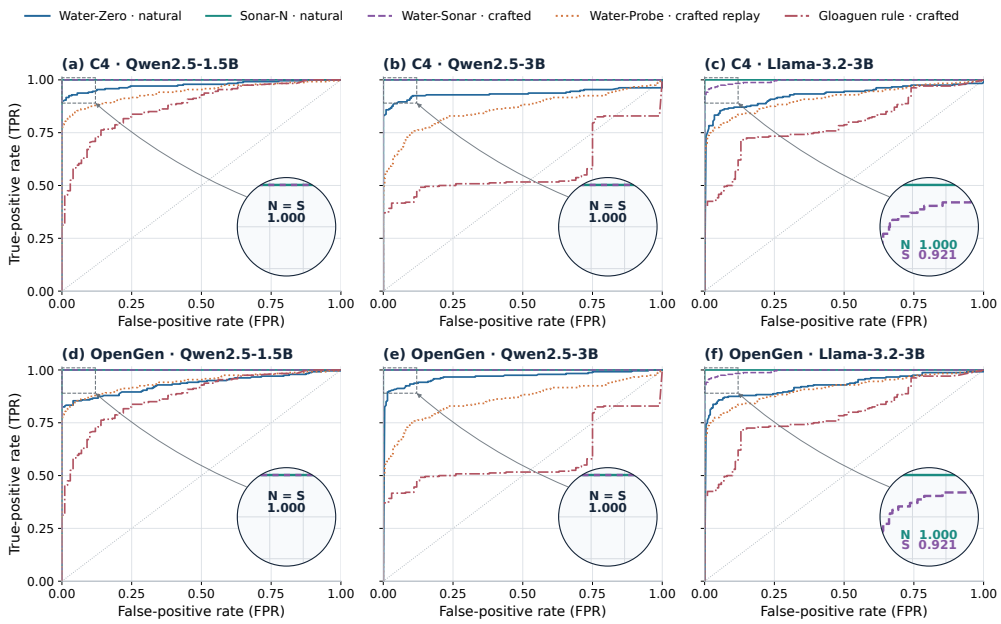

Figure 6: **SynthID ROC curves at 2,000 responses.** Rows use C4 and OpenGen natural prefixes; columns show three base models. Natural Water-Zero and Sonar-N change by row. The separate crafted-prompt Water-Sonar, matched Water-Probe replay, and locally measured Gloaguen paper rule repeat across rows. Circles enlarge FPR 0–0.12 and TPR 0.89–1.01; arrows mark their source. In the circles, N is Sonar-N and S is Water-Sonar. Qwen N and S curves coincide in these views.

Table 7: **Crafted-prompt detection across watermark families and decision levels.** Each TPR averages nine models, with 240 marked audits per model–family cell. Plain FPR averages nine separate 1,000-audit panels. Water-Zero is the exact pooled rule; Water-Sonar is the complete learned system. The two α rows use the same scores and audits with different fixed cutoffs.

		2,000		5,000		10,000		20,000	
Family	α	Zero	Sonar	Zero	Sonar	Zero	Sonar	Zero	Sonar
Support-dependent watermark									
SynthID	0.050	0.358	0.997	0.458	0.959	0.496	0.935	0.537	0.926
	0.010	0.317	0.994	0.436	0.955	0.482	0.935	0.528	0.926
N-gram or context-keyed watermarks									
KGW	0.050	0.045	0.981	0.062	0.979	0.054	0.975	0.044	0.975
	0.010	0.008	0.977	0.013	0.975	0.006	0.973	0.006	0.974
Aar	0.050	0.688	1.000	0.945	1.000	0.993	1.000	1.000	1.000
	0.010	0.514	1.000	0.881	1.000	0.981	1.000	0.998	1.000
DiPMark	0.050	0.039	1.000	0.048	1.000	0.054	1.000	0.046	1.000
	0.010	0.004	1.000	0.012	1.000	0.010	1.000	0.008	1.000
γ-RW	0.050	0.048	1.000	0.049	1.000	0.053	1.000	0.040	1.000
	0.010	0.010	1.000	0.008	1.000	0.011	1.000	0.006	1.000
KGW-Min	0.050	0.044	0.998	0.045	0.993	0.053	1.000	0.047	1.000
	0.010	0.010	0.991	0.007	0.987	0.007	0.999	0.007	1.000
KGW-Skip	0.050	0.041	1.000	0.050	1.000	0.046	1.000	0.045	1.000
	0.010	0.011	0.999	0.010	0.998	0.012	1.000	0.003	1.000
Fixed-key-list watermarks									
EXP-Edit	0.050	0.045	0.044	0.061	0.096	0.064	0.193	0.071	0.371
	0.010	0.006	0.014	0.010	0.044	0.012	0.121	0.017	0.281
ITS-Edit	0.050	0.061	0.133	0.065	0.218	0.105	0.331	0.219	0.617
	0.010	0.011	0.052	0.012	0.123	0.027	0.214	0.085	0.470
Mean TPR	0.050	0.152	0.795	0.198	0.805	0.213	0.826	0.228	0.876
	0.010	0.099	0.781	0.154	0.787	0.172	0.805	0.184	0.850
Plain FPR	0.050	0.044	0.047	0.049	0.046	0.050	0.043	0.048	0.052
	0.010	0.008	0.010	0.009	0.010	0.008	0.010	0.010	0.009

F.7 RANKING WITHOUT A FIXED THRESHOLD

AUROC measures how often a marked audit scores above a plain audit. Table 8 gives nine-model means and ranges at 2,000 and 20,000 responses. EXP-Edit's ranking improves with budget. The full model-by-family heatmap remains in the archived figure file.

Table 8: **Threshold-free Water-Sonar ranking.** Nine-model AUROC mean and range at two response budgets. Each model–family cell has 240 marked and 1,000 separate plain audits; the corresponding fixed-cutoff TPR is in Table 7.

	2,000 responses			20,000 responses		
Family	Mean	Min.	Max.	Mean	Min.	Max.
Support-dependent watermark						
SynthID	0.999	0.995	1.000	0.926	0.833	1.000
N-gram or context-keyed watermark						
KGW	0.987	0.881	1.000	0.979	0.815	1.000
Aar	1.000	1.000	1.000	1.000	1.000	1.000
DiPMark	1.000	1.000	1.000	1.000	1.000	1.000
γ-Reweight	1.000	1.000	1.000	1.000	1.000	1.000
KGW-Min	0.999	0.994	1.000	1.000	1.000	1.000
KGW-Skip	1.000	1.000	1.000	1.000	1.000	1.000
Fixed-key-list watermark						
EXP-Edit	0.418	0.259	0.549	0.621	0.446	0.845
ITS-Edit	0.598	0.505	0.676	0.839	0.732	0.904

F.8 WATER-SONAR ACROSS MODELS AND WATERMARK FAMILIES

Table 9 puts each model's endpoint means beside its weakest final key. The fixed-key-list families show lower power, and five SynthID models have a key with no detected audits at 20,000 responses despite high means. Each mean weights 12 held-out keys equally; each key minimum uses 20 audits and describes only this tested panel. The 2,000-response minima and intermediate budgets remain in the packaged results. This is the complete crafted-prompt Sonar system; the table does not isolate any feature or the top-k change. Plain FPR is in Table 3, while natural-prefix Water-Zero is a separate study.

Table 9: **Water-Sonar TPR by model, family, and held-out key.** Means at 2,000 and 20,000 responses use 240 marked audits per cell. The 20,000-response minimum uses 20 audits for each of 12 keys; it is not a confidence bound. Crafted prompts, $\alpha = 0.05$.

Model	TPR view	SynthID	KGW	Aar	DiPMark	γ-RW	KGW-Min	KGW-Skip	EXP-Edit	ITS-Edit
Qwen models										
Qwen2.5-1.5B	2k mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.017	0.175
	20k mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.367	0.642
	20k min.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.150
Qwen2.5-3B	2k mean	1.000	1.000	1.000	1.000	1.000	0.983	1.000	0.017	0.067
	20k mean	0.917	1.000	1.000	1.000	1.000	1.000	1.000	0.304	0.496
	20k min.	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
Qwen2.5-7B	2k mean	1.000	1.000	1.000	1.000	1.000	0.996	1.000	0.013	0.142
	20k mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.163	0.683
	20k min.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.150
Llama models										
Llama-3.2-3B	2k mean	0.971	1.000	1.000	1.000	1.000	1.000	1.000	0.046	0.113
	20k mean	0.833	1.000	1.000	1.000	1.000	1.000	1.000	0.204	0.738
	20k min.	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.050
Llama-3.1-8B	2k mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.046	0.096
	20k mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.146	0.517
	20k min.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
Llama2-7B	2k mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.100	0.138
	20k mean	0.833	1.000	1.000	1.000	1.000	1.000	1.000	0.596	0.692
	20k min.	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
Llama2-13B	2k mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.075	0.283
	20k mean	0.833	1.000	1.000	1.000	1.000	1.000	1.000	0.629	0.683
	20k min.	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000

Continued on next page

Model	TPR view	SynthID	KGW	Aar	DiPMark	γ-RW	KGW-Min	KGW-Skip	EXP-Edit	ITS-Edit
Other open models										
OPT-2.7B	2k mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.062	0.092
	20k mean	0.917	1.000	1.000	1.000	1.000	1.000	1.000	0.650	0.625
	20k min.	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
Mistral-7B	2k mean	1.000	0.829	1.000	1.000	1.000	1.000	1.000	0.021	0.096
	20k mean	1.000	0.771	1.000	1.000	1.000	1.000	1.000	0.283	0.475
	20k min.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
Nine-model mean	2k mean	0.997	0.981	1.000	1.000	1.000	0.998	1.000	0.044	0.133
	20k mean	0.926	0.975	1.000	1.000	1.000	1.000	1.000	0.371	0.617
	20k min.	0.444	0.889	1.000	1.000	1.000	1.000	1.000	0.000	0.039

G TOP-k INTERVENTION ABLATIONS

G.1 WATER-ZERO: WHICH MENU PAIRS EXPOSE THE SIGNAL?

We kept the unedited C4 prefixes, final SynthID keys, exact test, and response budgets fixed, then tested ten top-k pairs chosen in advance. Each pair uses fresh responses and has its own plain and fixed-bias KGW controls. The main result uses $8/32$; Table 10 shows whether detection depends on that one choice. The new $8/32$ audits repeat the main setting with fresh responses rather than reuse its measurements.

Table 10: **Water-Zero across top-k pairs on unedited text.** Only the menu pair changes; the exact test stays fixed. Each model, pair, and budget has 12 SynthID keys with 40 audits each, plus 200 plain audits and 100 KGW controls. TPR averages keys, and KGW rejection averages models. B is a maximum: the $2/8$ test can stop early. Rows with identical menus in both arms are controls.

		SynthID TPR				Plain FPR			KGW
Menu pair	B	Qwen 1.5B	Qwen 3B	Llama 3B	Mean	Qwen 1.5B	Qwen 3B	Llama 3B	Mean
No menu change									
$8/8$	2k	0.060	0.040	0.046	0.049	0.070	0.040	0.050	0.047
$8/8$	5k	0.063	0.046	0.069	0.059	0.060	0.040	0.035	0.040
$32/32$	2k	0.046	0.040	0.052	0.046	0.080	0.030	0.060	0.040
$32/32$	5k	0.050	0.035	0.029	0.038	0.030	0.035	0.065	0.030
Small or moderate menus									
$2/8$	2k	0.588	0.517	0.344	0.483	0.035	0.060	0.065	0.037
$2/8$	5k	0.721	0.608	0.517	0.615	0.050	0.055	0.030	0.050
$4/16$	2k	0.817	0.862	0.733	0.804	0.065	0.065	0.050	0.043
$4/16$	5k	0.958	0.963	0.929	0.950	0.030	0.055	0.045	0.067
$8/16$	2k	0.840	0.862	0.800	0.834	0.050	0.015	0.075	0.037
$8/16$	5k	0.948	0.977	0.954	0.960	0.040	0.025	0.035	0.040
$8/32$	2k	0.902	0.904	0.890	0.899	0.060	0.065	0.025	0.050
$8/32$	5k	1.000	0.979	0.987	0.989	0.030	0.025	0.050	0.057
Wide menus									
$8/64$	2k	0.906	0.933	0.904	0.915	0.035	0.020	0.055	0.063
$8/64$	5k	0.990	0.987	0.994	0.990	0.055	0.030	0.060	0.043
$8/128$	2k	0.908	0.850	0.877	0.878	0.035	0.040	0.075	0.053
$8/128$	5k	0.990	0.987	0.994	0.990	0.045	0.085	0.075	0.037
$16/64$	2k	0.904	0.898	0.842	0.881	0.075	0.055	0.050	0.040
$16/64$	5k	0.996	0.988	0.987	0.990	0.050	0.045	0.075	0.060
$32/128$	2k	0.767	0.779	0.748	0.765	0.030	0.050	0.070	0.043
$32/128$	5k	0.971	0.963	0.967	0.967	0.040	0.045	0.040	0.037

When both requests use the same menu, mean SynthID TPR is 0.049 for $8/8$ and 0.046 for $32/32$ at 2,000 responses. These values are near the 0.05 decision level. Changing the pair to $8/16$ raises mean TPR to 0.834; $8/32$ reaches 0.899, and $8/64$ reaches 0.915. At 5,000 responses, the latter two reach 0.989 and 0.990. Detection improves because the allowed tokens change, and the result holds across several menu pairs. With the same menu in both arms, the odds between two shared tokens

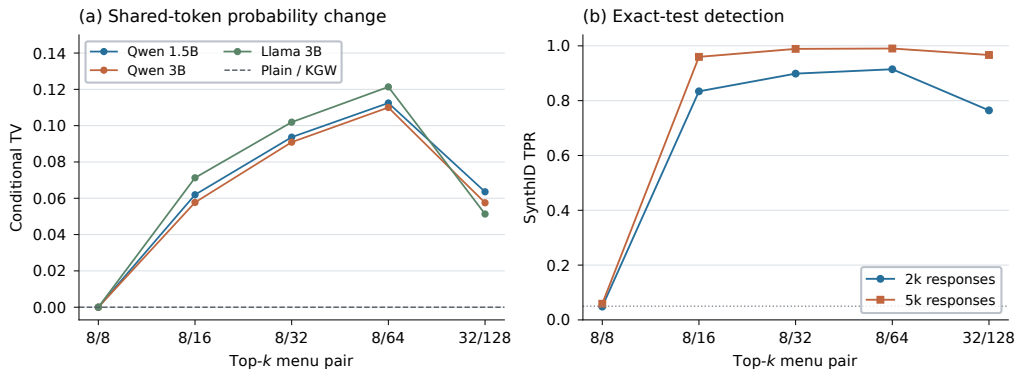

Figure 7: **Shared-token change and Water-Zero power across menu pairs.** (a) Mean shared-token conditional TV on 48 natural prefixes and 12 SynthID keys; plain and KGW controls remain near zero. (b) Three-model mean TPR at 2,000 and 5,000 responses, with 40 fresh audits per key and point. The $8/8$ control leaves the menu unchanged. Table 10 reports all ten pairs and their control rates.

cannot change between the arms. The $2/8$ pair is weaker: at a 2,000-response allowance it uses only 1,639–1,795 responses on average across models, has no usable group in 19.7–34.7% of audits, and has zero TPR on its weakest key in every model. In contrast, $8/32$ uses all 2,000 responses, has no zero-group audit, and has weakest-key TPR 0.60–0.75. Across pairs and budgets, the three-model mean plain FPR ranges from 0.033 to 0.068 and the mean KGW control rejection from 0.030 to 0.067. Absolute menu sizes matter as well as their ratio: at a fourfold ratio, mean 2,000-response TPR ranges from 0.765 for $32/128$ to 0.899 for $8/32$.

Why the menu pair matters. We examined the full next-token distributions behind the frozen natural panel. For each prompt and key, we kept only tokens available under both menus, made their probabilities sum to one again, and measured the difference between the resulting distributions using total variation (TV). This asks directly whether the relative chances of shared tokens change with the menu. Across three models, 48 prefixes, and 12 SynthID keys, mean TV rises from 0 for $8/8$ to 0.064, 0.095, and 0.115 for $8/16$, $8/32$, and $8/64$. It is 0.057 for $32/128$. Plain and fixed-bias KGW controls remain below 10^{-6} for every pair, matching their stable shared-token odds. In one of 1,728 SynthID prompt–key cells at $8/16$, the wide arm gives zero mass to the shared support; its conditional TV is undefined and excluded from that mean. The complete cell-level record includes this zero-mass case. Figure 7 places the token-distribution change beside detection rates from fresh audits. Menu pairs with larger changes also have higher detection here. Five pairs, however, are not enough to infer a general numerical rule linking TV to detection power.

On the locked 18-word OpenGen panel, mean SynthID shared-token TV at $8/32$ is 0.107, 0.092, and 0.093 across the three models (mean 0.097, versus 0.095 on C4). Plain and KGW controls stay below 10^{-6}, and the $8/8$ control has zero TV. This confirms the mechanism behind the transfer in Table 1; the archived transfer records give all five menu pairs and their cell-level values.

G.2 Water-Sonar: matched fixed-menu and refit controls

Table 11 compares the original top-k pair with two tests that use the same menu twice. Budgets and audit IDs are matched.

Table 11: **What changes when Sonar uses the same top-k twice?** All three settings use matched budgets and audit IDs. Mix is the original $32/128$ score; Frozen applies it to $128/128$ responses; Refit trains a new score on $128/128$ development audits. Δ is Mix minus Refit. Rates average nine models, each with 240 marked audits per family. TPR and FPR use the 0.05 decision; AUROC uses all score thresholds.

	2,000 responses				20,000 responses			
Watermark family	Mix	Frozen	Refit	Δ	Mix	Frozen	Refit	Δ
Support-dependent watermark								
SynthID	0.997	0.462	0.959	+0.038	0.926	0.028	0.998	-0.072
N-gram or context-keyed watermarks								
KGW	0.981	0.389	0.978	+0.003	0.975	0.000	1.000	-0.025
Aar	1.000	0.884	1.000	0.000	1.000	0.019	1.000	0.000
DiPMark	1.000	0.629	1.000	0.000	1.000	0.000	1.000	0.000
γ-Reweight	1.000	0.646	1.000	0.000	1.000	0.000	1.000	0.000
KGW-Min	0.998	0.320	0.998	0.000	1.000	0.000	1.000	0.000
KGW-Skip	1.000	0.442	0.999	+0.001	1.000	0.000	1.000	0.000
Fixed-key-list watermarks								
EXP-Edit	0.044	0.062	0.081	-0.037	0.371	0.516	0.026	+0.345
ITS-Edit	0.133	0.108	0.102	+0.031	0.617	0.252	0.875	-0.258
Overall summary								
Mean TPR	0.795	0.438	0.791	+0.004	0.876	0.091	0.878	-0.001
Mean AUROC	0.889	0.684	0.892	-0.003	0.929	0.148	0.923	+0.006
Plain FPR	0.047	0.047	0.051	-0.004	0.052	0.046	0.049	+0.003

For each final audit, we keep the prompt, watermark key, response budget, and saved top-128 responses. We replace only its top-32 quarter with new, independent top-128 responses. This gives $32/128$ and $128/128$ records with the same budget and audit IDs. We score the $128/128$ records twice: first with the original frozen 16-feature model, then with a new model trained on $128/128$ development records. The new fit keeps the original three development models, keys 48–59, and ridge value, without further tuning; final keys 72–83 and all final audit records remain held out. Each variant has 1,000 separate clean calibration and 1,000 plain evaluation audits per model and budget. Every model–family–budget cell contains the same 240 marked audit IDs.

At 2,000 responses, mean TPR is 0.795 for the original mixed-menu learner, 0.438 when its score is frozen on fixed-menu records, and 0.791 after fitting the fixed-menu score. At 20,000 responses these values are 0.876, 0.091, and 0.878. After training the new score, mean AUROC rises from 0.684 to 0.892 at 2,000 responses and from 0.148 to 0.923 at 20,000; the original scores give 0.889 and 0.929. Plain FPR stays near 0.05 in all three settings. The original trained score uses top-k labels, yet the complete crafted-prompt system keeps nearly the same mean detection with identical menus if its score is trained for that setting.

Table 11 also shows the family-level differences. At 20,000 responses, the mixed-menu learner detects EXP-Edit at 0.371 versus 0.026 for the refit; for ITS-Edit the refit reaches 0.875 versus 0.617 for the mixed-menu learner. For the six context-keyed families, both trained variants are near or at 1.000. Water-Zero establishes a direct top-k signal, while Water-Sonar can also learn useful patterns from the crafted prompts.